UNIVERSIDAD DE SALAMANCA

DEPARTAMENTO DE ESTADÍSTICA

MÉTODOS MULTIVARIANTES PARA EVALUAR PATRONES DE ESTABILIDAD Y CAMBIO DESDE UNA PERSPECTIVA BIPLOT

SUSANA LUISA DA CUSTODIA MACHADO MENDES

2011

UNIVERSITY OF SALAMANCA

DEPARTMENT OF STATISTICS

MULTIVARIATE METHODS TO ASSESS PATTERNS OF STABILITY AND CHANGE FROM A BIPLOT PERSPECTIVE

PhD Thesis "Doctor Europaeus"

SUSANA LUISA DA CUSTODIA MACHADO MENDES

2011

MÉTODOS MULTIVARIANTES PARA EVALUAR PATRONES DE ESTABILIDAD Y CAMBIO DESDE UNA PERSPECTIVA BIPLOT

Memoria que, para optar al Grado de Doctor por el Departamento de Estadística de la Universidad de Salamanca, presenta:

Susana Luísa da Custódia Machado Mendes

Salamanca

2011

Universidad de Salamanca

Departamento de Estadística

Mª PURIFICACIÓN GALINDO VILLARDÓN

y

Mª JOSÉ FERNÁNDEZ GÓMEZ

Profesoras titulares del Departamento de Estadística de la Universidad de Salamanca

CERTIFICAN:

Que Dña. Susana Luísa da Custódia Machado Mendes, Licenciada en Matemática, Campo de las Ciencias Actuariales, ha realizado en el Departamento de Estadística de la Universidad de Salamanca, bajo su dirección, el trabajo que para optar al Grado de Doctor, presenta con el título: ***Métodos multivariantes para evaluar patrones de estabilidad y cambio desde una perspectiva biplot***; y para que conste, firman el presente certificado en Salamanca, a 12 de Diciembre de 2011

Fdo.: Mª Purificación Galindo Villardón Fdo.: Mª José Fernández Gómez

"Your work is going to fill a large part of your life, and the only way to be truly satisfied is to do what you believe is great work. And the only way to do great work is to love what you do. If you haven't found it yet, keep looking, and don't settle. As with all matters of the heart, you'll know when you find it."

(Steve Jobs)

AGRADECIMIENTOS

En primero lugar, me gustaría agradecer a mis directoras de tesis, Profesora Doctora Mª Purificación Galindo Villardón, Directora del Departamento de Estadística de la Universidad de Salamanca, y Profesora Doctora Mª José Fernández Gómez, por haberme dado la oportunidad de realizar la tesis bajo su supervisión. Agradezco la confianza depositada en mí, la amistad sincera y la esperanza de haber cumplido todas las expectativas. Vuestra dedicación, tiempo y experiencia fueron fundamentales para llegar hasta aquí. ¡Muchas gracias!

Mi especial gratitud a todos los profesores del Departamento de Estadística de la Universidad de Salamanca, por la generosidad, por sus valiosos conocimientos y confianza en mi trabajo y siempre dispuestos a ayudar.

Um agradecimento muito especial ao Professor Doutor Ulisses. Por tudo! Desde o primeiro instante (e até ao último, dos últimos momentos…), a tua disponibilidade tem sido infinita! Desde a cedência dos dados, a ajuda na compreensão destes e o constante empenho em me abrires as portas em várias vertentes da investigação. Muito obrigado! Igualmente agradeço as horas de sono perdidas e os dias a tentar perceber o que "estas setinhas" e "pontos" têm a dizer sobre "os bichos"! Nunca me esquecerei de uma observação tua (e já lá vão uns anitos!) … "A vantagem de trabalhar com um matemático, é que de um momento para o outro dá a volta ao cubo e resolve-nos o problema!"... E cá está, após umas tantas voltas a vários cubos de dados, chego até aqui! O mais sincero e profundo obrigado!

Um agradecimento especial aos professores do Centro de Ecologia Funcional (CEF) da Universidade de Coimbra, Professora Doutora Helena Freitas e Professor Doutor Miguel Pardal, pela preciosa disponibilidade na colaboração deste trabalho.

Devo gratidão ao Professor Doutor Mário Pereira, à Doutora Paula Resende, à Professora Doutora Paula Bacelar-Nicolau, à Doutora Ana Marta Gonçalves, pela generosidade em fornecer dados para as diversas aplicações incluídas nesta tese.

Uma palavra consideravelmente especial à Sónia (chegou a hora da festarola!). Sem dúvida que conhecer-te e partilhar contigo tantos desabafos (científicos, mas não só!), foi uma das grandes portas que esta tese abriu… a par da disponibilidade científica (sem a tua *acartia tonsa* nada disto seria possível!), a enorme amizade, contínua compreensão e total confiança que desde o primeiro momento depositaste no meu trabalho, foram determinantes! Tenho a certeza que só agora estamos a começar… e o (zoo/fito) plâncton que nos aguarde, pois nunca mais será o mesmo!

Ao meu "chefinho" Sérgio, um grande e forte abraço! No meio de tanto trabalho, no meio de tanta hora na ESTM, no meio de tantas viagens entre Peniche e Condeixa, sempre tiveste tempo para me ajudar! E sempre com um sorriso nos lábios! O empenho que colocas em tudo o que fazes e a capacidade de trabalho inacreditável faz de ti um exemplo a seguir!

Aos meus dois grandes amigos e eternos companheiros, Rui Pedrosa e Mário Carvalho. Eu sei que sabem o quanto gosto de vocês, e o quanto foram importantes neste percurso da minha vida, mas acreditem, a vossa amizade fez toda a diferença! Mário, obrigado por te lembrares de mim sempre, e em todos os cantos do mundo! "A miúda" sabe/sente bem o quanto tens sido importante! Rui, o número de horas que temos partilhado no

nosso magnífico 10 (além de muitas!) foi um dos maiores incentivos (senão o maior!) para chegar até este momento. Tens tido sempre a palavra certa, na hora certa! Muito obrigado pelo teu incondicional apoio!

Mi especial gratitud a mis compañeros Luz Estela, Carlos, Guillermo, Miguel y Zaida, por vuestra importante e inmensa ayuda, semana tras semana.

Uma palavra muito especial para os meus colegas, mas sobretudo amigos Gilberto, João Costa, Ângela e Ana Santos (serás sempre a Ana Santos!). Obrigado pela amizade, companhia e constante motivação. Gil, um dia ainda irás fazer (e explicar-me) tudo sobre um biplot!

Aos meus alunos e amigos… vocês fazem a diferença estatisticamente significativa!

Aos meus colegas e amigos da ESTM, Pestaninha (força, também estás quase!), Roberto Gamboa, Teresa Mouga, Maria José Rodrigues, Graça Ezequiel, Teresa Baptista, Sofia Eurico, Paulo Almeida, Rui Ganhão, João Vasconcelos, Pedro Gonçalves, Maria Manuel Gil, Francisco Domingos (sempre disponível!), Américo Rodrigues (chegou o momento da nossa Barca Velha!), Marco Lemos, Anabela Almeida, João Paulo Jorge… sei que ficará sempre alguém de fora… Obrigado pelo apoio e amizade que tornaram esta etapa mais fácil de alcançar.

À Célia, um enorme abraço… tu foste uma das responsáveis pela tua (eterna) caloirinha estar aqui a escrever estas palavras! Obrigado amiga… mesmo distante, sempre presente!

Aos meus pais, para os quais não tenho palavras para agradecer o inesgotável apoio e confiança. Acreditem que nem sempre é fácil demonstrar a gratidão que tenho por

nunca deixarem de estar ao meu lado. Se não tivesse sido assim, acreditem que não tinha sido possível chegar até aqui. Obrigado!

À Caua e à Xanda. Se cheguei aonde cheguei, a vocês o devo! Espero conseguir um dia agradecer-vos tanta paciência e compreensão. Vocês são muito importantes para mim… Obrigado!

Aos meus sobrinhos Tiago e Miguel, agradeço por ainda saberem quem eu sou, apesar das minhas inúmeras ausências durante todo este tempo. Aos meus ricos cunhadinhos, Loures e Coelhone, muito obrigado pela vossa presença, partilha e amizade.

Aos "Avós" Maranhões, ao Tio Matos e à Tia Fernanda, à Ana, ao Eduardo e ao Ricardo, vocês sabem o enorme e importante papel que têm tido para a minha felicidade! Nunca terei palavras para agradecer o vosso infinito apoio e carinho.

Aos dois besnicos, Guilherme e Pedro, a vossa presença, espontaneidade e alegria foram fundamentais!

Ao meu Buddy, que acreditou sempre que a "fito-estatística" tinha futuro. Sempre com a palavra certa, no momento em que estava a fazer falta! Obrigado por me fazeres ver a luz tantas e tantas vezes! Tem sido longo este percurso, mas tem sido muito feliz! Obrigado… esta tese também é tua!

Muito obrigado a todos os que de uma forma ou de outra contribuíram para realizar esta tese.

Agradezco a todas las personas que de una u otra forma contribuyeron a la realización de esta tesis doctoral.

Tudo farei para que o ponto final que aqui coloco seja agora um ponto de partida!

Este foi sem dúvida um mergulho mágico na minha vida!

INDEX

NOTATION

In order to make the description of the methods comparable to that used in three-way analysis, some adjustments to the standard notation were introduced.

X - two-way data array of dimensions (size e.g. $(I \times J)$)

x_{ij} - value of the measurement referring to the ith subject and jth variable

U - matrix of eigenvectors of the matrix XX^T

Λ - matrix of eigenvalues of the matrix XX^T

V^T - matrix of eigenvectors of the matrix X^TX

$\tilde{J}$ - markers for the columns of the matrix X

$\tilde{H}$ - markers for the rows of the matrix X

u_k - weighted average of k species with respect to any gradient x

y_{ik} - abundance of k species in site i $(i = 1, \dots, n; k = 1, \dots, p)$

x_i - value of gradient x at site i

y_{+k} - total abundances of k species

z_{ij} - value of the environmental variable j in site i

c_j - optimal weights $(j = 1, \dots, q)$

δ - weighted variance of species scores

$\mathbb{X}$ - three-way data array of dimensions $(I \times J \times K)$

x_{ijk} - value of the measurement referring to the ith subject, jth variable and kth condition

i – subjects elements; $i = 1, \dots, I$

j – variables elements; $j = 1, \dots, J$

k – conditions elements; $r = 1, \dots, K$

I - total number of elements in first mode (subjects)

J - total number of elements in second mode (variables)

K - total number of elements in third mode (conditions)

p - number of components selected to describe the first mode of the data array (subjects); $p = 1, \dots, P$

q - number of components selected to describe the second mode of the data array (variables); $q = 1, \dots, Q$

r - number of components selected to describe the third mode of the data array (conditions); $r = 1, \dots, R$

P - total number of components in first mode (subjects)

Q - total number of components in second mode (variables)

R - total number of components in third mode (conditions)

S - total number of components

A - component matrix of dimensions $(I \times P)$ with coefficients of the subjects (first mode)

B - component matrix of dimensions $(J \times Q)$ with coefficients of the variables (second mode)

C - component matrix of dimensions $(K \times R)$ with coefficients of the conditions (third mode)

a_{ip} - elements of the components matrices $A(I \times P)$

b_{jq} - elements of the components matrices $B(J \times Q)$

c_{kr} - elements of the components matrices $C(K \times R)$

$\mathbb{E}$ - three-way data error array of dimensions $(I \times J \times K)$

e_{ijk} - error term associated with the description of x_{ijk}

$\mathbb{G}$ - core array of dimensions $(P \times Q \times R)$

g_{pqr} - elements of the core array $\mathbb{G}$

$SS(TOT)$ - total sum of squares

$SS(FIT)$ - fitted sum of squares

$SS(RES)$ - residual sum of squares or deviance (d)

d - deviance

$d_{\mathbb{i}}$ - deviance of a model $\mathbb{i}$

$df_{\mathbb{i}}$ - degrees of freedom of model $\mathbb{i}$

$\tilde{x}_{ijk}$ - profile preprocessed element referring to the ith subject, jth variable and kth condition

b_s - salience value

$st_{\mathbb{i}}$ - st-criterion of a of model $\mathbb{i}$

$\hat{x}_{ijk}$ - structural image element referring to the ith subject, jth variable and kth condition

$G_1, \ldots, G_p$ - horizontal core slice of core array $\mathbb{G}$

$G_1, \ldots, G_q$ - lateral core slice of core array $\mathbb{G}$

$G_1, \ldots, G_r$ - frontal core slice of core array $\mathbb{G}$

f_{pjk} - component score of individual k at occasion j on component p of the first mode (in interactive biplots)

$\mathcal{O}_k$ - matrix of operators between variables

D_I - diagonal matrix of dimensions $(I \times I)$

D_J - diagonal matrix of dimensions $(J \times J)$

X_w - compromise table

$\mathcal{U}$ - matrix of the eigenvectors of the analysis of the compromise

$\mathcal{R}_k$ - coordinates of the rows of the table X_k

$\mathcal{C}_k$ - coordinates of the columns of the table X_k

W_k - kth cross product table

INTRODUCTION

INTRODUCTION

In standard multivariate analysis, data are arranged in a two-way structure: a table or a matrix. A typical example is a table in which each row corresponds to a sample (or object) and each column to a measured variable. The two-way structure explicitly implies that for every sample the variable is measure and vice-versa. Thus, the data can be indexed by two indices: one defining the sample number and one defining the variable number. This arrangement is closely connected to the techniques subsequently used for analysis of the data (as Principal Component Analysis, for example). However, for a wide variety of data a more appropriate structure would be a three-way table or an array. An example could be a situation where for every sample the variables are considered at several different conditions. In this case every data element can be logically indexed by three indices: one identifying the sample number, one the measured variable and one the observed condition. Clearly, three-way data occur frequently, but are often not recognized as such due to lack of awareness. Moreover, researchers generally become interested in three-way analysis because they already have three-way data and they need to use multi-way methods to obtain and understand their results.

The problems described here are mostly stem from the ecological experiments. This is not coincidental as the data analytical problems arising in the ecological area can be complex. Thus, a typical multivariate data set in ecology usually has counts of several species (columns, say) at each of several sites or in each of several observation units (rows) (for example, Orlóci, 1975, Gauch, 1982, Digby and Kempton, 1987, Jongman et al., 1987, Legendre and Legendre, 1998). Hence, the nature of the data arising from this area can be very different, which tends to complicate the data analysis.

On the other hand, ecology is undergoing a profound change because structure–function relationships between communities and their environment are starting to be investigated at the field, regional, and even continental scales. This opened a wide-ranging challenge when interpreting the observed diversity patterns as a function of contextual environmental parameters. At the same time, the development of new analytical techniques, powerful mathematical and statistical methods as well as by the increasing power and storage capacity of personal computers (which are now capable of handling sets of data and of applying algorithms that a few years ago were limited) contributed for a great impetus of these two branches of science. One of the main consequences of such developments has been the constant increase of size and complexity of data sets (and which can produce matrices of data for each sample analyzed or even three-way arrays). Furthermore, when dealing with data analysis in the ecological area is important to have access to a diverse set of methodologies, in order to deal with problems in a sensible way.

The analytical problem most often involves the investigation of interactions or effects in natural systems on the (species) communities of a particular area. Such ecological assemblages consist of many interacting species, and each species is generally treated as a separate response variable in the analysis. Thus, the main objective in these studies is to obtain information by means of species-environment relationships, which gives an understanding of what are the parameters that influence the community composition and its structure.

Most researchers (in Ecology, but not only) solve this problem through the use of Canonical Correspondence Analysis (CCA, ter Braak, 1986). However, it is important to

relay the point that temporal and spatial variability are in general present in these ecological studies and as consequence ecological sampling programs or experimental designs need to take this into account. Yet, one of the problems that arise in methods such as CCA is that the variability over tables (or matrices which are usually times or sites) is mixed with the variability within the tables. Therefore, by this technique (widely used) the study of spatial and temporal gradients is not carried out optimally and often fails to find the underlying structures in multi-way arrays.

The methods of analysis for three-way data, which studying the (three) dimensions, objects, variables and events, are certainly an alternative and whose interest by researchers is gaining some relevance. However, its use is still minority compared to the CCA, particularly given the difficulty associated with the use of algebraic language which is implicit. From mathematical point of view this is one practical problem, when handling multi-way data. Besides that, there is also the question of notation that is often used by the authors: frequently use special symbols to simplify the notation, which sometimes makes the work difficult to read.

However, multi-way analysis is the natural extension of standard multivariate analysis, when data are arranged in three-(or higher) way arrays. This in itself provides a justification for the use of multi-way methods. Moreover, the power not only to separate the effect of three dimensions, but to evaluate the effect of the interaction between them in the dynamics of the species-environment also justifies the interest in this field. Actuality, multi-way methods provide a logical and advantageous tool in many different problems. Besides, the exploration and development of multi-way models makes it

possible to obtain information that more adequately describes the intrinsic and complex reality.

This research focuses on the comparative study of various methods of three-way data analysis. At the same time, emphasis is given to the several aspects in which comparisons between the methods are based, as well as its applicability to analyze data from ecological studies in time and space.

The most common methods are the Triadic Analysis (Jaffrenou, 1978; Thioulouse and Chessel, 1987) and Tucker models (Tucker, 1966). The first seeks a consensus configuration over time or space and therefore aims to evaluate the stable part in the dynamics of relationships. In the second, the models are adjusted to achieve the goal of explaining the interaction.

This work aims to produce a general methodological framework, including different techniques for studying multi-way matrices, and develop a specific technique to facilitate the use of these practices by researchers.

The work described in this thesis is concerned to meet the following objectives:

(1) Give a general outline of the most popular methods for analyzing data when they are organized into one or two arrays of data;

(2) Make a revision of the methods from the French school (specifically, the STATIS family methods), that is, when data are organized as data cubes;

(3) Performing a description of the models from school Dutch/English (specifically, the Tucker3 model) for the treatment of three-way tables/three modes;

(4) Perform an exhaustive search of the literature in the field of ecology to understand the impact of these methods in this field;

(5) Propose an new method that really solve the problem of describing not only the stable part of the dynamics of structure–function relationships between communities and their environment (in different locations and/or at different times), but also the interactions and changes associated with the ecosystems' dynamics;

(6) Apply the new proposal to a real data set collected in arrays of three ways (sites × environmental variables/species × time or time × environmental variables/species × sites) and compare the results with those found with classical techniques.

The data analytical techniques covered in this thesis are also applicable in many other areas, as evidenced by many works of applications in other areas which are emerging in the literature.

CONTENT OUTLINE

CONTENT OUTLINE

The main part of the thesis is divided into six chapters.

CHAPTER 1 initiates with a briefly review and discussion of some issues related to the use of indirect gradient analysis, namely, Correspondence Analysis. Next, a brief introduction to the biplot theory, namely the variant of the standard biplot technique, the HJ-Biplot will be explored. It will be demonstrated that HJ-Biplot is a powerful tool to find hidden patterns in a data matrix. Useful characteristics like the guarantee of providing high quality of representation for both lines as for the columns are outlined and also discussed. In order to better understand all the interpretation under the use of HJ-Biplot, a comprehensive example will be presented.

A published application with bacterioplankton data (*"Bacterioplankton dynamics in the Berlengas Archipelago (West coast of Portugal) using the HJ-biplot method"*) on the international peer-reviewed journal, ARQUIPELAGO LIFE AND MARINE SCIENCES, will be provided in the appendix section **PUBLISHED ARTICLES BY THE AUTHOR** (Paper I).

CHAPTER 2 describes one of the most successful methods for studying species-environment relationship, i.e. Canonical Correspondence Analysis. Additionally, the main advantages and drawbacks will be described. Also, a concise description of the Co-Inertia Analysis will be presented, in view of the fact that is a simple and robust alternative to Canonical Correspondence Analysis when the number of samples is low compared to the number of explanatory (environmental) variables.

In order to better understand all the interpretation under the use of these methods, comprehensive examples will be presented.

CHAPTER 3 begins with a revision of data coupling methods, namely within in STATIS family methods (French school). The principals of alignment algorithms of Partial Triadic Analysis and STATICO will be introduced, since their methodological bases and objectives are related to the new method proposed in **CHAPTER 5**.

Related to Partial Triadic Analysis, published applications with plankton data on the international peer-reviewed journals BIOMETRICAL LETTERS (*"The efficiency of Partial Triadic Analysis method: an ecological application"*) and FRESENIUS ENVIRONMENTAL BULLETIN (*"Zooplankton distribution in a Marine Protected Area: the Berlengas Natural Reserve (Western coast of Portugal"*) are presented in the appendix section **PUBLISHED ARTICLES BY THE AUTHOR** (Papers II and III, respectevely).

On the other hand, and associated to STATICO, published applications with plankton data on the international peer-reviewed journals ESTUARINE, COASTAL AND SHELF SCIENCE (*"Spatio-temporal structure of diatom assemblages in a temperate estuary"* and *"Diel vertical behavior of Copepoda community (naupliar, copepodites and adults) at the boundary of a temperate estuary and coastal waters"*), ZOOLOGICAL STUDIES (*"Response to Climatic variability of Copepoda life history stages in a southern European temperate estuary"*) and JOURNAL OF APPLIED STATISTICS (*"An empirical comparison of Canonical Correspondence Analysis and STATICO in identification of spatio-temporal ecological relationships"*) are presented in the appendix section **PUBLISHED ARTICLES BY THE AUTHOR** (Papers IV, VI, VII and V, respectively).

CHAPTER 4 begins with a revision of three-way methods with special emphasis to the terminology that characterize these techniques. General considerations about the Tucker3 model (English/Dutch school) for the treatment of three-way tables/three modes will be presented.

CHAPTER 5 presents a new proposal, the CO-TUCKER. The new method is a combined approach of STATICO and Tucker3 techniques. It is an alternative method which aims to solve the problem of describing not only the stable part of the dynamics of structure–function relationships between communities and their environment (for several conditions such as different sites and/or different times), but also the interactions and changes associated with the ecosystems dynamics'. Throughout the chapter, special emphasis will be given to the advantages of CO-TUCKER model with regard to the possibility of extracting hidden structures and capture the underlying correlations between variables in a multi-array. In addition, aspects related to the data preprocessing, strategies in the selection of components and the quality assessment of the fit of the levels of a mode will be highlighted. Particular attention is made for the core array and its suppleness in order to explain an interaction between each component with any component in the other modes. Also, joint biplots will be explored with the aim of capture multilinear structures in the (higher-order) datasets.

The CO-TUCKER technique will be applied on a real data set. The results and the inherent advantages will be presented and discussed. Also, comparative considerations, advantages and drawbacks are explored and discussed with STATICO technique.

Finally, **CHAPTER 6** presents the conclusions and perspectives of future work.

CHAPTER 1

CHAPTER 1 – THE ANALYSIS OF A MULTIVARIATE DATA MATRIX

1.1 PRINCIPAL COMPONENT ANALYSIS AND CORRESPONDENCE ANALYSIS

Statistical applications in ecology are numerous and include many of the classical and modern methods, as well as some novel methodology.

Researchers in ecology commonly use multivariate analysis to interpret patterns in ecological data and relate these patterns to environmental stressors.

Methods are classified as being direct or indirect. When a direct method is used, the data consist of species information and measurement of the stressors. The direct method relates these two sets of information. When an indirect method is used, the stressors are typically unknown and the analysis is used to produce the stress gradients from data on abundance.

In the indirect perspective, two methods that have found common use in the early analysis of one ecological data matrix (Fig. 1) were Principal Components Analysis (PCA, Hotelling, 1933) and Correspondence Analysis (CA, Benzécri, 1973).

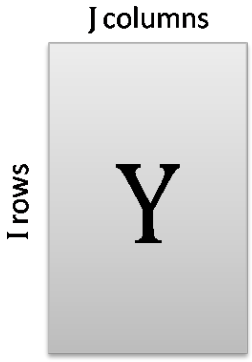


Figure 1. Y two-way data matrix of dimensions ($I \times J$)

The PCA is usually applied to reduce data (usually a matrix of continuous variables) on numerous species into several components analysis that summarize the primary information in the data.

One difficulty with using PCA is that it works best when the relationships between species and the environment are linear. Unfortunately, many species environment relationships are nonlinear. The relationship between abundance and the gradient is typically modeled as a Gaussian curve rather than a linear relationship (Fig. 2).

Another limitation arises since PCA does not allow the jointly representation of rows (for example, sites) and columns (for example, species) in the same diagram.

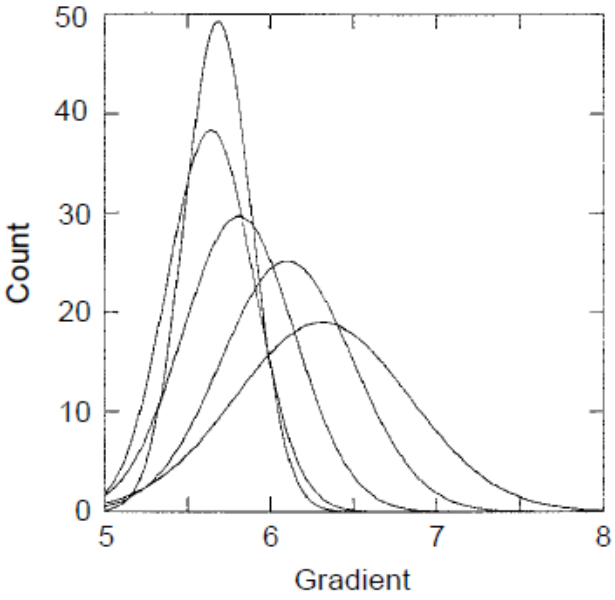


Figure 2. Plot of hypothetical species abundance curves along a gradient for line species (adapted from Smith (2002))

On the other hand, the CA has found extensive use in ecology as a method for visualizing the patterns of association in a table of frequencies or nonnegative count data (for example, species counts). Inherent to the method is the expression of the data in each row or each column relative to their respective totals, and it is these sets of relative values (called profiles) that are visualized. CA is routinely applied as a tool for the analysis of multivariate species count data, for example a matrix of species counts at different locations, in order to obtain an ordination diagram of species and sites (i.e., a spatial diagram of the important dimensions in the data) (Fig. 3).

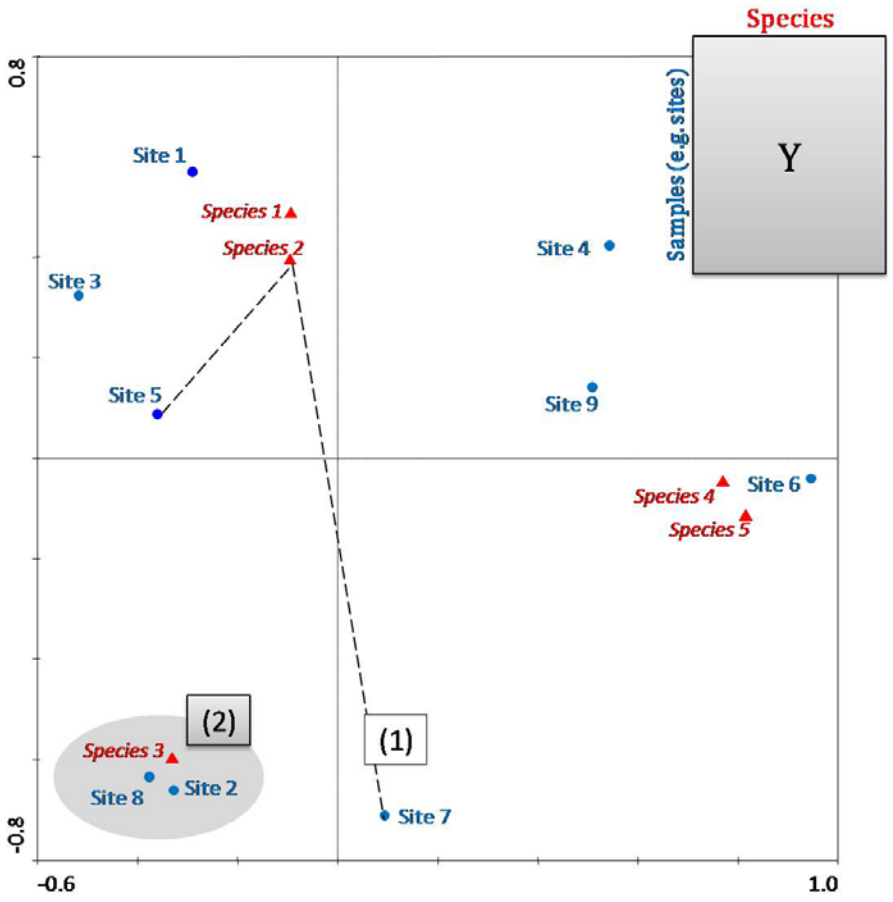


Figure 3. Ordination diagram of hypothetical species and sites samples. The distance between a site point and species points approximates the (predicted) relative frequency (or probability of occurrence) of species in that site (see example [1]). Similarly, sites where a particular species has higher frequency tend to have their symbols clustered around that species point (see example [2]). Sites are represented as Site 1 to Site 9 and species as *Species 1* to *Species 5*.

Such ordination diagrams were initially made by ecologists using reciprocal averaging algorithm, which is equivalent to CA (Hill, 1974).

In geometrical terms, the profile vectors form a cloud of points in high-dimensional space and the objective is to fit an optimal diagram that best approximates this cloud in a least squares sense. Thus, proximities between points and other geometric features of the diagram indicate associations between rows, between columns and between rows and columns (Fig. 3).

This relativization of the data makes perfect sense when the margins of the table represent samples from subpopulations of inherently different sizes. But in some ecological applications sampling is performed on equal areas or equal volumes so that the absolute levels of the observed occurrences may be of relevance, in which case relativization may not be required (Greenacre, 2010).

These sorts of relationships and sampling designs are expected in ecological work.

However, when the data are not given in the form of nonnegative count data or the objective is a joint representation of rows and columns of a data matrix, the biplots methods appear as a good solution to analyze one data matrix.

This chapter presents a brief account of the biplot theory, namely the variant of the standard biplot technique, the HJ-Biplot (Galindo 1986), following (for the most part) notation of Gabriel (1971) and Galindo (1986, 1985).

A published application (with bacterioplankton data) on the international peer-reviewed journal, ARQUIPELAGO LIFE AND MARINE SCIENCES, is provided in the appendix section **PUBLISHED ARTICLES BY THE AUTHOR** (Paper I).

1.2 HJ-BIPLOT

Biplots can be viewed as the multivariate analogue of scatter plots, where samples (or cases) are plotted as points on two variables. This is an exploratory data analysis method that looks for hidden patterns in a data matrix[1]. Biplots were introduced by Gabriel (1971) in the context of PCA (Hotelling, 1933).

The fundamental characteristic that makes differentiation of the diverse graphical representations associated with classical methods of dimensionality reduction is the fact that in a biplot, is possible to obtain a joint representation of rows and columns of a data matrix. Hence, with biplots, the multivariate distribution of a set of variables can be approximated in a low dimensional space (usually two dimensions) giving a useful visualization of the entire structure.

From the algebraic point of view, the biplot is based on the same principles that support most of the factorial techniques of dimensionality reduction, i.e., uses the singular value decomposition (SVD)[2]. By this technique it is possible find a good representation in low-dimensional space. Moreover, SVD provides the coordinates on dimensions (or directions in space), which are usually called (in mathematical sense) singular vectors.

[1] Throughout this work, and where the context so requires, the terms "matrix" and "table" will be used as having the same meaning.

[2] The singular value decomposition (SVD) consists in the decomposition of a matrix X into a left set of orthonormal vectors U, a right set of orthornormal vectors V, and a diagonal matrix Λ of singular values, which are the square roots of the eigenvalues of both $X^T X$ and XX^T; thus $X = U\Lambda \mathrm{V}^{\mathrm{T}}$.

The dimensions are arranged in such a way that they are orthogonal[3] (i.e., at right angles), and successively represent as much of the variation as possible. In addition, the technique provides measurements (singular values), which, if squared, indicate the amount of variability explained by each dimension (Kroonenberg, 2008). So, to display the main variability in a two-dimensional graph, the first two dimensions should be used.

The fundamental difference is that in this case the data is reproduced and then incorporated in a joint representation of rows and columns.

The two most important biplot factorizations proposed by Gabriel (1971) were designated: GH-Biplot and JK-Biplot. The GH-Biplot gets high quality representation for the J columns (variables) and not so high for the I rows (objects[4]), while the JK-Biplot achieves high quality of representation for rows, and not so high for the columns.

Using an appropriate choice of markers for the rows and columns, Galindo (1986, 1985) proposed a variant of the standard biplot techniques, the HJ-Biplot, which enables simultaneous representation on the same coordinate system. Additionally, Galindo (1986, 1985) has shown that this alternative ensures high quality of representation for both rows as for the columns.

[3] Two vectors are orthogonal if they are perpendicular to each other irrespective of their lengths; a $I \times J$ matrix X is (column-wise) orthogonal if $X^T X = \Lambda_J$, where Λ is a $I \times J$ diagonal matrix with arbitrary nonnegative numbers.

[4] Throughout this work, the terms “subject” and “object” should be read as having the same meaning.

In fact, this is the core advantage of this ordination method providing better results than other conventional multivariate techniques (Galindo, 1986; Galindo and Cuadras, 1986). HJ-Biplot is applied in several investigations (see for example Rivas-Gonzalo et al., 1993; Santos et al., 1991; Martínez-Ruiz et al., 2001; Martínez-Ruiz and Fernández-Santos, 2005; Mendes et al., 2009a). Thus, the theoretical background for the HJ-Biplot technique is briefly presented.

The HJ-Biplot is a symmetric simultaneous representation technique that to a certain extent resembles CA (Benzécri, 1973), but is not restricted to frequency data. It is a joint representation, in a low dimensional vector space (usually, in two dimensions), of the rows (I) and columns (J) of the two-way data matrix X, using markers (points/vectors), for its rows ($\tilde{a}_1, \dots, \tilde{a}_I$) and for its columns ($\tilde{b}_1, \dots, \tilde{b}_J$). The markers are obtained from the usual SVD of the data matrix.

The initial matrix $X(I \times J)$ can be written according to the SVD defined as,

$$X = U\Lambda \mathrm{V}^{\mathrm{T}} \tag{1.2.1}$$

where U is the matrix of eigenvectors of the matrix XX^T; Λ is the matrix of eigenvalues of the previous matrix arranged in decreasing order of magnitude, and V^T is the matrix of eigenvectors of the matrix X^TX. Suitable choice of markers for the rows (I) and for the columns (J) in a low-dimension subspace affords the biplot representation.

For the case of HJ-Biplot, the following factorization is chosen,

$$\tilde{J} = U\Lambda \tag{1.2.2}$$

and

$$\widetilde{H} = V\Lambda \tag{1.2.3}$$

where $\tilde{J}$ are the markers for the columns of the matrix X, and $\widetilde{H}$ the markers for the rows of the matrix X.

Both markers can be represented in the same reference system. Indeed, taking into account the relationship between U and V, that is, $U = XV\Lambda^{-1}$:

$$U\Lambda = (XV\Lambda^{-1})\Lambda = XV \tag{1.2.4}$$

The markers chosen for the rows coincide with the projection of the I points in the space of the principal components of the columns.

Similarly, $V = X^T U\Lambda^{-1}$, and hence:

$$V\Lambda = (X^T U\Lambda^{-1})\Lambda = X^T U \tag{1.2.5}$$

That is, the markers chosen for the columns coincide with the projection of the J points in the space of the principal components of the rows.

Both representations are related. Let be, $\tilde{A} = XV$ and $\tilde{B} = X^T U$, then:

$$\tilde{B} = X^T U = X^T XV\Lambda^{-1} = X^T \tilde{A}\Lambda^{-1} \tag{1.2.6}$$

and

$$\tilde{A} = XV = XX^T U\Lambda^{-1} = X\tilde{B}\Lambda^{-1} \tag{1.2.7}$$

Once has that the hth coordinate of the jth column can be expressed as a function of the hth coordinates of the I-rows.

The goodness of fit for the first two dimensions of HJ-Biplot is achieved by

$$\frac{\lambda_1 + \lambda_2}{\sum_s \lambda_s} \tag{1.2.8}$$

where λ_s are the eigenvalues.

And so, the HJ-Biplot (as mentioned previously), achieves an optimum quality of representation for both rows and columns, since both are represented on the same reference system. Besides, it is closely related to PCA, since variance and covariance matrices are plotted on a plan that account for most of the inertia (Galindo, 1986; Galindo and Cuadras, 1986).

The rules for the interpretation of the HJ-Biplot are a combination of the rules used in other multidimensional scaling techniques (as CA, Factor Analysis and classical biplots) and are based on very simple geometric concepts. Below those basic interpretational rules are presented (for details see Section 1.2.2, Galindo (1986) and Mendes et al. (2009a)):

– the distances among row markers are interpreted as an inverse function of similarities, in such a way that closer markers (objects) are more similar. This property allows for the identification of clusters of objects with similar profiles;

- the lengths of the column markers (vectors) approximate the standard deviation of the variables;

- the cosines of the angles among the column vectors approximate the correlations among variables in such a way that small acute angles are associated with variables with high positive correlations; obtuse angles near to the straight angle are associated with variables with high negative correlations and right angles are associated with non-correlated variables. In the same way, the cosines of the angles among the variable markers and the axes (Principal Components) approximate the correlations between them. For standardized data, these approximate the factor loadings[5] in factor analysis;

- the order of the orthogonal projections of the row markers (points) onto a column marker (vector) approximates the order of the row elements (values) in that column. The larger the projection of an individual point onto a variable vector is, the more this individual deviates from the average of that variable.

Several measurements are essential for a correct HJ-Biplot interpretation (Galindo, 1986; Galindo and Cuadras, 1986).

For instance,

- the absolute contribution of the column element *j*th (variable) to the factor *s*th,

[5] Throughout this work, the terms "loadings" and "scores" should be read as having the same meaning.

$$CAE_j F_s = \tilde{b}_{js}^2 \tag{1.2.9}$$

where $\tilde{b}_{js}$ denotes the projection of the variable j in the sth dimension (Fig. 4);

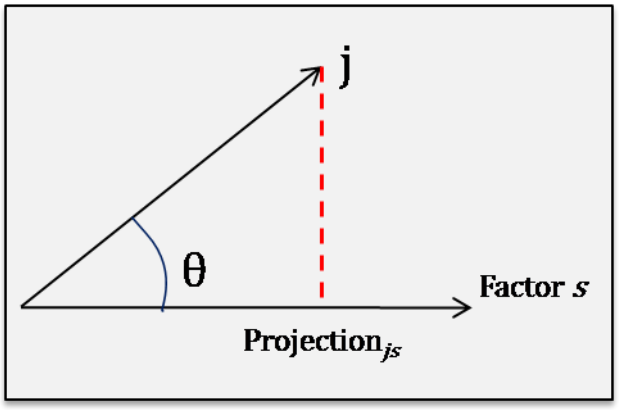


Figure 4. Geometric representation of the projection of the variable j in the sth dimension.

– the relative contribution of the column element jth (variables) to the factor sth,

$$CRE_j F_s = \frac{\tilde{b}_{js}^2}{\lambda_s} \tag{1.2.10}$$

which expresses the importance of the variable to explain a factor or dimension;

– the relative contribution of the factor sth to the column element jth (variables),

$$CRF_s E_j = \frac{\tilde{b}_{js}^2}{\sum_s \tilde{b}_{js}^2} \tag{1.2.11}$$

which expresses the relative variability of the variable explained by a factor or dimension.

By analogy, for the row elements,

– the absolute contribution of the row element ith (objects) to the factor sth,

$$CAE_iF_s = \tilde{a}_{is}^2 \tag{1.2.12}$$

where $\tilde{a}_{js}$ denotes the projection of the row i in the sth dimension (Fig. 5);

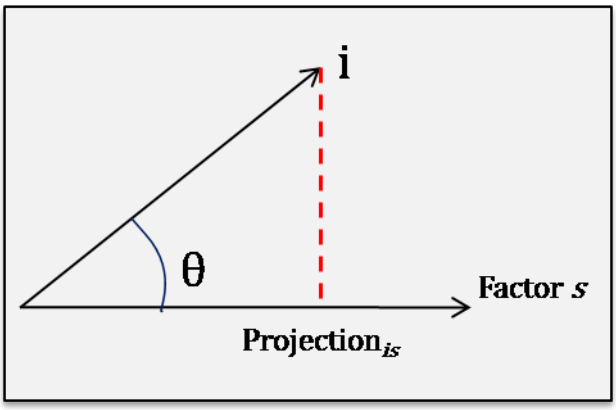


Figure 5. Geometric representation of the projection of the row i in the sth dimension.

– the relative contribution of the row element ith (objects) to the factor sth,

$$CRE_iF_s = \frac{\tilde{a}_{is}^2}{\lambda_s} \tag{1.2.13}$$

which expresses the importance of the row to explain a factor or dimension;

– the relative contribution of the factor s to the row element i (objects),

$$CRF_sE_i = \frac{\tilde{a}_{is}^2}{\sum_s \tilde{a}_{is}^2} \tag{1.2.14}$$

which expresses the relative variability of the row explained by a factor or dimension.

Also, for a point (row or column) on a factorial plan, the quality of representation (QR) can be defined by adding the two relative contributions of these factors to the element, i.e,

– for the column element j,

$$QRE_{j_{(1-2)}} = CRF_1E_j + CRF_2E_j = \cos^2\theta_{j(1-2)} \quad (1.2.15)$$

– for the row element i,

$$QRE_{i_{(1-2)}} = CRF_1E_i + CRF_2E_i = \cos^2\theta_{i(1-2)} \quad (1.2.16)$$

Only points with a high quality of representation can be properly interpreted (Galindo and Cuadras, 1986).

1.2.1 VISUALIZING HJ-BIPLOT RESULTS

In the appendix section **PUBLISHED ARTICLES BY THE AUTHOR** (Paper I), an application will be shown illustrating many of the aspects covered in this chapter.

The work was already published on an international peer-reviewed journal, namely,

> ***Mendes S****, Fernández-Gómez MJ, Galindo-Villardón MP, Morgado F, Maranhão P, Azeiteiro UM, Bacelar-Nicolau P. 2009. Bacterioplankton dynamics in the Berlengas Archipelago (West coast of Portugal) using the HJ-biplot method. Arquipélago Life and Marine Sciences 26: 25-35.*

It is the hope, that the problem described here will enable researchers to generalize also to other types of data and problems than those actually shown. A briefly resume of the work is here expressed.

ABSTRACT

"The relationship between bacterioplankton and environmental forcing in the Berlengas Archipelago (Western Coast of Portugal) were studied between February 2006 and February 2007 in two sampling stations: Berlenga and Canal, using an HJ-biplot. The HJ-biplot showed a simultaneous display of the three main metabolic groups of bacteria involved in carbon cycling (aerobic heterotrophic bacteria, sulphate-reducing bacteria and nitrate-reducing bacteria) and environmental parameters, in low dimensions. Our results indicated that bacterial dynamics are mainly affected by temporal gradients (seasonal gradients with a clear winter versus summer opposition), and less by the spatial structure (Berlenga and Canal). The yearly variation in the abundance of aerobic heterotrophic bacteria were positively correlated with those in chlorophyll a concentration, whereas ammonium concentration and temperature decreased with increasing phosphates and nitrites concentration. The relationship between aerobic heterotrophic bacteria, chlorophyll a and ammonium reveals that phytoplankton is an important source of organic substrates for bacteria."

KEYWORDS: *Berlengas Natural Reserve, aerobic heterotrophic bacteria, sulphate-reducing bacteria, nitrate-reducing bacteria.*

CHAPTER 2

CHAPTER 2 – THE ANALYSIS OF THE RELATIONSHIP BETWEEN TWO MULTIVARIATE DATA MATRICES

2.1 INTRODUCTION

In the previous chapter, methods that analyze one data matrix were pointed out. In particular, HJ-Biplot was performed to get an optimal picture of the data.

However, the study of the relationship between species and its environment usually leads to two sets of data: (i) a species array that contains the abundance or the occurrence of a number of *taxa* in a set of sites; and (ii) an environmental array that includes quantitative or categorical measurements from the same sites.

In this sense, a direct gradient analysis comes out as a way to relate species abundance data to environmental variables.

Two methods that have found common use for the evaluation of relationships between ecological data and environmental variables are Redundancy Analysis (RDA, Legendre and Legendre, 1998) and Canonical Correspondence Analysis (CCA, ter Braak, 1986).

CCA is a variation of CA: CA derives combinations of the species that explain distance and CCA relates combinations of species to environmental variables. The method is related to Canonical Correlation Analysis (CANCOR, Hotelling, 1936; ter Braak and Looman, 1986). CANCOR is commonly used to relate one set of variables to another by

forming combinations of variables. The combinations are linear combinations of variables with weight chosen to maximize the correlation between the linear combinations.

RDA is similar to CCA, but it is a constrained PCA instead of a constrained CA. Thus, RDA is similar to PCA in that linear combinations of the variables (species) are formed. The new axes are formed not to maximize the variance of the species abundance data, but are constrained to be linear combinations of the environmental variables.

CCA is a form of direct gradient analysis and has become an important multivariate technique in ecology (Palmer, 1993). It uses the environmental information simultaneously with the species data when theoretical gradients are extracted. One of the mainstays of the method is the assumption of a nonlinear relationship between (linearity combined) environmental axes and species abundance, known as the unimodal response model (ter Braak, 1985).

On the other hand, if two tables are linked by the same individuals, one can find a structure, a co-structure to study the relationship between the two set of variables. Thus, Co-Inertia Analysis (Dolédec and Chessel, 1994) is a simple and robust alternative to CCA when the number of samples is low compared to the number of explanatory (environmental) variables. Co-Inertia Analysis can be seen as the PCA of the table of cross-covariances between the variables of the two tables.

The purpose of this chapter is to present a summary of CCA (parting from the point of view of ter Braak and Verdonschot (1995) and ter Braak (1987)), followed by a concise description of the Co-Inertia Analysis.

2.2 CANONICAL CORRESPONDENCE ANALYSIS

Canonical Correspondence Analysis (CCA) (ter Braak, 1986) is the standard multivariate method (and often regarded as fundamental) in the field of Ecology. Commonly used to elucidate the relationships between two pair of tables, namely to relate the biological assemblages of species to their environment.

The method is designed to extract synthetic environmental gradients from ecological data sets. CCA operates on (field) data on abundances (usually counts of individuals of species) and data on the explanatory variables (usually environmental variables) at several occurrences (usually sites) (ter Braak and Verdonschot, 1995). CCA stems directly from the classical ordination method of CA (Benzécri, 1973); in this sense CCA can be seen as a constrained CA. In fact, CCA selects ordination axes as in CA but imposing that they are a linear combination of explanatory variables. CCA chooses then the best weights for those explanatory variables so as to maximize the dispersion of abundances scores. CCA is therefore restricted to the case where explanatory variables are linearly independent and not too many. From the point of view of statistical methods, CCA can be designated as having a predictive strategy (Thioulouse, 2010), which means that is oriented towards the prediction of "explained" (i.e., dependent) variables by "explanatory" (i.e., or independent) ones. Thus, as the CCA involves a regression step, this means that has the matrix inversion step. This matrix inversion step means that "explanatory" variables must be independent (in the statistical sense), and that is to say, they must be linearly independent (since the rank of their correlation matrix must not be less than its dimension and this implies that the number of samples (sites) can't be less that the number of explanatory variables) (Thioulouse, 2010).

Moreover, when the number of explanatory variables included in the analyses is approximately equal to the number of samples collected, CCA is equivalent to CA.

Being a restriction of CA, CCA shares properties and drawbacks with CA but some drawbacks can be more easily solved in CCA than in CA (Gauch, 1982; Jongman et al., 1987). The statistical model underlying CCA is that data abundance or frequency is a unimodal function of its position along explanatory variables gradients. CCA is an approximation to Gaussian regression under a certain set of simplifying assumptions, and is robust to violations of those assumptions (ter Braak and Prentice, 1988). CCA is inappropriate for extremely short gradients, in which data abundance or frequency is a linear or monotonic function of gradients (ter Braak and Prentice, 1988; ter Braak, 1987).

From a more theoretical point of view, it is possible to arrive at the basic equations of CCA from different perspectives, most of them described in the literature. For instance, CCA is a maximization of the dispersion of the species scores using a linear restriction on the site scores (Johnson and Altman, 1999; ter Braak, 1987).

Alternatively, CCA has been stated to be weighted least squares approximation to the weighted averages of the species with respect to the environmental variables (ter Braak, 1986). It is possible to put out CCA in the framework of reciprocal averaging, where the reciprocal averaging algorithm is combined with the regression of site scores onto environmental variables.

CCA has also been formulated as a weighted principal component analysis of a matrix of weighted averages (ter Braak, 1987).

The purpose here is to give a description of CCA following the viewpoint of ter Braak (1987) and ter Braak and Verdonschot (1995), who described the method as being a combination of CA and multiple regression. It is also important to stress that CCA is characterized by having a dual purpose; i.e., CCA aims to find the agreement between the typology of sites from the standpoint of the matrix structure of community composition (i.e., from the point of view of content in terms of species), and the typology of sites from the point of view of the structure of the environmental matrix (i.e., from the standpoint of environmental value that occur in those sites).

Given the importance of this method in the field of Ecology, its description will be made from this point of view, i.e., from the point of view of species-environment relationship. For this reason it is decided here to keep the notation usually followed to describe this method.

Let, in the literature,

$$u_k = \sum_{i=1}^{n} \frac{y_{ik}}{y_{+k}} x_i \qquad (2.2.1)$$

the weighted average of k-th species with respect to any gradient x (environmental variable, synthetic gradient or ordination axis). In other words, it is the weighted average of the gradient values of the sites at which the species occurs. Note that y_{ik} is

the abundance of k-th species at site i $(i = 1, \dots, n$ and $k = 1, \dots, p)$, x_i the value of gradient x at site i, and

$$y_{+k} = \sum_{i=1}^{n} y_{ik} \tag{2.2.2}$$

the total abundance of k-th species. For a standardized gradient x, i.e. a gradient for which

$$\sum_{i=1}^{n} \frac{y_{i+}}{y_{++}} x_i = 0 \tag{2.2.3}$$

and

$$\sum_{i=1}^{n} \frac{y_{i+}}{y_{++}} x_i^2 = 0 \tag{2.2.4}$$

the weighted variance (dispersion) of species scores u_k $(k = 1, \dots, p)$ is defined by

$$\delta = \sum_{k=1}^{p} \frac{y_{+k}}{y_{++}} u_k^2 \tag{2.2.5}$$

Let x be a synthetic gradient, i.e., a linear combination of environmental variables

$$x_i = c_0 + c_1 z_{i1} + c_2 z_{i2} + \cdots + + c_q z_{iq} \tag{2.2.6}$$

where z_{ij} is the value of the environmental variable j at site i and c_j $(j = 1, \dots, q)$ the coefficient or weight.

Then, CCA is the method that chooses the optimal weights c_j, i.e., the weights that result in a gradient x for which the weighted variance of the species scores, δ, is maximum. Mathematically, the synthetic gradient x can be obtained by solving an eigenvalue

problem; x is the first eigenvector x_1 with the maximum eigenvalue δ (ter Braak, 1987). The optimized weights (c_j) are termed canonical coefficients. Each subsequent eigenvector $x_s = (x_{1s}, \dots, x_{ns})^T (s > 1)$ maximizes δ subject to constraint

$$x_i = \sum_{j=1}^{q} c_q \, z_{ij} \tag{2.2.7}$$

and the extra constraint that it is uncorrelated with previous eigenvectors, i.e.,

$$\sum_{i=1}^{n} y_{i+} \, x_{is'} x_{is'} = 0 \; (s' < s) \tag{2.2.8}$$

In practice CCA is used to detect, interpret and predict the underlying structure of the data set based on the explanatory variables (for example, environmental variables). Furthermore, analyzing the CCA ordination diagram obtained from the most popular package (the CANOCO software) (ter Braak and Smilauer, 2002), it shows more accurately the dissimilarities between the patterns of occurrence of different species (i.e., the focus scaling is on inter-species distances). Moreover, the scaling type is by biplot scaling meaning that the interpretation is made by the biplot rule. These two options (inter-species distances and biplot scaling) lead to the χ^2 distance as a measure of dissimilarity.

2.2.1 VISUALISING CANONICAL CORRESPONDENCE ANALYSIS RESULTS

The graphical results obtained by the CCA provide an ordination diagram with species, environmental variables and sites. In this diagram the variation patterns of species community, which are better explained by the environmental variables are revealed. Species and sites appear as points, while environmental variables are represented by vectors.

In order to illustrate the interpretation of the results obtained by CCA (following the biplot scaling type), a hypothetical example is showed (Fig. 6).

The individual interpretation of species points and environmental variables, as well as their combinations are presented.

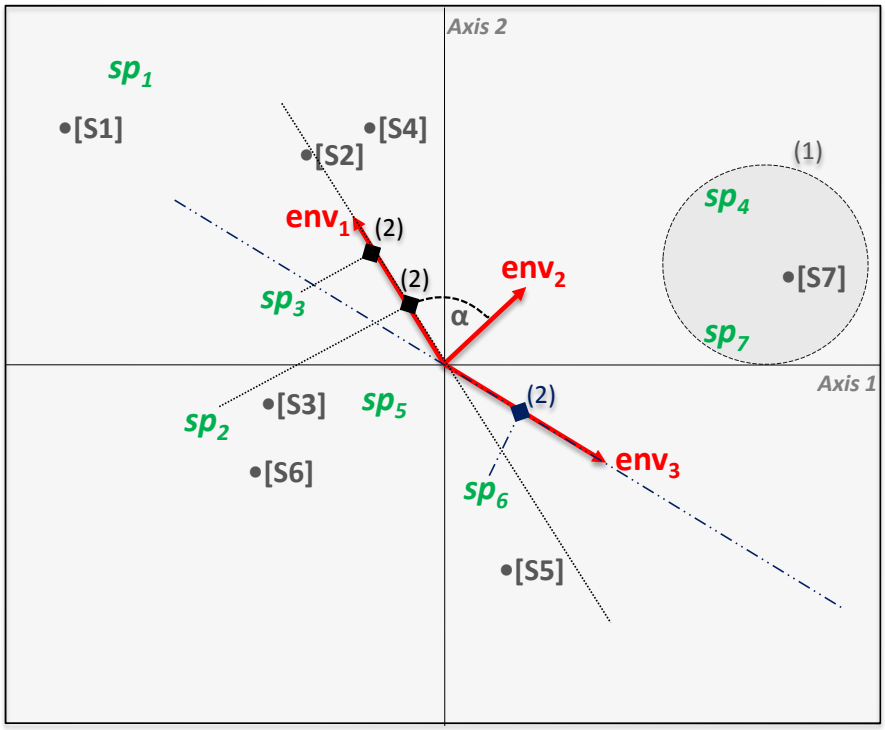


Figure 6. CCA triplot diagram 1-2 of hypothetical environmental variables, species and sites samples. Species points are coded as *sp*$_1$ to *sp*$_8$, environmental vectors as env$_1$ to env$_3$ and sites points as [S1] to [S7].

For species points (*sp1*, *sp2*, *sp3*, *sp4*, *sp5*, *sp6*, *sp7* and *sp8*):

- the distance between the *sp* "symbols" in the diagram approximates the dissimilarity of distribution of relative abundance of those species across the samples, measured by their χ^2 distance. Points in proximity correspond to species often occurring together (as it is illustrated by (1)).

For environmental variable arrows (env_1, env_2 and env_3):

- each arrow points in the expected direction of the steepest increase of values of environmental variable. The angles between arrows (as appear α in figure 6) indicate correlations between individual environmental variables. More precisely, the approximated correlations of one environmental variable with the others can be interpreted by projecting their arrowheads onto the imaginary axis running in the direction of that variable's arrow.

For the species-environmental relationship:

- the species can be projected perpendicularly onto the line overlaying the arrow of particular environmental variable (as it is illustrated by (2)). These projections can be used to approximate the optima of individual species in respect to values of that environmental variable. Species projection points are in the order of the predicted increase of optimum value for that variable.

In conclusion, it is important to refer that CCA is effective at detecting relationships between environmental variables and abundance data when there is sufficient range in the environmental variables (which is determined by the variation of the variables and the distribution of their values) and the abundances (which is determined by the length

of the gradient, i.e., by a measure of how much change occurs in species abundances). In these cases, other approaches should be used (Verdonschot and ter Braak, 1994).

2.3 CO-INERTIA ANALYSIS

Co-Inertia Analysis (Dolédec and Chessel, 1994) is a multivariate method that performs a simultaneous analysis of two tables (Fig. 7).

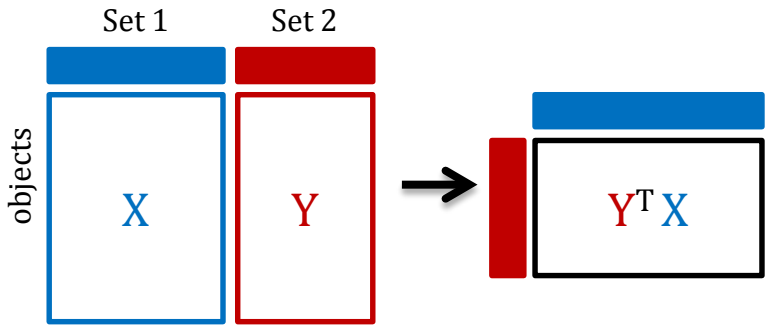


Figure 7. Co-Inertia Analysis initial scheme.

In other words, it describes the co-structure between two tables by summarizing as well as possible the squared covariances between first and second tables (Fig. 8). Moreover, Co-Inertia Analysis is considered an extension of the inter-battery analysis, which is the first step of partial least squares regression (Partial Least Squares). The difference between these two methods is that Co-Inertia Analysis works on a covariance matrix instead of a correlation matrix (Dolédec and Chessel, 1994).

The Co-Inertia Analysis is a very flexible technique and is a generalization of CANCOR as well as of RDA, both with additional restrictions (Dray et al., 2003).

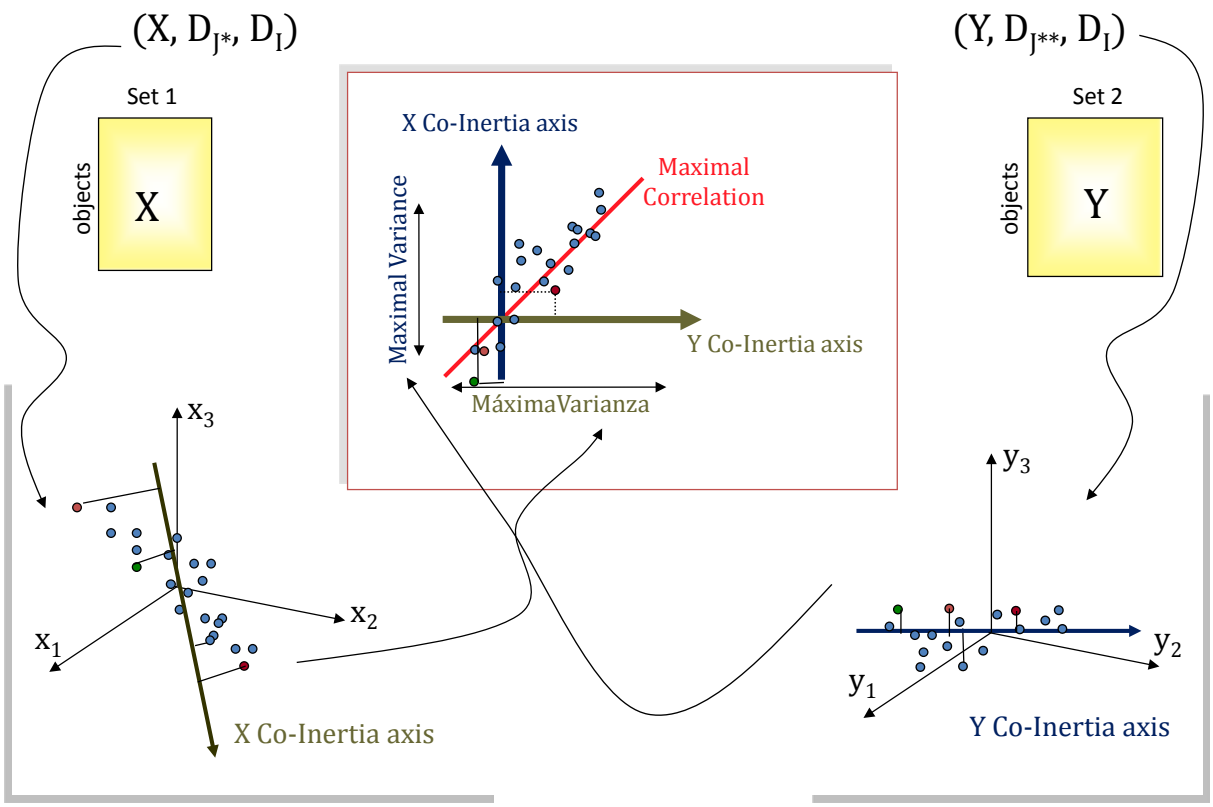


Figure 8. Principles of Co-Inertia Analysis. The two data tables *X* and *Y* produce representations of the *I* objects in two hyperspaces. Separate analyses find axes maximizing inertia in each hyperspace (first factorial axis). Co-Inertia Analysis aims to find a couple of co-inertia axes on which the *I* objects are projected. Co-Inertia Analysis maximizes the square covariance between projections of the objects on the co-inertia axes.

Co-Inertia Analysis has been successfully employed to demonstrate the co-structure (relation) between two data matrices in several areas (see for example, Pottier et al., 2007; Lair, 2005; Carassou et al., 2010; Twatwa et al., 2005; Pavoine and Bailly, 2007; Covain et al., 2008).

The mathematical description of Co-Inertia Analysis is presented using simple matrix notations (Fig. 8).

Let X be the first table, with I rows (objects) and J^* columns (first group of variables), and let Y the second table, with the same I rows, and J^{**} columns (second group of variables). X^T and Y^T are the transpose of X and Y, respectively. Let D_I be the diagonal matrix ($I \times I$) of objects weights, where $D_I = Diag(w_1, \dots, w_I)$, and D_{J^*} and $D_{J^{**}}$ be two metrics of hyperspace of first and second tables, respectively. Before perform the Co-Inertia Analysis, the two tables are analyzed separately.

Let (X_k, D_{J^*}, D_I) and $(Y_k, D_{J^{**}}, D_I)$ be the two statistical triplets. A generalized PCA of these triplets corresponds to the spectral decomposition of $X^T D_I X D_{J^*}$ and $Y^T D_I Y D_{J^{**}}$. When D_I is the matrix of uniform row weights ($w_l = {}^1/_I$), and D_{J^*} and $D_{J^{**}}$ are identity (euclidean metrics), then these analysis are simple PCA.

Co-Inertia Analysis is the eigenanalysis of matrix $X^T D_I Y D_{J^{**}} Y^T D_I X D_{J^*}$. Indeed, it is the analysis of a cross product table, and its triplet notation is $\left(Y^T D_I X, D_{J^*}, D_{J^{**}}\right)$. If the columns of both tables are centered, then the total inertia of each table is simply a sum of variances:

$$Iner_X = trace\left(X D_{J^*} X^T D_I\right) \text{ and } Iner_Y = trace\left(Y D_{J^{**}} Y^T D_I\right)$$

and the co-inertia between X and Y is in this case a sum of squared covariances:

$$CoIner_{XY} = trace\left(X D_{J^*} X^T D_I Y D_{J^{**}} Y^T D_I\right) \quad (2.3.1)$$

Co-Inertia Analysis maximizes the covariance between the row scores of the two tables (Dray et al., 2003). Hence, the meaning of the co-structure between the two data tables

is defined as follows: co-inertia is high when the values in both tables are high simultaneously (or when they vary inversely) and low when they vary independently or when they do not vary.

One of the most positive and interesting points of this method comes from the fact that maximize the covariance and this ensures that the scores do not have very small variances, and therefore have a good percentage of explained variance in each space (Thioulouse, 2010).

2.3.1 VISUALIZING CO-INERTIA ANALYSIS RESULTS

In order to illustrate the interpretation of the results obtained by Co-Inertia Analysis, a hypothetical example is showed (Fig. 9).

The first dataset contains the environmental variables measured (X) and the second dataset contains the species abundance (Y).

The relationship between the environmental dataset and the species abundance dataset is provided by the Co-Inertia Analysis.

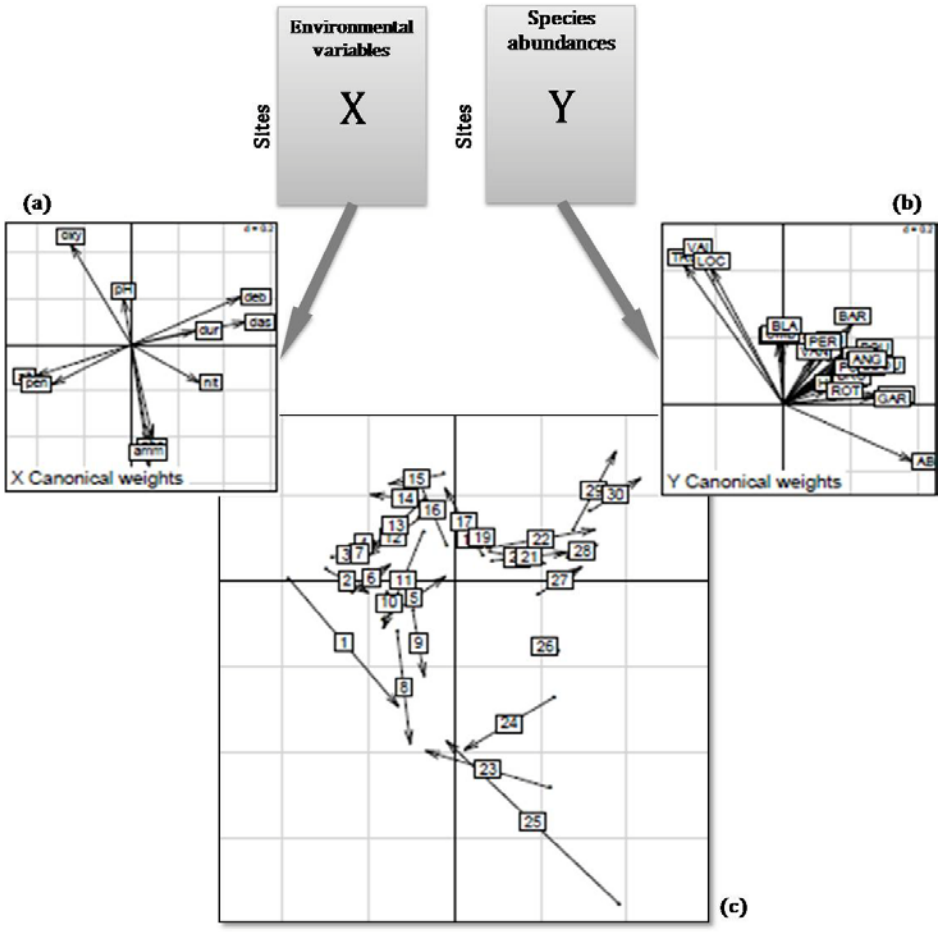


Figure 9. Canonical weights. The scatter plots (a) and (b) represent the coefficients of the combinations of the variables for each table to define the co-inertia axes. Co-structure scatter plot (c) with arrows is specific to the Co-Inertia Analysis, and represents the individuals (in this case, the sites sampled). The beginning of the arrow is the position of the site described by the environmental data set; the end of the arrow is the position of the site described by the species abundances (adapted from Dufour (2009)).

For example, sites 23 to 26 are apart from the other sites. Although being apart, they linked to nitrates (nit) and ammonium (amm) (see *X* canonical weights graph). In these sites (23-26), there are also less species (see *Y* canonical weights graph). Interestingly, the abundance of species in 23-26 is similar to the abundance of species in sites 1, 8 and 9.

In addition, the length of the connecting line reveals the disagreement or the consensus between the two profiles (species–environment), i.e., the length of the line is proportional to the divergence between the datasets. When the datasets agree very strongly, the arrows will be short. Likewise, a long arrow demonstrates a locally weak relationship between the environment and diatoms features for that case.

For instance, site 26 (similar to 27 and 28 sites) presented short arrow which means that the environmental factors explained well the distribution of species for that case.
On the other hand, and notwithstanding the strong dispersion of the environmental points and a poor fit between the species and environment (long arrows), sites 23, 24 and 25 were grouped together meaning a similar pattern from species point of view (the end of the arrow are close and points in same direction).

In conclusion, matching two tables is a complex procedure and there are several ways to proceed. A Co-Inertia Analysis is the match between two tables and their associated triplets. Each triplet may be a PCA, a CA or a multiple CA.

Experience shows that, in many cases, different methods will give similar results, but that in particular situations the results of a study can greatly depend on the choice of the multivariate method (Dray et al., 2003). At the end, the choice always depends on the initial question.

CHAPTER 3

CHAPTER 3 – METHODS FOR DESCRIBING THREE-WAY MULTIVARIATE DATA: ANALYSIS OF THE STRUCTURES' STABILITY

3.1 INTRODUCTION

When sampling is repeated in time (or in space), a sequence of tables are obtained. When a pair of ecological table is repeated, the result is a pair of *k*-table, or two data cubes. One sequence of tables makes one data cube, and a sequence of pairs of tables makes a pair of data cubes (a species data cube plus an environmental data cube). Analyzing the relationships between the two cubes can give useful insights into the evolution of species-environment relationships and *K*-table methods are required.

The *K*-table analysis methods are used to analyse series of tables. They belong to three families (Thioulouse, 2010): STATIS (Structuration des Tableaux a Trois Indices de la Statistique, L´Hermier des Plantes, 1976; Lavit, 1988, Lavit et al., 1994) (Fig. 10), Multiple Factor Analysis (MFA, Escofier and Pagès, 1984; Escofier and Pagès, 1994), and Multiple Co-Inertia Analysis (MCOIA, Chessel and Hanafi, 1996). Partial Triadic Analysis (PTA, Jaffrenou, 1978; Thioulouse and Chessel, 1987; Blanc et al., 1998) is one of the simplest analyses of the STATIS family, and it can be seen as the PCA of a series of PCA's (Thioulouse, 2010).

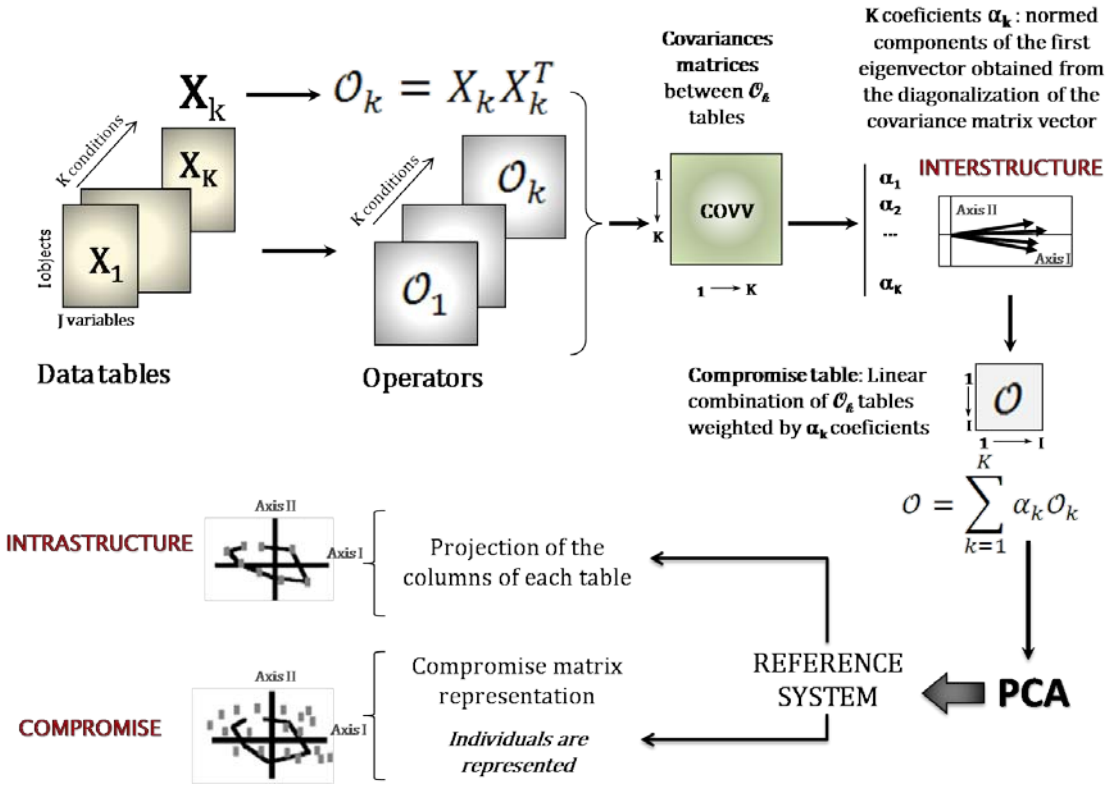


Figure 10. Simplified scheme of the principle of the STATIS method. STATIS is a flexible technique for handling three-way data and it consists in a three-step method. The method is especially geared toward the situation where the same individuals have measures on a set of variables (which may vary) under a number of conditions. Here only a simplified description is provided (for full details see (Lavit, 1988, Lavit et al., 1994). X_k= initial tables; J = number of variables; I = number of objects; $\mathcal{O}_k$ = similarity matrix between subjects; α_k = coefficients of the first eigenvector.

Multivariate analysis methods for pairs of data cubes are not widespread. Two of them are based on Co-Inertia (Thioulouse, 2010): the first one is called Between-Group Co-Inertia Analysis (BGCOIA, Franquet et al., 1995), and the second one is the STATICO method (Simier et al., 1999; Thioulouse et al., 2004).

Besides the several strategies that exist to study a pair of data tables, the intent of this chapter is to give an exposition of PTA, as well as STATICO, since their methodological bases and objectives are related to the proposed method in **CHAPTER 5**.

Published applications (with plankton data) on the international peer-reviewed journals BIOMETRICAL LETTERS, FRESENIUS ENVIRONMENTAL BULLETIN, ESTUARINE, COASTAL AND SHELF SCIENCE, ZOOLOGICAL STUDIES and JOURNAL OF APPLIED STATISTICS are presented in the appendix section **PUBLISHED ARTICLES BY THE AUTHOR** (Papers II, III, IV, V, VII and VI, respectively).

3.2 PARTIAL TRIADIC ANALYSIS

The PTA (Jaffrenou, 1978; Thioulouse and Chessel, 1987; Blanc et al., 1998) is a multi-table technique allowing the analysis of a set of data tables recorded through a series of sampling conditions (Fig. 11). PTA belongs to the STATIS family of the *k*-table analysis methods and aims to analyse a three-way data array (i.e., series of *k*-tables) seen as a sequence of two-way tables (i.e., the *k*-table is summarized by a matrix). In PTA all *k*-tables must have the same rows and the same columns (as a cube data set), which mean that the same variables must be measured at the same sampling occurrences, several times or conditions. Its main advantage or potential is related to the fact that it works with original data instead of operators (as STATIS performs), which permits to do all the interpretations of the results in a direct way. In comparison to the STATIS methods, the PTA is one of the simplest. It can be seen as the PCA of series of PCA's (Thioulouse, 2010) and allows the optimal projection of trajectories.

The PTA applied to the study of a series of *k*-tables has been successfully employed to interpret this kind of cube data sets (see for example, Mendes et al., 2011b; Mendes et al., 2010; Rolland et al., 2009; Decäens et al., 2009; Rossi, 2006; Ernoult et al., 2006; Gaertner, 2000).

The mathematical description of PTA is presented using simple matrix notations (Fig. 11).

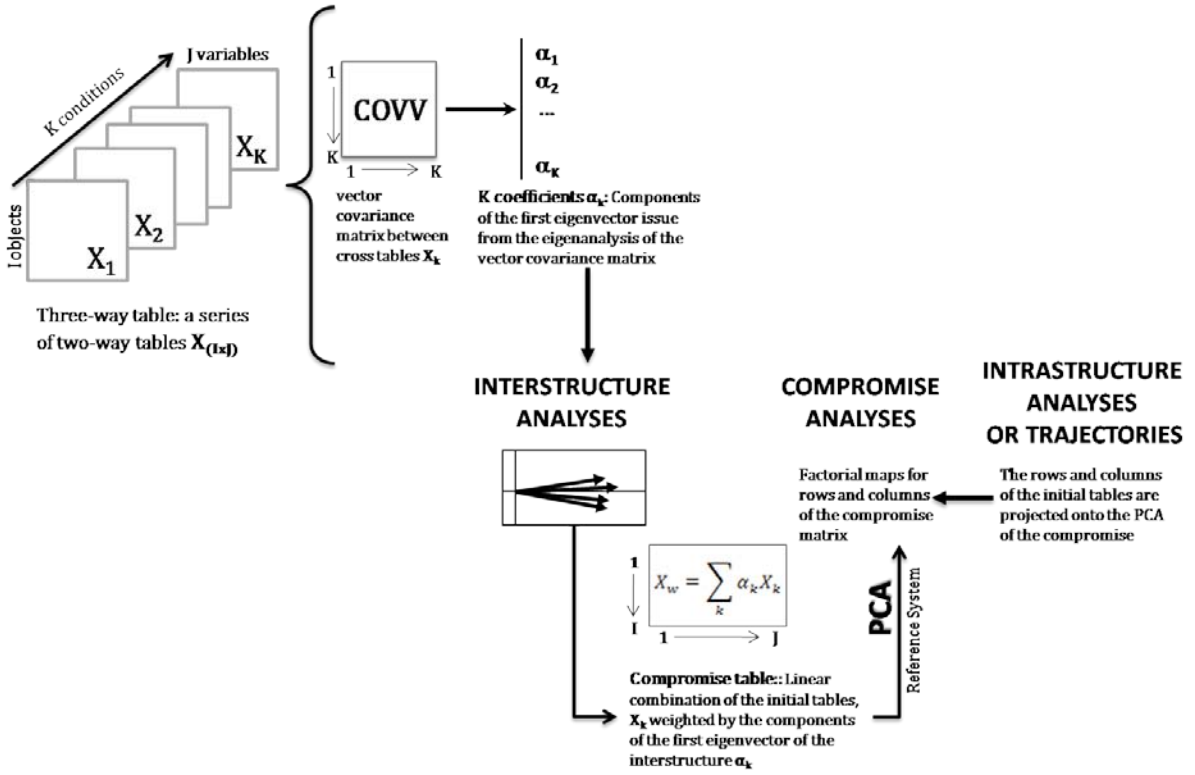


Figure 11. General scheme of the PTA: construction of the interstructure matrix and extraction of the compromise table.

The principle of PTA is to identify the structure which is common to the series of tables having the same rows (I) and same columns (J). More precisely, it searches for structures that are stable among the sequence of tables.

After the initial transformations (by centralization, normalization, etc.) the K tables, X_k, are obtained. Each X_k is a data matrix of J quantitative variables measured on the same I objects. According to the PTA methodology, (X_k, D_J, D_I) defines a statistical triplet where D_J and D_I are positive definite weighting matrices for variables and objects and whose positive diagonal elements add up to 1. The PTA, like all the STATIS family methods, is a three-step procedure (Lavit et al., 1994), namely the interstructure, the compromise and the intrastructure analyses (also called trajectories).

First a matrix of scalar products between tables is computed (i.e., the $k \times k$ matrix which elements are: $COVV(X_k, X_l) = trace\left(X_k^T D_I X_l D_J\right)$). The diagonalization of this matrix provides eigenvectors. The k coefficients, α_k, of the first eigenvector are then used to weight the k tables in the calculation of the compromise table. Then, a PCA of this $k \times k$ table is performed in order to establish the ordination of the different matrices. Alternatively, a matrix of vector correlations (RV) can be used to rescale the importance of the tables. Each element in this table is:

$$RV(X_k, X_l) = \frac{COVV(X_k, X_l)}{\sqrt{VAV(X_k)VAV(X_l)}} \quad (3.2.1)$$

where $VAV(X_k)$ is the variance of the vector obtained by putting all the columns of table X_k one below the other. It is basically the vector variance of table X_k, i.e. $VAV(X_k) = trace\left(X_k^T D_I X_k D_J\right)$. The vector correlation matrix and the vector covariance matrix are linked by the same relationships as the normal correlation and covariance matrices. Each table is projected onto the factorial diagram obtained from the PCA analysis and represented by an arrow, in order to establish the ordination of the different tables, which summarizes the global structure and the relationships between them. This configuration (based upon the covariance matrix) allows an overall graphical comparison of the tables and shows proximities between the configurations of the same observations.

The second step of this method is the analysis of the compromise (X_w), a fictitious table which is computed as the weighted mean of all the tables of the series, using the components of the first eigenvector of the interstructure as weights (i.e. issued from the eigenvalues of the vector covariance matrix) (Thioulouse et al., 2004). In other words, it

consists in calculating a linear combination of the k initial tables with the aim of constructing a mean table of maximum inertia (Gaertner, 2000):

$$X_w = \sum_k \alpha_k X_k \tag{3.2.2}$$

The main property of this compromise is that maximizes the similarity with all initial tables, as measured by the sum of their squared dot product $\sum_k \langle X_w, X_k \rangle^2$ (when the tables are normed, the dot product is the RV coefficient).

The weight of each table is proportional to its inertia, so tables that are unstructured from the others will be downweighted. This property ensures that the compromise will resemble all the tables of the sequence as closely as possible in least square sense (Thioulouse, 2010).

Once obtained, X_w (which has the same dimensions and the same structure and meaning as the tables of the original series) is then analyzed by a PCA and the rows and columns of the individual matrices can be also projected onto the analysis as supplementary individuals and supplementary variables, respectively. Thus, the analysis of the compromise gives a factor map that can be used to interpret the structures of this compromise. In other words, gives a picture of the structures which are common to all the tables (Thioulouse et al., 2004).

The third step summarizes the variability of the succession of tables in comparison to the common structure defined by the compromise. The rows and columns of all the tables of the three dimensional array are projected on the factor map of the PCA of the

compromise as additional elements (Thioulouse, 2010; Thioulouse et al., 2004) in order to summarize the reproducibility of the structure across the series of tables. Denote $\mathcal{U}$ by the matrix of the eigenvectors of the analysis of the compromise. The coordinates of the rows of the table X_k are:

$$\mathcal{R}_k = X_k D_J \mathcal{U} \tag{3.2.3}$$

and the coordinates of its columns are

$$\mathcal{C}_k = X_k^T D_I X_w D_J \mathcal{U} \Lambda^{-1/2} \tag{3.2.4}$$

$\Lambda^{-1/2}$ being the diagonal matrix of the inverses of the square root of the eigenvalues of the analysis of the compromise.

Each row of each table is represented by a point in the space of its J columns and can be projected as supplementary individual onto the principal axes of compromise. The same procedure is applied (similarly) for the columns (Simier et al., 1999). The points can then be linked, for example by lines, to underline their trajectories. Their study concludes the third step of the method.

In conclusion, PTA offers the possibility to study the stable structure in a sequence of tables and the compromise displays (optimally) this stable structure (when exists).

Moreover, the study of the dynamics trajectories shows how each table moves away from that optimal compromise solution.

3.2.1 VISUALIZING PARTIAL TRIADIC ANALYSIS RESULTS

In the appendix section **PUBLISHED ARTICLES BY THE AUTHOR** (Papers II and III), two applications will be shown illustrating many of the aspects covered in this chapter. Both works were already published on international peer-reviewed journal s, namely BIOMETRICAL LETTERS and FRESENIUS ENVIRONMENTAL BULLETIN.

The first work,

> ***Mendes S****, Fernández-Gómez MJ, Pereira MJ, Azeiteiro U, Galindo-Villardón MP. 2010. The efficiency of Partial Triadic Analysis method: an ecological application. Biometrical Letters 47: 83-106.*

leads with phytoplankton data and demonstrates the efficiency of PTA in the simultaneous analysis of several data tables. Moreover, the results also showed how PTA is well-adapted to the treatment of spatio-temporal data.

A synopsis of the study and their results is here presented:

ABSTRACT

"In this paper we present a Partial Triadic Analysis (PTA) method that can be applied to the analysis of series of ecological tables. The aim of this method is to analyze a three-way table, seen as a sequence of two-way tables. PTA belongs to the family of STATIS methods and comprises three steps: the interstructure, the compromise and the trajectories. The advantage of this method is related to the fact that it works with

original data instead of operators, which permits all the interpretations to be performed in a directly way. In this study we present an efficient application of the PTA method in the simultaneous analysis of several data tables and show how well-adapted it is to the treatment of spatio-temporal data. Two kinds of matrices were constructed: a species abundance table and an environmental variables table. Both matrices had the sampling sites in rows. All computations and graphical displays were performed with the free software ADE-4. An example with phytoplankton and environmental factors data is analyzed, and the results are discussed to show how this method can be used to extract the stable part of species and environment relationships".

KEYWORDS: *Partial triadic analysis, multi-table analysis, STATIS, species abundance, environmental factors.*

The second work,

> ***Mendes S****, Marques SC, Azeiteiro UM, Fernández-Goméz MJ, Galindo-Villardón MP, Maranhão P, Morgado F, Leandro SM. 2011. Zooplankton distribution in a Marine Protected Area: the Berlengas Natural Reserve (Western coast of Portugal). Fresenius Environmental Bulletin 20: 496-505.*

leads with zooplankton data and offered the possibility to study the three-dimensional array in two ways: (A) the spatial variability of the zooplankton community and its dynamics in time (data were organized as a series of tables for each date, where each column corresponded to the species density and each row corresponded to a sample) and (B) the dynamic trajectories of the zooplankton community per site (data were

considered as a series of tables for each site, where each column corresponded to the species density and each row corresponded to a sample).

A summary of the study and their results is here presented:

ABSTRACT

"Zooplankton distribution in the Berlengas Natural Reserve (Portugal) was studied over a period of one year (February 2006 to February 2007). Monthly sampling was performed at 6 stations, differentiated according to depth and distance to the coastline. The aim of this study was to investigate the overall zooplankton variability through its different dimensions (space vs. time). The Partial Triadic Analysis (PTA) was used to study the spatial variability of the zooplankton community and its dynamics in time and the dynamic trajectories of the zooplankton species for each site. It was possible to distinguish a neritic-ocean gradient of the zooplankton composition and a temporal variability. Four distinct periods can be highlighted considering the distribution of the dates and the arrangement of the species: (i) the first one comprised August to November, (ii) the second one was related to June and July, (iii) the third one associated with spring (April and May) and, (iv) the latest one was related to winter (February, March and December 2006 and January and February 2007). The PTA method showed the similarities between the successive data tables and proved to be useful for investigating biotic structures and detecting spatial-temporal patterns in zooplankton distribution."

KEYWORDS: *Berlengas Natural Reserve, partial triadic analysis, spatio-temporal distribution, zooplankton.*

3.3 STATICO

The name of STATICO derives from the two methods STATIS and Co-Inertia, and has previously been widely explored in several areas (see for example, Gonçalves et al., 2011b; Mendes et al., 2011a; Thioulouse, 2010; Mendes et al., 2009b; Gonçalves et al., 2011a; Carassou and Ponton, 2007; Simier et al., 2006; Thioulouse et al., 2004).

The STATICO method is particularly suited to study the stable relationship between pairs of tables when both vary over k conditions (Fig. 12). It is an exploratory tool for three-way data analysis and is based on the PTA (Jaffrenou, 1978; Thioulouse and Chessel, 1987; Blanc et al., 1998) of a sequence of cross product tables.

The principle of STATICO method is briefly explained and its mathematical description is presented using simple matrix notations (Fig. 12).

Starting from the sequence of paired tables, each cross product table is computed using the pair of tables at each condition. All the tables of the sequence do not need to have the same number of rows, but they need to have the same number of columns across conditions. This means that the number of sampling occurrences can vary from one condition to another, but the number of columns (J^*) of the first group of matrices ($\mathbb{X}$) must be the same for all the conditions (K), and the number of columns (J^{**}) of the second group of matrices ($\mathbb{Y}$) must also be the same for all conditions. Therefore, all the cross product tables have the same number of rows ($I = J^{**}$) and columns ($J = J^*$). They contain the covariances between the columns of the two tables.

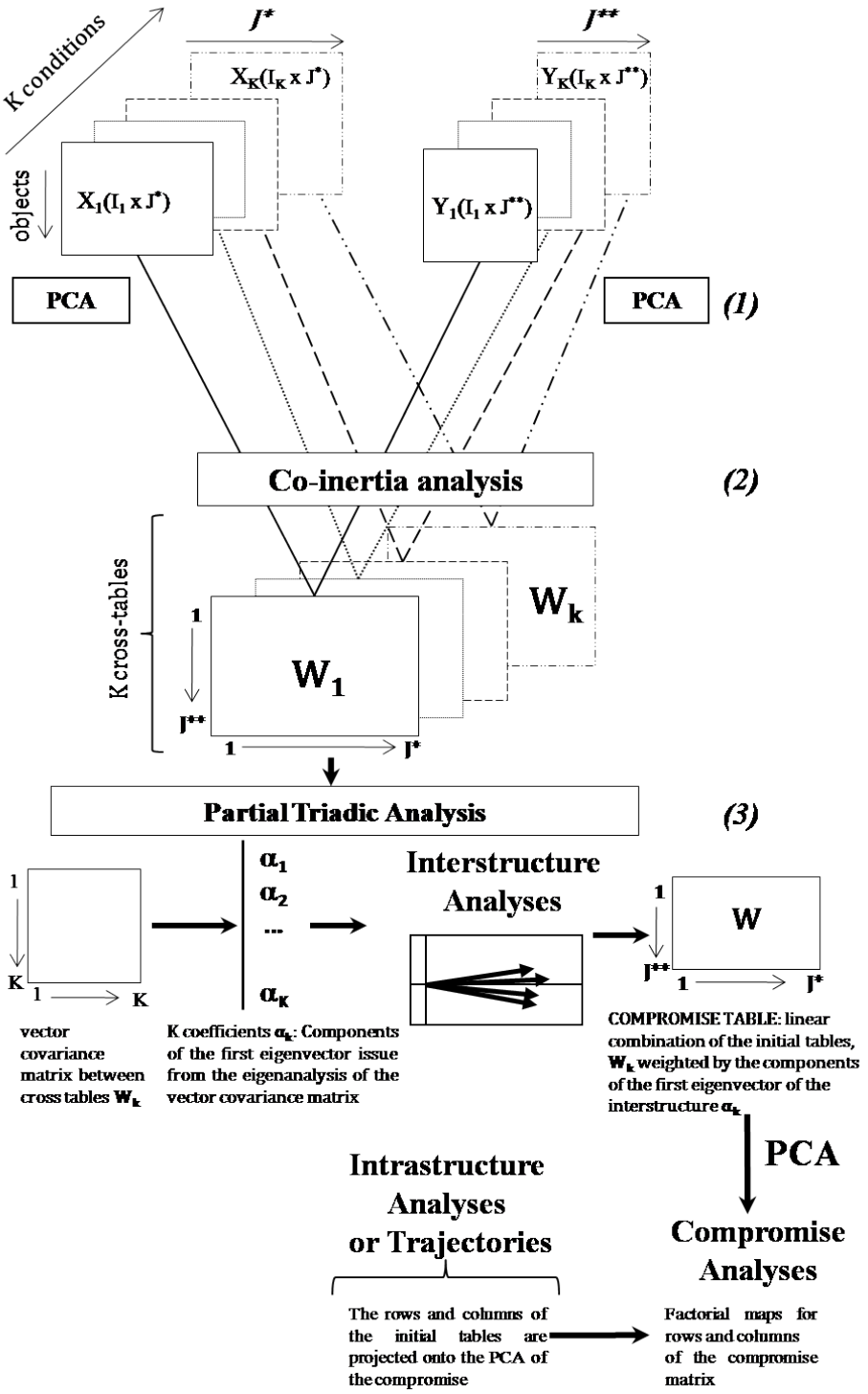


Figure 12. STATICO general scheme. The data structure is a sequence of K paired tables. X_k and Y_k are respectively the first group (dimension $I_K \times J^*$) and the second group (dimension $I_K \times J^{**}$) of tables in the pair. W_k is the cross-table at K occurences; J^* is the number of variables of $\mathbb{X}$, J^{**} is the number of variables of $\mathbb{Y}$, I_K is the number of rows at kth different conditions. (1) Basic analyses (PCA for $\mathbb{X}$ and $\mathbb{Y}$ tables) are performed on each table; (2) Co-inertia analyses allow linkage of the pairs of PCA-PCA, producing a sequence of K cross-tables; (3) PTA is finally used to analyze this sequence.

Let (X_k, D_{J^*}, D_{I_k}) and $(Y_k, D_{J^{**}}, D_{I_k})$ be the pair of triplets at condition k. The frontal slice X_k is the table of variables (first group of matrices) measured at condition k, and the frontal slice Y_k is the other table of variables (second group of matrices) observed at the same condition k. D_{J^*} and $D_{J^{**}}$ are diagonal matrices of dimensions $J^* \times J^*$ and $J^{**} \times J^{**}$, respectively and they are the same for all the conditions. Finally, D_{I_k} is equal to $Diag\left(\frac{1}{I_k}\right)$ and is the same for both tables (I is here assumed as the number of samples to be assessed).

Let W_k be the kth cross product table, that is, $W_k = Y_k^T D_{I_k} X_k$ and (W_k, D_I, D_J), $I = J^{**}$ and $J = J^*$ the Co-inertia Analysis triplet at condition k. Hence, those cross product tables are computed and the obtained data cube is then analyzed by means of a PTA (for more detail, see Chapter 3.2). So, STATICO method is a three main step procedure (Fig. 12): (1) separated analysis of each table, (2) Co-Inertia Analysis (for details see Chapter 2.3) of each paired tables, and (3) the PTA analysis of the Co-Inertia cube.

First, a global comparison between the conditions is made in order to assess the interstructure. This step gives optimal weights α_k such that the inertia of the triplet $\left(\sum_{k=1}^{K} \alpha_k W_k , D_I, D_J\right)$ is maximum with the constraint $\sum_{k=1}^{K} \alpha_k^2 = 1$.

The second step is to construct the compromise solution (W) for the K data matrices W_k. Therefore, the compromise is determined by weighting (on average) the cross product tables using the weights α_k; thus $W = \sum_{k=1}^{K} \alpha_k W_k$ (Simier et al., 1999). This compromise solution W is then subjected to an analysis (through a PCA) to evaluate the similarity between the subjects of the compromise structure. Moreover, a

graphical representation is displayed, and rows and columns of W are plotted (Thioulouse, 2010).

The third step, the intrastructure (or trajectories) is to evaluate to what extent the compromise structure is reflected in the individual data matrices W_k. This can be done by projecting the rows and columns of each table of the sequence in the analysis of the compromise (i.e., into the space of the compromise solution), with usual supplementary element projection technique (Thioulouse, 2010). This gives a display of the variables of the first group of matrices at each condition, of the variables of second group of matrices at each condition, and two displays of the sampling occurrences at each condition (one from the point of view of J^{**} variables, I rows of W_k, and one from the point of view of J^{*} variables, J columns of W_k).

3.3.1 VISUALIZING STATICO METHOD RESULTS

In the appendix section **PUBLISHED ARTICLES BY THE AUTHOR** (Papers IV, V, VI and VII), four applications will be shown illustrating many of the aspects covered in this chapter. All the papers were already published on international peer-reviewed journals, namely ESTUARINE, COASTAL AND SHELF SCIENCE, JOURNAL OF APPLIED STATISTICS and ZOOLOGICAL STUDIES.

The first work,

> ***Mendes S****, Fernández-Gómez MJ, Resende P, Pereira MJ, Galindo-Villardón MP, Azeiteiro UM. 2009. Spatio-temporal structure of diatom assemblages in a temperate estuary. A STATICO analysis. Estuarine, Coastal and Shelf Science 84: 637-664.*

leads with phytoplankton data (previously studied, and published, by means of CCA) and confirmed the applicability of the STATICO method to study the diatom estuarine dynamics. Moreover, the work also highlighted the advantages of STATICO over the classical CCA technique.

A summary of the study and their results is here presented:

ABSTRACT

"This study examines the spatio-temporal structure of diatom assemblages in a temperate estuary (Ria de Aveiro, Western Portugal). Eighteen monthly surveys were conducted, from January 2002 to June 2003, at three sampling sites (at both high and low tide) along the estuarine salinity gradient. The relationship of diatom assemblages and environmental variables was analysed using the STATICO method, which has been designed for the simultaneous analysis of paired ecological tables. This method allowed examination of the stable part of the environment-diatom relationship, and also the variations of this relationship through time. The interstructure factor map showed that the relationship between the 11 environmental variables and the abundance of the 231 diatom species considered was strongest in the months May and September 2002 and

January, February and May 2003. The stable part of the species-environment relationships mainly consisted of a combined phosphate, chlorophyll a and salinity gradient linked to a freshwater-marine species gradient. A more pronounced gradient was observed in January, February and May 2003. Diatom assemblages showed clear longitudinal patterns due to the presence of both marine and freshwater components. May and September 2002 had the least structured gradients with marine-estuarine species appearing in the freshwater side of the gradient. The most complete gradient in February 2003 could be considered, in terms of bio-ecological categories, as the most structured period of the year, with a combination of strong marine influence in the lower zone and freshwater influence in the upper. The best-structured gradients were during periods of a diatom bloom. Stable diatom assemblages (with a strong structure and a good fit between the diatoms and environment) are described and characterized. This study shows the efficiency of the STATICO analysis. The inclusion of space-time data analysis tools in ecological studies may therefore improve the knowledge of the dynamics of species–environmental assemblages."

KEYWORDS: *environmental forcing, Diatoms, Estuaries, multitable analysis, STATICO*

The second work,

> ***Mendes S****, Fernández-Gómez MJ, Pereira MJ, Azeiteiro UM, Galindo-Villardón MP. 2011. An empirical comparison of Canonical Correspondence Analysis and STATICO in identification of spatio-temporal ecological relationships. Journal of Applied Statistics* (First online: 18 November 2011; http://dx.doi.org/10.1080/02664763.2011.634393).

presented a comparison between CCA and STATICO technique and showed the advantages and disadvantages of each method from a practical standpoint, on their respective graphical outputs and on their global properties. The treated ecological data (already studied by Resende et al. (2005) and Mendes et al. (2009b)) are a sequence of pairs of ecological tables, where species abundances and environmental variables are measured at different, specified location, over the course of time.

A synopsis of the study and their results is here presented:

ABSTRACT

"The wide-ranging and rapidly evolving nature of ecological studies means that it is not possible to cover all existing and emerging techniques for analyzing multivariate data. However, two important methods enticed many followers: the Canonical Correspondence Analysis and the STATICO analysis. Despite the particular characteristics of each, they have similarities and differences, which analyzed properly, can, together, provide important complementary results to those that are usually exploited by researchers. If on one hand the use of Canonical Correspondence Analysis is completely generalized and implemented, solving many problems formulated by ecologists, on the other hand this method has some weaknesses mainly caused by the imposition of the number of variables that is required to be applied (much higher in comparison with samples). Also, the STATICO method has no such restrictions, but requires that the number of variables (species or environment) is the same in each time or space. Yet, STATICO method presents information that can be more detailed since it allows visualizing the variability within groups (either in time or space). In this study,

the data needed for implementing these methods is sketched, as well as the comparison is made showing the advantages and disadvantages of each method. The treated ecological data are a sequence of pairs of ecological tables, where species abundances and environmental variables are measured at different, specified location, over the course of time."

KEYWORDS: *multitable analysis, multivariate data, ecological data analysis, simultaneous analysis, paired tables*

The third work,

> *Gonçalves AMM, Pardal MA, Marques SC,* ***Mendes S****, Fernández-Gómez MJ, Galindo-Villardón MP, Azeiteiro UM. 2011. Diel vertical behavior of Copepoda community (naupliar, copepodites and adults) at the boundary of a temperate estuary and coastal waters. Estuarine, Coastal and Shelf Science* (First online: November 2011; doi:10.1016/j.ecss.2011.11.018).

leads with copepod communities data and provided important results on their composition, structure and dynamics, as well as in terms of their short-term changes under influence of an estuarine hydrographic regime.

A synopsis of the study and their results is here presented:

ABSTRACT

"Presently, despite an apparent growing interest on diel vertical migration research topic, scarce efforts were developed in southern European marine coastal systems. This study investigated copepod assemblages' distribution patterns and determined the main structuring hydrological and physical factors at different temporal scales. Seasonal, tidal, lunar and diel vertical migrations accomplished by horizontal movements were examined on the main copepod fraction of the Mondego estuary. Seasonal samples were conducted hourly at the mouth of the estuary, during diel cycles (25h), both over neap and spring tides, at the bottom and surface, using a 63µm and 335µm mesh size nets. Simultaneously, four sampling sites inside the estuary were conducted during flood tide to evaluate and compare copepods species' distribution along the estuary. A first attempt to characterize species' life cycles was conducted. Spring-spring tide revealed to be the situation that best expresses the stable part of copepod-environment dynamics. Acartia tonsa and Oithona nana distributed mainly at the bottom during ebb tides. A clear resident estuarine performance was noticeable in O. nana proving the estuarine preferences of the species. Neritic species showed preferences by saline waters, whereas the resident species distributed mainly at estuarine areas. Copepodites stages showed a similar distribution pattern as estuarine species, avoiding leaving the estuary. In contrast nauplii and O. plumifera showed higher densities at surface flood tides. Indeed, vertical migrations accomplished by horizontal movements were mainly influenced by depth and tidal cycles, whereas day and night were not ecologically significant."

KEYWORDS: *Copepoda, Life history stages, Vertical migration, Horizontal movements, Temperate estuarine system.*

Finally, the four work,

> *Gonçalves AMM, Pardal MA, Marques SC,* ***Mendes S****, Fernández-Gómez MJ, Galindo-Villardón MP, Azeiteiro UM. 2011. Response to Climatic variability of Copepoda life history stages in a southern European temperate estuary. Zoological Studies (in press).*

leads with copepods data and by means of STATICO the results revealed the environmental parameters which best explained copepods dynamics, highlighted the seasonal variations in the distribution patterns of copepods in dependence to hydrological factors and stressed inter-annual variations of copepods assemblages during an extreme drought.

A synopsis of the study and their results is here presented:

ABSTRACT

"This study aims to investigate the effects of an extreme climate event (severe drought) on Copepoda ecology. Monthly samples were conducted from 2005 to 2007, at five stations, using a 63 and 335µm mesh nets. Calanoida were represented mainly by Acartia clausi, Temora longicornis and Acartia tonsa and Cyclopoida by Oithona plumifera and Acanthocyclops robusts. A. clausi and T. longicornis dominated at the mouth and middle estuary; A. tonsa and A. robustus were associated to the upper estuary while O. plumifera showed the highest densities at the downstream section. Nauplii occurred in higher densities at the mouth. The relationship of copepod assemblages and environmental factors was analyzed using the STATICO method which

allowed distinguishing the combination factors that mostly contributed to this relationship. Winter was characterized by high concentrations of nutrients, cold waters and low salinities while summer was related, in general, by high values of phosphate, salinity and temperature. Marine and estuarine species (mainly copepodites) showed high densities in summer. Freshwater species occurred at maximal densities in winter, coincidently with higher river flow. Copepoda assemblages showed a clear seasonal pattern that superimposed to the inter-annual variability. Moreover, the severe drought was responsible for the predominantly dominance of marine species."

KEYWORDS: *Copepods; life stages; Mondego estuary; Seasonal and inter-annual variability; STATICO.*

CHAPTER 4

CHAPTER 4 – MODELS FOR ANALYZING THREE-WAY MULTIVARIATE DATA: DYNAMIC ANALYSIS OF STRUCTURES

4.1 INTRODUCTION

Environmental data sets are, in general, multidimensional and have a complex structure. The nature of the data arising from these areas can be very different, which tends to complicate the data analysis. The analytical problems are often further complicated by biological and ecological variations. Hence, in dealing with data analysis in the environmental subject it is important to have access to a diverse set of methodologies in order to be able to deal with the problems in a sensible way.

Usually, data are collected as sets (tables) of objects and variables obtained under different experimental circumstances or for various sampling periods, etc. Putting all tables together results in data with three-way structure (Rizzi and Vichi, 1995). An example for such data is when samples collected at different sampling sites and the environmental parameters or species densities are measured during certain period of time (sites × parameters or species densities × time). There are many tools helping to explore and interpret three (or higher) way structure of the data.

Tucker (1966) developed a series of three-way models, which are now known as the Tucker1, Tucker2, and Tucker3 models (Kroonenberg and de Leeuw 1980). The models are also collectively called three-mode principal component analysis or originally three-

mode factor analysis though it would be conceived as a component model (Tucker, 1966; Tucker, 1963).

The most popular is the Tucker3 model. In fact, Tucker2 can be seen as a specific case of the Tucker3 model and the Tucker1 model simply corresponds to unfolding[6] the array to a two-way matrix followed by a PCA. Hence, the focus will be on the Tucker3 model.

Many of the limitations associated with the resolution of a single matrix are partially or completely overcome when several matrices (the so-called three-way data sets) are treated together (Tauler, 1995). Three-way resolution methods always introduce a significant improvement in the recovery of the true response profiles and have the additional benefit of providing quantitative information.

Two tendencies prevail within the family of three-way resolution methods: the use of non-iterative procedures, which solutions are based on the resolution of a generalized eigenvalue/eigenvector problem (Juan et al., 1998); and the iterative methods, focused on the optimization of initial estimates by using suitable data structure and a few constraints (Juan et al., 1998; Tauler, 1995). Tucker3 model is a good representative of the former latter tendency, i.e., an iterative method.

General advantages and drawbacks are recognized for both trends. Contrary to three (or higher)-way methods, the non-iterative methods algorithms are very computer time

[6] Note that the term unfolding as used here should not be confused with the meaning of unfolding in psychometrics. Unfolding as used here is also known as reorganization or concatenation.

efficient, user-friendly and furnish unique solutions. However, these methods cannot be generalized to more than three-way data and they are only exploratory tools since they cannot be used in a modeling context as can be done with other three (or higher)-way models. Moreover, they consist of several steps that cannot be fused into a single loss function. Thus, it is difficult to evaluate the overall solution. However several fit measures, such as explained variances, can be derived both for separate analysis and for the fit of the individual matrices to the common compromise structure (Kroonenberg, 2008).

Tucker3 optionally is a more flexibility model. This flexibility is due to the core array $\mathbb{G}$, which allows an interaction between a factor of one mode with any factor in the other modes. The core array enables the exploration of the underlying structure of a multiway data set but at the same time presents some drawbacks.

First, a Tucker3 model cannot determine component matrices uniquely; infinity of other equivalent solutions can be obtained by rotating the result without changing the fit of the model.

Second, the iterative optimization slows down the resolution process and more demanding user intervention is required.

Finally, Tucker3 can provide more in-depth and relevant information for the interpretation of the results, however their understanding for applied researchers is

sometimes much more difficult compared to the classical methods (like STATIS family methods, among others).

This chapter shows the basic procedures and ideas for explaining what a three-way data is, which types of elements are commonly defined in a three-way array, and introduce the basis of the Tucker3 model (which their methodological bases and objectives are entirely related to the proposed method in the next chapter).

4.2 THREE-WAY METHODS: GENERAL CONSIDERATIONS

4.2.1 TERMINOLOGY AND NOTATION

A common form of three-way data consists of objects by variables by conditions where objects, variables and conditions should be taken as generic terms that have a different content in different areas of study. Such data can be arranged into a three-dimensional block and this three-way data block is referred as a three-way data array $\mathbb{X}$ (Fig. 13).

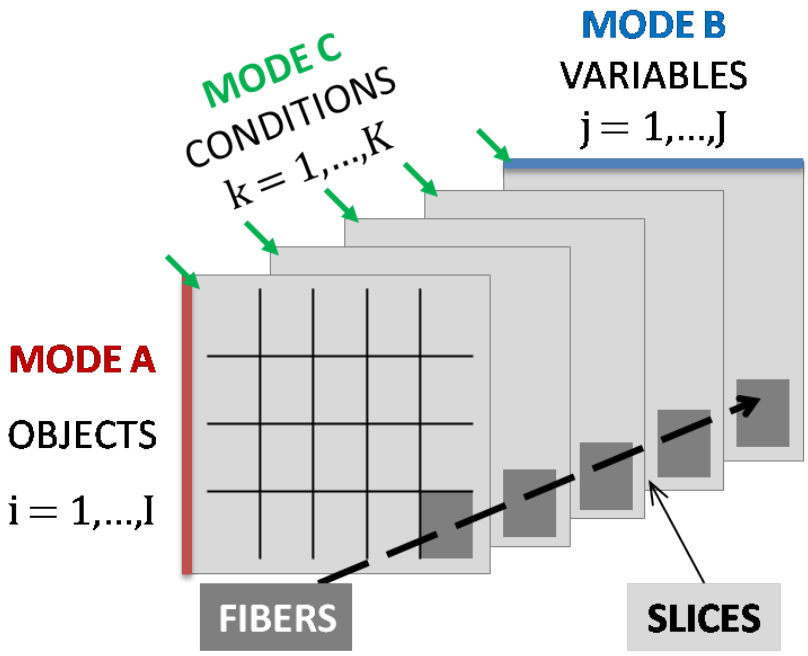


Figure 13. Three-way data array.

The three-way data $\mathbb{X}$ have dimensions I, J and K, which corresponds to the number of rows, columns and tables, respectively. Thus, an element of $\mathbb{X}$ is x_{ijk}, where $i = 1, \ldots, I$, $j = 1, \ldots, J$ and $k = 1, \ldots, K$.

It should be noted that an array is the multiway analogue term of a vector and a matrix (since a vector is a single-subscripted one-way array, a matrix is a double-subscripted two-way and a multiway array has three or more subscripts).

Modes, slices and fibers were defined as elements of a three-way data array. Each direction in a three-way array is called a way or a mode, and the number of levels in the mode is called the dimension of that mode. A table X_k is a slice of $\mathbb{X}$ of dimension $I \times J_k$ obtained by fixing the index in the third mode. The tables X_i and X_j are called horizontal and vertical slices, and can be obtained by fixing the first or the second mode index. In other words, a slice is a matrix and can also be named a frontal slab or layer. Fibers (referred as rows, x_{ki}, columns, x_{jk}, or tubes, x_{ij}) are the column vectors of the slices (Fig. 14).

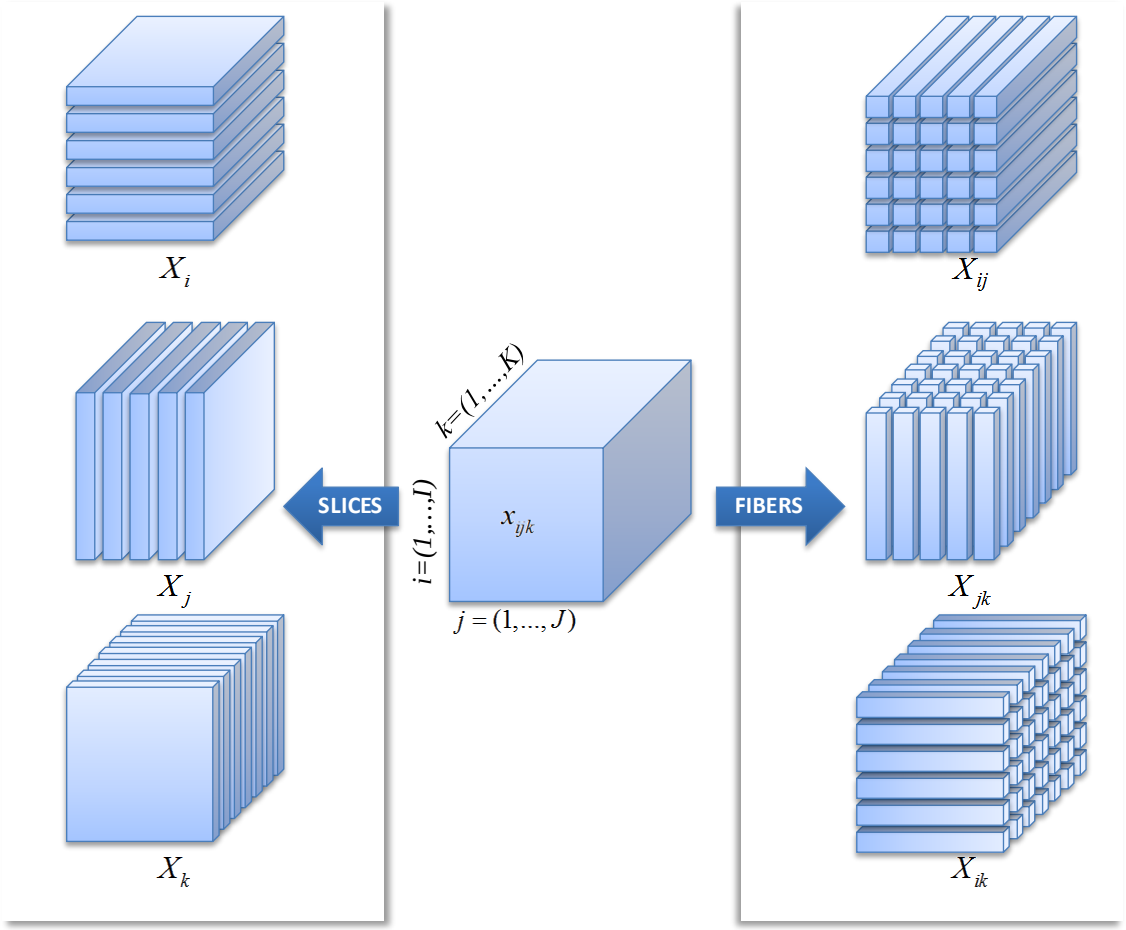


Figure 14. Three-way data array $\mathbb{X}$ with slices (horizontal X_i, vertical X_j and frontal X_k) and fibers (tubes X_{ij}, columns X_{jk} and rows X_{ik}) designated.

The number of levels in each way is I, J and K for first, second and third mode, respectively. The first way (objects) has index i running along the vertical axis, the second way (variables) has index j running along the horizontal axis and the third way (conditions) has index k running along the "depth axis" (Kroonenberg, 2008).

4.2.2 UNFOLDING

Unfolding is an important concept in multi-way analysis. It is simply a way of rearranging a multi-way array to a matrix, and in that respect not very complicated.

In figure 15 the principle of unfolding is illustrated graphically for a three-way array showing one possible unfolding. Unfolding is accomplished by concatenating matrices for different levels of, for example, the third mode next to each other. Notice, that the column-dimension of the generated matrix becomes quite large in the mode consisting of two prior modes. This is because the variables of the original modes are combined. There is not one new variable referring to one original variable, but rather a set of variables.

The profound effect of unfolding arises when the multi-way structure of the data is ignored, and the data treated as an ordinary two-way data set.

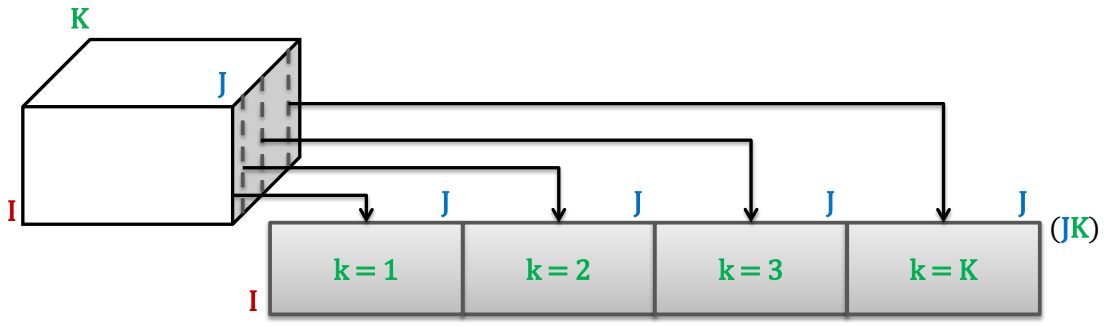


Figure 15. The principle of unfolding. Here the first mode is preserved while the second and third are confounded, i.e., the left-most matrix in the unfolded array is the $I \times J$ matrix equal to the first frontal slice ($k = 1$).

For a specific three-way problem, consider a hypothetical two-way matrix consisting of typical data with rows and columns equal to the first and second mode of the three-way array. For instance, if the array is structured as $sites \times species\ density \times time$, then

consider a matrix with sites in the row mode and species density in the column mode. Such a matrix could be adequately modeled by a bilinear model. Consider next a typical matrix of modes one and three ($sites \times time$) as well as mode two and three ($species\ density \times time$). If all these hypothetical two-way problems are adequately modeled by a bilinear model, then likely, a three-way model will be suitable for modeling the three-way data. Though the problem of deciding which model to use is complicated, this rule of thumb does provide rough means for assessing the appropriateness of multi-way models for a specific problem (Bro, 1998).

4.3 TUCKER3 MODEL

Tucker (1966) was the first to propose what he called “three-mode factor analysis”, but what is now called three-mode principal component analysis. The proposed method is an algebraic solution in the sense that if an exact solution exists, this method will produce it. An approximate solution is obtained when not all components are required that are necessary to produce an exact solution (Kroonenberg, 2008). In the following, the main elements of the theory about the Tucker3 model are briefly recalled. Such model is a three-way components analysis of a three-way data set. Such data can, for instance, consist of scores of a number of subjects on a number of variables, measured in a number of different conditions. The aim of Tucker3 analysis is to summarize the three sets of entities constituting the three-way data set in such a way that the main information in the data can be summarized by means of a limited number of components for each set of entities. It is therefore a generalization of two-way principal components analysis (PCA) (Fig. 16).

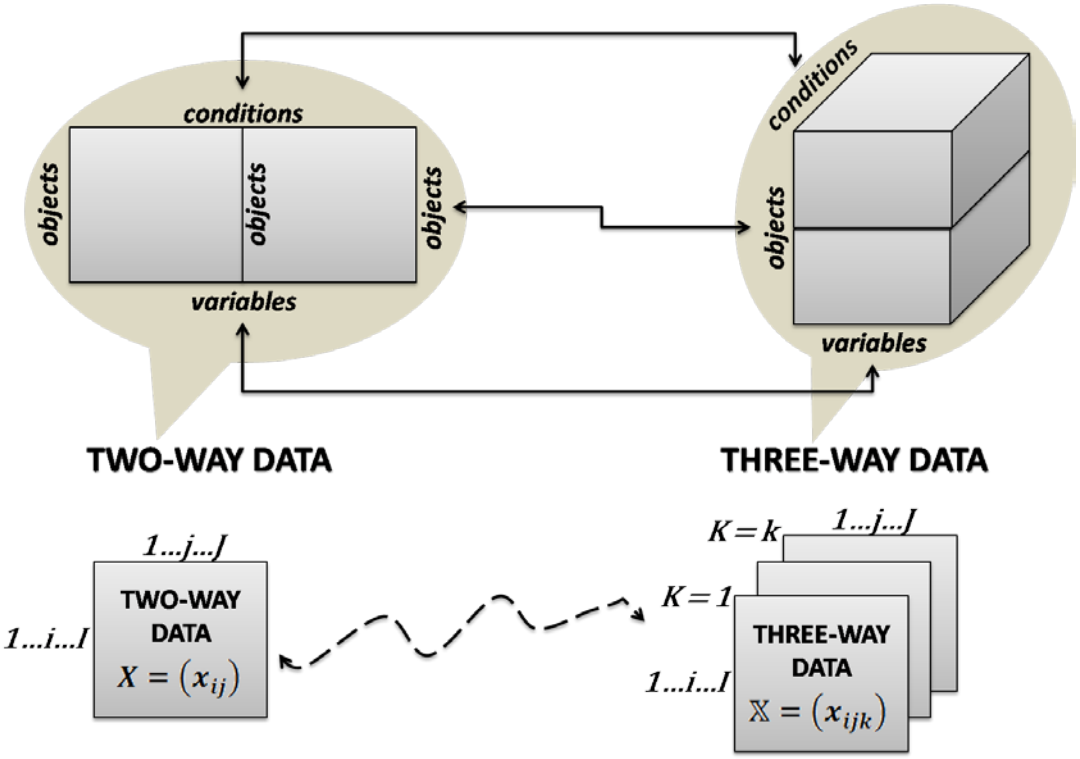


Figure 16. Extending two-mode PCA to three-mode PCA.

4.3.1 DESCRIPTION

The Tucker3 model (Tucker, 1966) is the decomposition of the three-way data array $\mathbb{X} = (x_{ijk})$, where it is assumed that first mode consists of subjects, the second of variables, and the third of conditions (Fig. 17). This decomposition has the form (Kiers and der Kinderen, 2003; Kroonenberg, 2008; Barbieri et al., 1999),

$$x_{ijk} = \sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{r=1}^{R} a_{ip}b_{jq}c_{kr}g_{pqr} + e_{ijk}, i = 1, \dots, I;\ j = 1, \dots, J;\ k = 1, \dots, K \qquad (4.3.1.1)$$

where the coefficients a_{ip}, b_{jq} and $c_{kr}$are the elements of the component matrices A, B and C, respectively, the g_{pqr} are the elements of the three-way core array $\mathbb{G}$, and the e_{ijk} are the errors of approximation collected in three-way array $\mathbb{E}$. A is the $(I \times P)$ matrix with the coefficients of the subjects on the subject components. B is the $(J \times Q)$ matrix with coefficients of the variables, and C is the $(K \times R)$ coefficient matrix of the conditions. The values p, q, r are the number of components selected to describe the first, second and third mode of the data array, respectively, and are not necessarily the same for each mode.

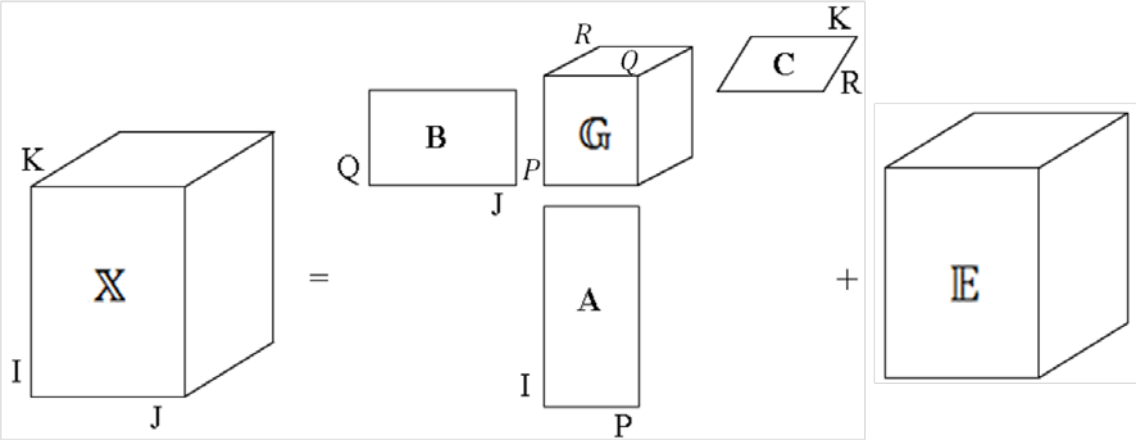


Figure 17. The Tucker 3 model scheme.

4.3.2 DATA PREPROCESSING

Data preprocessing consists in the application of procedures to a multiway data set before a multiway model is fitted. This step takes an important place in data analysis (Stanimirova et al., 2004; Bro and Smilde, 2003). Several types of data pretreatment are known, the most usual ones in two-way data analysis are centering and scaling. Centering removes the differences in the size of rows and/or columns. In row centering the corresponding row mean is subtracted from each element of the data matrix.

Column centering is done by subtracting from each data element the corresponding column mean. Centering can be done sequentially on rows and columns, which is known as double centering. The order of centering does not affect the final result. Scaling can also be done on rows and columns of the data matrix. Centering and scaling can be combined. This is the case for autoscaling, which is applied among the others when the variables are in different units. Its aim is to give the variables the same importance, by making the standard deviation of each variable equal to 1. These preprocessing methods can be extended to N-way arrays. The preprocessing of N-way data requires more caution, and several rules on how to do this can be found in the literature (Kroonenberg, 2008; Bro and Smilde, 2003). Single data centering is done across one mode and can be followed, if necessary, by sequential centering in other modes. Scaling in one mode will not change the data structure, whereas scaling to unit standard deviation in two modes is not possible. Combinations of centering and scaling for a given mode can be performed in the same way as for the two-way data. They are not problematic, when scaling within one mode is combined with centering across other modes (Bro and Smilde, 2003). Several scalings can be performed sequentially, in both preprocessing of two-way and preprocessing of N-way data, but will generally need iterations and may not converge. If centering is applied when this is not necessary (Bro and Smilde, 2003), it can contribute additional artificial variation, which will destroy the data structure and will lead to erroneous results obtained by both two-way and N-way methods (Stanimirova et al., 2004).

In the present work, and because most data treated here consist of what are called three-way profile data, the preprocessing will be addressed to the standard way to

preprocess such data, which is by removing the means per variable-time-point combination (jk) and normalising each variable (j) across all measurements of the variable ($k = 1, \dots K$):

$$\tilde{x}_{ijk} = \frac{\left(x_{ijk} - \bar{x}_{\cdot jk}\right)}{\sqrt{\sum_{ik} \frac{1}{IK} \left(x_{ijk} - \bar{x}_{\cdot jk}\right)^2}} \qquad (4.3.1.2)$$

This type of preprocessing usually is called profile preprocessing (Kroonenberg, 2008).

An important point to consider is that the decision to choose the "appropriate" preprocessing option mainly depends on the researcher's assessment of the origin of the variability of the data, that is, which means and which of variances that can be meaningfully interpreted.

4.3.3 PROPERTIES

4.3.3.1 DECOMPOSITION

The Tucker3 model possesses some of the properties of the SVD (for details see Section 1.2); in particular, it allows for a complete decomposition of the three-way array. It is also possible, given orthonormal components, to make a complete component-wise partitioning of the variability. This property extends to the partitioning per level of a mode and per element of the core array (details of this will be explored in Chapter 5, Section 5.2.5). The model has subspace uniqueness. Properties this model lacks are

component uniqueness (any transformation of the components and/or core array with the inverse transformation of the appropriate other matrices will not affect the fit) and nesting of solutions.

4.3.3.2 RESTRICTIONS ON THE NUMBER OF COMPONENTS

Even thought the number of components for the three modes may differ, there are some built-in restrictions in the model. In particular, the minimum-product rule says that the product of the number of the components of two modes must always be equal or larger than that of the third mode, so that $P \times Q \geq R$, $P \times R \geq Q$ and $Q \times R \geq P$. The restriction is a consequence of the three-way orthogonality of the core array.

Nearly all concepts mentioned in this chapter are treated in more technical detail (among others) in Smilde et al. (2004), Kroonenberg (1983, 2008), and Bro's thesis (1998).

CHAPTER 5

CHAPTER 5 – TUCKER-CO: A NEW METHOD FOR THE SIMULTANEOUS ANALYSIS OF A SEQUENCE OF PAIRED TABLES

5. 1 INTRODUCTION

Successful applications of STATIS family methods and Tucker3 model have been reported. Comparisons between three-mode principal component analysis, STATIS and various other three-way methods can be found in Peré-Trepat et al. (2007), Thioulouse (2010), Mechelen and Smilde (2011), Kroonenberg and ten Berge (2011), Sinha et al. (2009), Pardo et al. (2008), Stanimirova et al. (2004), Vivien and Sabatier (2004), Juan et al. (1998), Kroonenberg (1989, 1992), Carlier et al. (1989) and Kiers (1988), but there is a lack of comparative literature on the performance of some specific methods belonging to the STATIS family methods, more specifically, the PTA method and Tucker3 model. These comparisons play an important role given that, understanding the performance of the methods, allows a better choice and suitability to the problems raised by the investigators.

However these methodologies only take account a unique data structure, i.e., a single data cube. Hence, when the objective is to study a pair of tables (for example, sets of tables, repeated in time or space, to study the relationships between species and their environment) these methods are not able to do it.

Thus, the choice of methods that achieve this objective (i.e., the study of a pair of tables), leads researchers to work with data cube coupling methods (for details, see Chapter 3). Among these, stand out the Between-Group Co-Inertia Analysis (BGCOIA, Franquet et al., 1995), the STATICO method (Simier et al., 1999; Thioulouse et al., 2004) and more recently the COSTATIS (Thioulouse, 2010) (which means "CO-inertia and STATIS").

BGCOIA is a between-group Co-Inertia Analysis, considering each table (for example, date, site, condition, etc.) of the sequence as a group (that is, first the mean of the variables in each table is computed and arranged in two new tables and then a Co-Inertia Analysis is done on this couple of new tables).

On the other hand, STATICO is a PTA on the series of cross product tables obtained by crossing the two tables of a pair at each date (that is, first the k cross-covariance tables are computed from the two k-tables, resulting in a new k-table and then a PTA is done on this new k-table) (for details see Chapter 3, Section 3.3).

Finally, the COSTATIS is a co-inertia analysis of the compromises computed by the PTA of the two k-tables. Hence, in this technique, two PTA are used to compute the compromises of the two k-tables and then a Co-Inertia Analysis is used to analyze the relationships between these two compromises.

Of the three methods mentioned, and for the present work, the choice falls on the STATICO. This choice lies in the fact that STATICO is much valued by researchers, given the potential associated with it (since it provides a complete and consistent analysis

framework, with a much stronger mathematical background (Thioulouse et al., 2004), but also because it "only" displays the stable component of data relationship variations, without revealing what is different or interact (as the Tucker3 model puts in evidence through the core matrix and related joint biplots). On the other hand, three-mode models, as Tucker3 model, has for objective to represent measured data as a linear combination of a small number of optimal, orthogonal components following an appropriate generalization of principal component analysis applied to multi-way data arrays (for details, see Chapter 4). Thus, Tucker3 model can lead to an easier and more straightforward interpretation of the relevant information contained in a three-way data set. Actuality, one of the major advantages of modeling three-way data sets with three-mode models is the gain in interpretability.

Hence, what is proposed here, more than a simple comparison of methods, is a new technique (and alternative), the CO-TUCKER, that combines the STATICO (more specifically, the Co-Inertia Analysis) and Tucker3 to analyze two paired k-tables. In this sense, all the methodological bases will be given. Additionally, the efficiency of what one may expect by its exploration will be also provided by an application on a real data set.

Note that despite new the proposed technique, CO-TUCKER, is illustrated by a species-environment data set, other types of the three-way data can be handled in a similar manner. The option to use this type of data (i.e, ecological data) allowed proper comparison with the STATICO technique, whose use is widely held to ecological problems.

5.2 THE CO-TUCKER METHOD

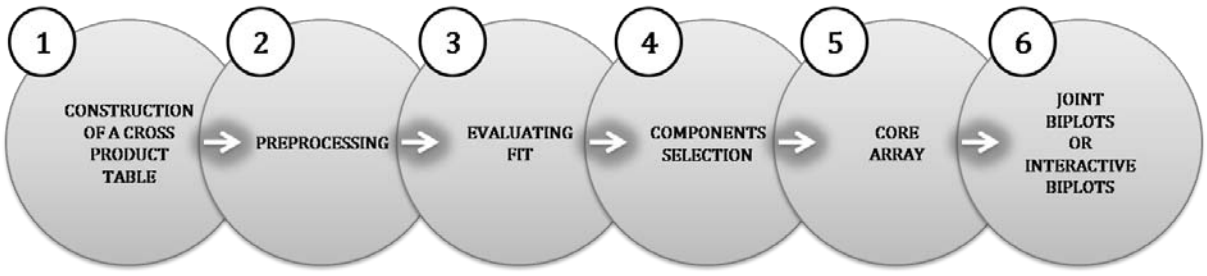
The CO-TUCKER is a combined algorithm of the Tucker3 performed on the series of cross product tables (the k-tables, W_k) each one calculated by Co-Inertia Analysis. Indeed, CO-TUCKER aims to analyze two cubes of paired data and will benefits from the advantages of both STATICO and Tucker3.

The proposed method, CO-TUCKER, consists in two main steps:

(1) in a first step, after the individual analysis of each matrix for each data cube (by means of a PCA, CA or MCA, as applicable), k Co-Inertia Analysis are computed in order to calculate the sequence the cross-covariance k tables (W_k), i.e., it consists in perform k Co-Inertia Analysis of k-cross product tables ($\left(W_k, D_I, D_J\right)$ where $W_k = Y_k^T D_{I_k} X_k$, and X_k is the first data set, Y_k is the second data set, $D_I(I \times I)$ and $D_J(J \times J)$ the diagonal matrices and $D_{I_k} = Diag(1/I_k)$ the weight matrix);

(2) in a second step, a Tucker3 model is used in order to analyze this new k-table or array ($\mathbb{W}$).

The CO-TUCKER general scheme is presented in figure 18 and can be schemed following the next steps:



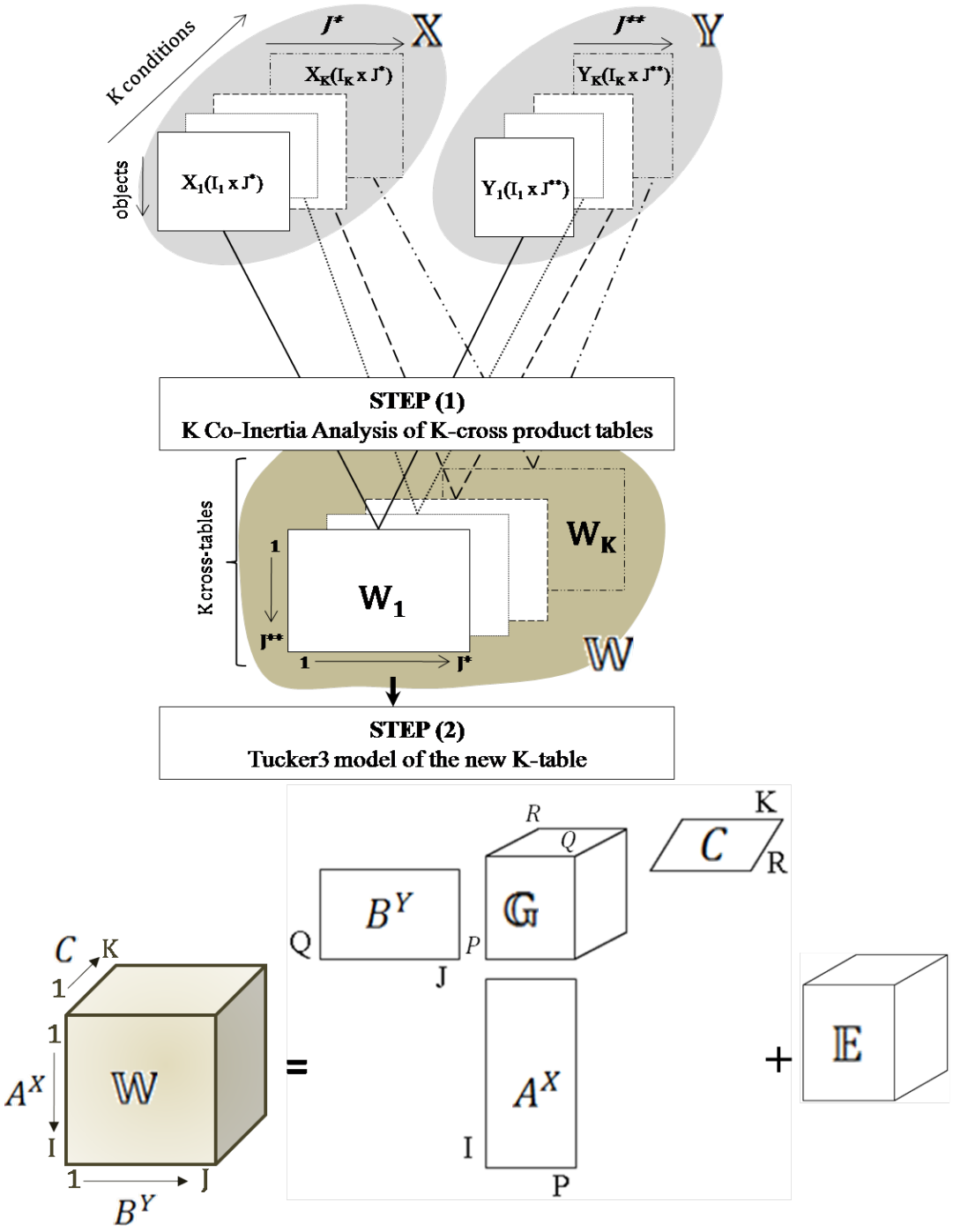


Figure 18. CO-TUCKER general scheme. The data structure is a sequence of K paired tables. X_k and Y_k are respectively the first group (dimension $I_K \times J^*$) and the second group (dimension $I_K \times J^{**}$) of tables in the pair. W_K is the cross-table at K occurences; J^* is the number of variables of $\mathbb{X}$, J^{**} is the number of variables of $\mathbb{Y}$, I_K is the number of rows at k different conditions. In the first step, after the individual analysis of each matrix for each data cube (by means of a PCA, CA or MCA, as applicable), K Co-Inertia Analysis are computed in order to calculate the sequence the cross-covariance K tables (W_K); then, in a second step, a Tucker3 model is used in order to analyze this new k-table or array ($\mathbb{W}$).

As mentioned, the first step of CO-TUCKER consists in performing a Co-Inertia Analysis, i.e., it consists in the analysis of a cross product table, and its triplet notation is: $\left(W_k, D_I, D_J\right)$, where $W_k = Y_k^T D_{I_k} X_k$.

The second step is to analyze these matrices by means of Tucker3 model. This means that the k tables (or matrices, which represents k conditions) have I rows and J columns. These k matrices, W_k, are decomposed as shown in figure 18, and it can be formulated as a factorization of the three-way data matrix, $\mathbb{W}(I \times J \times K)$, such that:

$$w_{ijk} = \sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{r=1}^{R} a_{ip}^X b_{jq}^Y c_{kr} g_{pqr} + e_{ijk}\,, i = 1, \dots, I;\ j = 1, \dots, J;\ k = 1, \dots, K \qquad (5.2.1)$$

where, by a commonly-used formulation for a Tucker3 model applied on the three-way array (for details see Chapter 4, Section 4.3.1):

- the a_{ip}^X, b_{jq}^Y and c_{kr} are the elements of the $A^X(I \times P)$, $B^Y(J \times Q)$ and $C(K \times R)$ loading matrices, respectively, and p, q, r denote the number of components in the modes one, two and three[7];
- the e_{ijk} is an element of the three-way residual matrix $\mathbb{E}$, which denotes an error term associated with the description of w_{ijk} (which is the value of the measurement referring to the ith subject, jth variable and kth condition);

[7] The use of the notation a_{ip}^X and b_{jq}^Y (analogous to $A^X(I \times P)$ and $B^Y(J \times Q)$) is adopted to identify the elements that come from matrices X and Y, respectively.

- the elements g_{pqr} weights the products between component p of the subjects (first mode), component q of the variables (second mode), and component r of the different conditions (third mode) and explains the interaction among p, q, r factors of each of the modes. Those elements are stored in the core array $\mathbb{G}$, an array of dimensions $(P \times Q \times R)$. This matrix can be considered as a generalization of the diagonal matrix of the eigenvalues obtained from the SVD of a two-way matrix (note that, SVD forms the basis of the generalization to methods for multiway arrays). Moreover, the core array $\mathbb{G}$ is derived from the three components matrices $A^X(I \times P)$, $B^Y(J \times Q)$ and $C(K \times R)$ as follows:

$$g_{pqr} = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} a_{ip}^{X}\, b_{jq}^{Y} c_{kr} w_{ijk} \tag{5.2.2}$$

Generally, each of the A^X, B^Y and C matrices can be interpreted as loading matrix in the classical two-way PCA and are constrained to be orthogonal and the columns of the matrices are scaled to have unit length. In this way, the magnitude of the squared element of the core (g_{pqr}^2) indicates as to what is the importance of the interaction between the components p, q and r, in the model of $\mathbb{W}$. The core efficiently describes the main relations in the data, and the component matrices A^X, B^Y and C describes how the particular subjects, variables and conditions relate to their associated components. In addition, the product of the total number of components in each different mode $(P \times Q \times R)$ can be used as an indication of the number of possible interactions, and hence the complexity of the model. Note that, in the original data array $\mathbb{W}$ every element of the matrix represents the value of a specific combination of levels of the original modes then, in a similar way, each element of the core array represents the value,

(mutual) weight, or interaction of a specific combination of the components of the modes.

Equation (5.2.1) for three way unfolded data array, $\mathbb{W}(I \times JK)$, can also be written as:

$$\mathbb{W}^{(I\times JK)} = \mathrm{A}^{\mathrm{X}}\left(\mathbb{G}^{(P\times QR)}\right)(C \otimes B^{Y})^{T} + \mathbb{E} \tag{5.2.3}$$

where $\otimes$ is the Kronecker product[8], $\mathbb{E}$ is the matrix of model residuals, and $\mathbb{G}^{(P\times QR)}$ is the core array, $\mathbb{G}(P \times Q \times R)$ arranged as $P \times QR$ (Fig. 18).

Also, equation (5.2.2), can also be written as:

$$\mathbb{G}^{(P\times QR)} = (\mathrm{A}^{\mathrm{X}})^{\mathrm{T}}\left(\mathbb{W}^{(I\times JK)}\right)(C \otimes B^{Y}) \tag{5.2.4}$$

where $\mathbb{G}^{(P\times QR)}$ is the $P \times Q \times R$ core array $\mathbb{G}$ matricized to a $P \times QR$ matrix and $\mathbb{W}^{(I\times JK)}$ is the data array matricized to an $I \times JK$ matrix. If the fit is perfect, then $\mathbb{G}$ contains exactly the same information as $\mathbb{W}$ merely expressed using different coordinates.

One of the interesting properties of the CO-TUCKER model (as in Tucker3 model) is that the number of components for the different modes does not have to be the same. Furthermore, the original core ($\mathbb{G}$) may be rotated to optimize for super diagonality and to maximize the core variance in order to obtain a matrix with a limited number of

[8] The Kronecker product ($\otimes$) of a $I \times P$ matrix $A^{X} = (a_{ip}^{X})$ and $J \times Q$ matrix B^{Y} is the $IJ \times PQ$ matrix composed of $J \times Q$ submatrices of the form $a_{ip}^{X}B$.

elements having high absolute values, thus making the relation between them more understandable.

A limitation of this model (that arrives from Tucker3 model; see details in Section 4.3.1) is that the solution obtained is not structurally unique (due to its rotational freedom). Infinity of other equivalent solutions can be obtained by rotating A^X, B^Y, C and $\mathbb{G}$ matrices without changing the fit of the model. This emphasizes for straightforward interpretable core array obtained for the selected model of complexity. Finally, the components in each mode are constrained to orthogonality, which implies a fast convergence.

Since Tucker3 is a multilinear model (Bro and Andersson, 1998b; Estienne et al., 2001; Bro and Andersson, 1998a; Kroonenberg, 1989; Kroonenberg and De Leeuw, 1980; Kroonenberg, 1983, 2008; ten Berge et al., 1987; Paatero and Andersson, 1999; Barbieri et al., 1999; Kroonenberg and Brouwer, 1985), the matrices A^X, B^Y, C and $\mathbb{G}$ are built simultaneously during the Alternating Least Squares (ALS)[9] fitting process of the model in order to account for the multidimensional structure.

Hence, the basic algorithm presented here (using the appropriate notation) uses the Kroonenberg and De Leeuw (1980) approach to solving the estimation of the parameters.

[9] Most of the details of the procedure follows Kroonenberg (1989), were first described in Kroonenberg and De Leeuw (1980) and Kroonenberg (1983) and it is embodied in the program TUCKALS3 (Kroonenberg and Brouwer, 1985).

In the program, PCA is performed for each way after stringing-out the $I \times J \times K$ data matrix W in three different ways, i.e. as a $I \times JK$ matrix, a $J \times KI$ matrix, and a $K \times IJ$ matrix. The estimation procedure is rather involved as the least squares loss function $\|\mathbb{W} - A^X \mathbb{G}[C^T \otimes (B^Y)^T]\|$ ($\otimes$ is the Kronecker product) has to be minimized over the component matrix A^X, the variable component matrix B^Y, the component matrix C, and the core matrix $\mathbb{G}$. This minimization is carried out via an ALS procedure. The basic idea of this estimation method is that conditional on the other matrices, say $\mathbb{G}$, C, and B^Y, the remaining component matrix, A^X, can be found via a least squares procedure. By estimating the matrices one at a time and alternating, a series of least squares problems can be solved, of which it can be shown that they converge to a (local) minimum.

The estimation proposal proceeds as outlined in next scheme (Fig. 19):

SET INTERACTION COUNTER $\alpha = 0$

STARTING SOLUTION

Initialize A_0^X, B_0^Y, C_0 (Tucker (1966), Method I) (*)

⇓

UPDATE INTERACTION COUNTER $\alpha = \alpha + 1$

A^X-SUBSTEP

(1) Fix $B_{\alpha-1}^Y$ and $C_{\alpha-1}$

(2) Solve an eigenvalue-eigenvector problem for A^X to obtain a new A_α^X

⇓

B^Y-SUBSTEP

(1) Fix A_α^X and $C_{\alpha-1}$

(2) Solve an eigenvalue-eigenvector problem for B^Y to obtain a new B_α^Y

⇓

C-SUBSTEP

(1) Fix A_α^X and B_α^Y

(2) Solve an eigenvalue-eigenvector problem for C to obtain a new C_α

⇓(**)

CHECK CONVERGENCE

If

the differences between successive interactions with respect to the loss function $\|\mathcal{F}^{\alpha-1} - \mathcal{F}^{\alpha-1}\|^2$ are not small enough

or

the Euclidean norms of successive estimations of A^X, B^Y and C are not to small enough

then

continue with next iteration

else

end iterations

⇓

COMPUTE CORE ARRAY

(1) Use equation (5.2.2), $g_{pqr} = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} a_{ip}^X \; b_{jq}^Y \; c_{kr} \, w_{ijk}$

(2) Calculate the core array $\mathbb{G}$ from the data and the estimated A^X, B^Y and C

Figure 19. Iterative algorithm step of CO-TUCKER model.

(*) To initialize the algorithm A_0^X, B_0^Y, C_0 are chosen in such way that they will fit the model exactly if such an exact solution is available

(**) ESTIMATE MISSING DATA STEP: an optional step introduced, which uses the expectation-maximization (EM) approach (Dempster et al., 1977). It uses the current estimates A_α^X, B_α^Y and C_α and the data $\mathbb{W}$ to estimate the missing values according to the model equation (5.2.1).

So, CO-TUCKER is computed by iterative algorithms, and thus the solution allows the partitioning of the sum of squares of $\mathbb{W}$ as

$$SS(TOT) = SS(FIT) + SS(RES) \tag{5.2.5}$$

This means that the total sum of squares or variation ($SS(TOT)$) of the data can be divided into a fitted ($SS(FIT)$) and a residual part ($SS(RES)$).

The ratio,

$$SS(FIT)/SS(TOT) \tag{5.2.6}$$

can be used to evaluate the descriptive performance of the model.

Moreover, the ratio

$$SS(RES)/SS(TOT) \tag{5.2.7}$$

is the proportional or relative lack of fit of a level of a mode.

By comparing the relative residual sum of squares (the relative residual) within and across models, one can gauge to what extent the model adequately fits individual parts of data. Alternatively, such an investigation can demonstrate that certain levels of a mode are dominating the solution more than is acceptable, given their role in the investigation (Kroonenberg, 2008). Because not only the relative residual sum of squares but also the total sum of squares is known, it is possible to distinguish between two cases when the relative residual is large:

(1) when a large relative residual coupled with a small total sum of squares;

(2) when there is a large amount of variability but is not fitted very well.

In the first case does not raise great cause for concern, since this means that probably the level was not very important in terms of variability and did not contribute much anyway. However, for the second case, this means that there is a considerable variability that the model could not capture. The causes for this situation may be, among others, the fact that the model does not have enough components, the data do not conform to the model, or there is simply a large amount of random noise.

5.2.2 DATA PREPROCESSING

Special attention should be given to the preprocessing of data. In most studies carried out by use of the STATIS family methods, the preprocessing of the data consists in scaling the variables (mode two) across objects (mode one) at each condition (mode three). Their practice removes the condition to condition variability which might not be advisable, because this variability is something that should be explained as well. With this scaling, the scores are measured in different standard units across conditions which make them less comparable than might be desirable. It should be noted that the variability between conditions differ considerably, so that different scaling may lead to different results.

However, the preprocessing of N-way data requires more caution, and several rules on how to do this can be found in the literature (Stanimirova et al., 2004; Bro and Smilde, 2003). Single data centering is done across one mode and can be followed, if necessary, by sequential centering in other modes. Scaling in one mode will not change the data structure, whereas scaling to unit standard deviation in two modes is not possible.

Combinations of centering and scaling for a given mode can be performed in the same way as for the two-way data. They are not problematic when scaling within one mode is combined with centering across other modes (Stanimirova et al., 2004; Bro and Smilde, 2003). Several scalings can be performed sequentially, in both preprocessing of two-way and preprocessing of N-way data, but will generally need iterations and may not converge. If centering is applied when this is not necessary, i.e. when there is no offset to be corrected (Stanimirova et al., 2004; Bro and Smilde, 2003), it can contribute additional artificial variation, which will destroy the data structure and will lead to spurious results obtained by both two-way and N-way methods. Thus, in conclusion, for all the reasons mentioned and especially for this type of data, the profile data, the scaling of the variables across all objects and conditions should be reflected in order to assess whether expected (or not) be taken into account.

Nevertheless, whether by one or by another method, the preprocessing of data always depends on the problem at hand.

Yet, for the specific case of CO-TUCKER when prior to calculations of Co-Inertia Analysis, the columns of both tables (group one and two of matrices, i.e., X_k and Y_k) are centered (or have another kind of preprocessing), then when executing the model no preprocessing method is used before the data analysis.

5.2.3 COMPONENTS SELECTION

Similarly to what happens in Tucker3 model, also in CO-TUCKER choosing the numbers of components to use and thus select an "optimal" model in terms of complexity, assumes a very important step. Hence, the process of components selection in CO-TUCKER followed the methods already developed for the Tucker3 model.

In CO-TUCKER, and in contrast to PCA, solutions with different numbers of components cannot be derived from each other. Therefore, the numbers of components can vary independently for each mode, and for each combination of numbers, a new analysis has to be carried out. So, a strategy could be to carry out all conceivable analyses, compare all results, and then choose the most useful solution, where useful can be defined, for instance, in terms of parsimony, interpretability or stability. However, such an approach of comparing full solutions is practically unfeasible (Kiers and der Kinderen, 2003). Generally, and in accordance to what happens in the Tucker3 model, the optimal complexity is the one requiring the smallest number of components capturing a high fraction of data variance. In practice, a compromise between models describing a reasonable percentage of variation and models with fewer components is necessary. Note that the chosen model must be a reasonable model to explain patterns within and between modes (Stanimirova et al., 2006; Giussani et al., 2008; Kroonenberg, 2008). Moreover, the algorithm's convergence criterion of Tucker3 model is associated with stability in the solution which implies that a local, but not an optimal solution is obtained.

One way to select the "optimal" model is to compute the $DifFit$-criterion (developed for Tucker3 model by Timmerman and Kiers (2000)), which computes the fit for models with all sensible combinations of numbers of components, up to a chosen maximum value. Thus, the method propose select the optimum number of components by calculating the adjustment values for all possible solutions obtained from the model algorithm.

In this approach, a model is fitted to each solution to approximate the values of the three-way data array (here the three-way data array $\mathbb{W}$). Thus, an error and an explained value are associated with each solution, and the fit is precisely the same as that part of $\mathbb{W}$ that is explained by each solution. According to Timmerman and Kiers (2000), and in correspondence to Tucker 3 model, $P \leq QR$, $Q \leq PR$ and $R \leq PQ$ must hold for all possible solutions. It should be noted that this condition eliminates redundant solutions.

So, given a model with the same total number of components ($S = P + Q + R$) (following Kroonenberg (2008) and Timmerman and Kiers (2000), and according to the related concepts already demontrated for the Tucker3 models), a model that has the highest proportion of fitted sum of squares, $SS(FIT)$ (or equivalently the smallest residual sum of squares) is selected. Then, and to compare classes with different S, the difference

$$dif_s = proportional\, SS(FIT)_s - proportional\ \ SS(FIT)_{s-1} \qquad (5.2.3.1)$$

is computed (that is, the gain in fit by increasing the number of components in each mode is calculated). Only those differences, that are sequentially highest, are considered. Moreover, the differences in fit between one solution and another play the role of the eigenvalues in a two-way principal components analysis, which were ranked in descending order; this means that the gain in fit from one solution to another becomes smaller and smaller.

In the next step, the salience value (b_s), defined as a ratio between dif_s and dif_{s*} (where dif_{s*} has the next highest value after dif_s) is worked out. The $DifFit$-criterion ends designating that the chosen model is one that has the highest salience value. In other words, this step is to determine a subset of solution(s) for which all later solutions had a smaller difference in fit. Finally, a lower bound for dif_s is defined and models with values below that should be avoid.

The critical value is defined by

$$\frac{1}{[S(Min) - 3]} \tag{5.2.3.2}$$

where $S(Min) = MIN(I, JK) + MIN(J, KI) + MIN(K, IJ)$, and I, J, K the numbers of levels of the first, second and third mode, respectively (as in Tucker3 model, Timmerman and Kiers (2000)).

Thus, the method proposed by Timmerman and Kiers (2000) is based on eliminating solutions until an optimal solution is obtained, and can be outlined as follows:

(1) in the first step, all redundant solutions are eliminated, that is, all solutions for which there is one solution with a smaller total number of components (S) that explain the same amount of variability are discarded;

(2) in the second step, all solutions that do not achieve the condition by which increasing the number of components produces a smaller gain in fit (smaller variance explained) are discarded, as compared with the previous solution with fewer components.

This procedure allows finding the optimum balance between the number of components retained in each mode and the variability explained by the model (Timmerman and Kiers, 2000, Kroonenberg, 2008).

To visualize the procedure, it is useful to build the three-mode scree plot (Kroonenberg, 2008), in which the residual sum of squares is plotted against the total number of components S. In fact, the $DifFit$-criterion is essentially a way to define the convex hull in the three-mode scree plot, and models that are taken into consideration are those on the convex hull. Figure 20a shows an example of a three-mode scree plot.

One of the disadvantages of the Timmerman and Kiers (2000) procedure is that it is rather time-consuming. This is because each analysis is an iterative process of adjustment, which also is repeated starting from different initial configurations several times to avoid ending up in local minima (Kiers and der Kinderen, 2003). In this sense, Kiers and der Kinderen (2003) proposed an alternative procedure, which basically consists in a modified $DifFit$-criterion by using the approximate fit values rather than

optimal fit values. Specifically, in their work it is proposed to replace these fit computations by a procedure for computing approximate fit values. The authors showed that it is sufficient to approximate the models via a single computation using Tucker's original non least-squares method, and to calculate the residual sum of squares for the models under consideration using the core array. The alternative method gives solutions for all combinations of numbers of components in one go, making it considerably more efficient and thus an enormous saving in computer time (Kiers and der Kinderen; 2003, Kroonenberg, 2008).

Another alternative to the $DifFit$-criterion is the evaluation of the residual sum of squares (or deviance, d) together with the degrees of freedom (df). Again a plot can be constructed with the residual sum of squares of each model plotted versus the degrees of freedom (called as deviance plot). Figure 20b shows an example of a three-mode deviance plot.

The general idea is to drawn straight lines between the *loci* of models with constant d/df. If the origin is included in such a plot, these lines can be seen to fan out from the origin with a slope of d/df. If $d/df = 1$, then each parameter added or degree of freedom lost leads to a gain of one unit of deviance. If d/df is large, then, in exchange for a few degrees of freedom, one can get a large decrease in deviance or a considerably better fitting model. On the other hand, if d/df is small, one needs to sacrifice many degrees of freedom to get a reduction in deviance or an increase in fit. The principle of parsimony, which can be expressed here as a preference for models with good fit and

few parameters or many degrees of freedom, suggests that models in bottom right corner of the plot are the first to be taken into consideration for selection.

Finally, Ceulemans and Kiers (2006) proposed a numerical heuristic procudere to select models in deviance plots (Kroonenberg, 2008). The procedure called as the st-criterion is defined as:

$$st_{𝕚} = \left(\frac{d_{𝕚-1} - d_{𝕚}}{df_{𝕚-1} - df_{𝕚}}\right) \Big/ \left(\frac{d_{𝕚} - d_{𝕚+1}}{df_{𝕚} - df_{𝕚+1}}\right) \tag{5.2.3.3}$$

where $d_{𝕚}$ is the deviance and $df_{𝕚}$ the degrees of freedom of model 𝕚.

The process of selection using the st-criterion can be resumed as (Ceulemans and Kiers, 2006):

(1) determine the df and d values of all component solutions from one wishes to choose;

(2) for each of n observed df values, retain only the best fit solution;

(3) sort the n solutions by their df values and denote them by $s_{𝕚}$ ($𝕚 = 1, \dots, n$);

(4) exclude all solutions $s_{𝕚}$ for which a solution $s_{𝕛}$ ($𝕛 < i$) exists such that $d_{𝕛} < d_{𝕚}$;

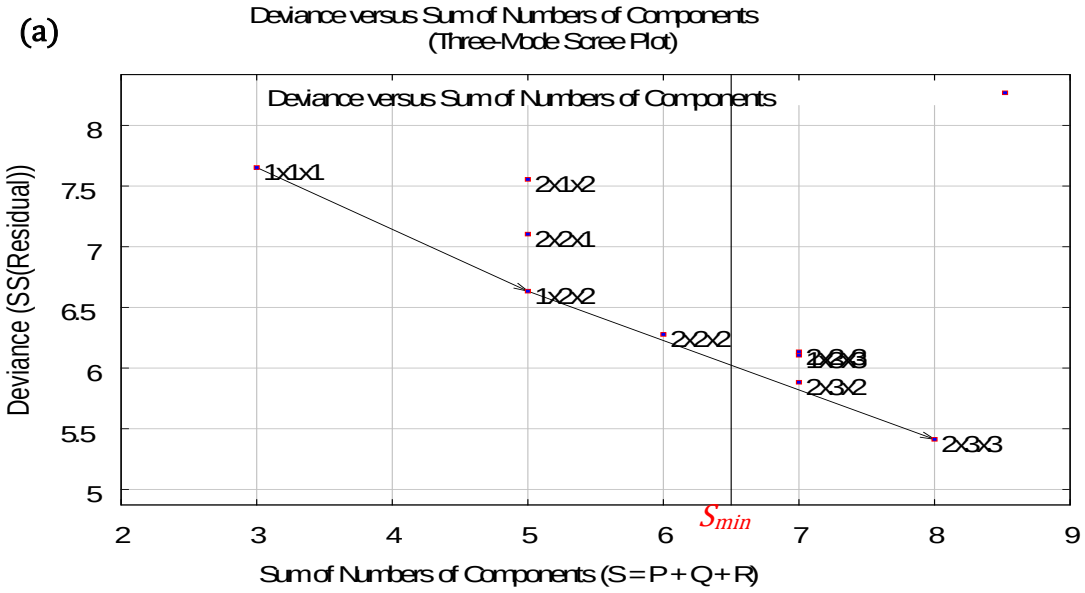
(5) consider all triplets of adjacent solutions and exclude the middle solution if its point is located above or on the line connecting its neighbors in the deviance plot;

(6) repeat the last step until no solution can be excluded;

(7) determine the *st* values of the convex hull solutions obtained;

(8) select the solution with the smallest *st* value.

In summary, steps (1)-(6) serve to determine which models lies on the convex hull, and steps (7)-(8) serve to find smallest angle $\measuredangle_{i-1,i+1}$ between the lines connecting a model i with its previous $i-1$ and subsequent $i+1$ models, as illustrated in figure 20b.



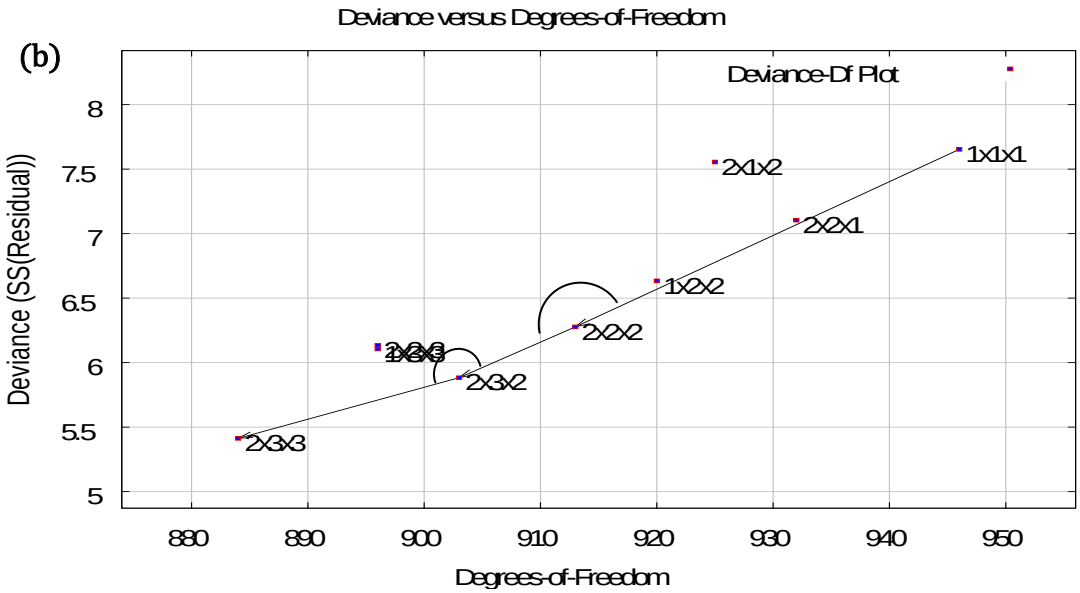


Figure 20. Scree plots produced for a hypothetical example model less than or equal to the $2 \times 3 \times 3$-mode. (a) Three-mode scree plot. The lower bond, S_{min}, has been drawn as a vertical line and all the models to the right of that line fail to make the minimum value (equivalent to the eigenvectors with eigenvalues smaller than 1 in the two mode case) (b) Three-mode deviance plot. Both graphs show that the $2 \times 3 \times 3$-model and the $2 \times 3 \times 2$-models are feasible, but a $2 \times 2 \times 2$-model is more economical. Actually the three-mode scree plot suggests that the first two models are overfitting the data. The st-criterion of Ceulemans and Kiers (2006) suggests that the $2 \times 3 \times 2$-model is the model of choice as the convex hull bends at that point (note that for the $2 \times 2 \times 2$-model a near 180° $angle$ is showed and the $2 \times 3 \times 2$-model $angle < 2 \times 2 \times 2$-model $angle$; in such case, it might be advantageous in practice to select the less parsimonious model $2 \times 3 \times 2$, which provides more detailed information about the data)

A model selection should not be followed too rigorously and substantive criteria can also play a role in a selection. Thus, the choice of a particular model should not only be based on numerical arguments and investigations, but also on content-based arguments. Proper regard should be paid to the information from all parts of the process of selection, so that a reasoned decision can be made. The interpretational qualities of a model (independently if it is the "optimal" model) should be considered as an important detail in the selection process.

5.2.4 EVALUATING FIT

To assess the quality of the fit of the levels of a mode, it is useful to observe at their residual sums of squares in conjunction with their fitted sums of squares (as illustrates the example in figure 21).

It is the major tool for evaluating structured residuals and it has for each level the $SS(RES)$ on the vertical axis and the $SS(FIT)$ on the horizontal axis. From the plot it is possible gauge to what extent the model succeeds in fitting the variability of a level.

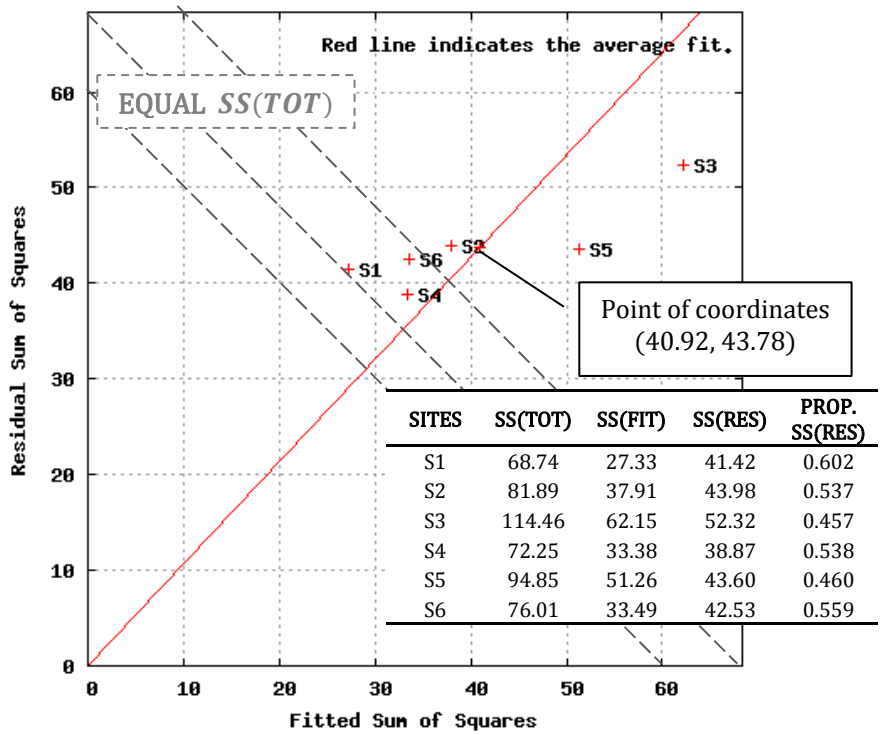
A standard reference line is drawn from the point (0,0) through the point with the average $SS(FIT)$ and the average $SS(RES)$. Points above the line fit worse compared to the overall fit and the points below the line fit better than the overall model. The further away are from the line, worse (or better) is the fit. So, by the sum-of-squares plots it is possible to assess the fit or lack of it for each level of each mode separately.

Additionally, by the inspection of $SS(TOT)$ of the elements per mode it is possible to detect elements with very large influence on the overall solution (i.e., with large $SS(TOT)$ values) and elements that did not play a role in the solution (i.e., with small $SS(TOT)$ values). As a result, the further the levels of a mode are toward the northeast corner of the plot, the higher their total sums of squares. Thus, levels with large amounts of variability can be immediately spotted from their location on the plot. On the other hand, if a level has such a comparatively large relative sum of squares (i.e., $SS(FIT)/SS(RES)$), it will be located in the southeast corner.

In particular, large values indicate that the level contains more fit than error information, and vice-versa. Also, the relative differences between residual sums of squares can be gauged by the investigation of the relative residual sums of squares (i.e., $SS(RES)/SS(TOT)$). So, if a level has a high relative residual, a limited amount of the data is fitted by the model.

Also a negatively sloping dashed line (and all lines that could be drawn parallel to it) can be drawn. It represents points with an equal $SS(TOT)$, and the further away a line is from origin, the larger total sum of squares. In conclusion, the $SS(RES)$ and $SS(FIT)$ as well as their relationships can be shown directly in the sums-of-squares plot.

Figure 21 will be used to illustrate the sums-of-squared plots interpretation.



SITES	SS(TOT)	SS(FIT)	SS(RES)	PROP. SS(RES)
S1	68.74	27.33	41.42	0.602
S2	81.89	37.91	43.98	0.537
S3	114.46	62.15	52.32	0.457
S4	72.25	33.38	38.87	0.538
S5	94.85	51.26	43.60	0.460
S6	76.01	33.49	42.53	0.559

Figure 21. Sums-of-squares plot of a hypothetical example corresponding to the sites mode of a 2 × 3 × 3 model. S1 to S6 represents six sampling sites.

In the preceding figure (Fig. 21) the red line indicates the average fit. Lines at an angle of -45 degrees are *loci* of points with equal $SS(TOT)$. The line from the point (0,0) through the point with coordinates (40.92, 43.78) (that marks the average fit) separates the ill-fitting points (above the line, i.e., S1, S2, S4 and S6) from the well-fitting points (below the line, i.e., S3 and S5, with a fit/residual ratio of 1.19 and 1.18, respectively). Sampling site S3 presented the largest $SS(TOT)$ (114.46), and S1 has the smallest (68.74). Since S1, S2, S4 and S6 are ill represented in the solution then only very partial statements can be made about their influence in the study.

5.2.5 CORE ARRAY

The core array $\mathbb{G}$ can be considered as a generalization of the matrix of eigenvalues that is associated with two-way SVD with the difference that the components can be related[10] (Fig. 22). The core array represents the value by which the single component product is weighted. Therefore, the value and the sign of each core element, give information about the entity of the interaction among the components of the different modes[11] (for details see comprehensive explanation and example in Section 5.2.8.1).

[10] Recall Section 5.2, equation (5.2.2), $\mathbb{G}^{(P\times QR)} = \left(\mathrm{A}^{\mathrm{X}}\right)^{\mathrm{T}}\left(\mathbb{W}^{(I\times JK)}\right)(C \otimes B^{Y})$ where $\mathbb{G}^{(P\times QR)}$ is the $P \times Q \times R$ core array $\mathbb{G}$ matricized to a $P \times QR$ matrix and $\mathbb{W}^{(I\times JK)}$ is the data array matricized to an $I \times JK$ matrix. If the fit is perfect, then $\mathbb{G}$ contains exactly the same information as $\mathbb{W}$ merely expressed using different coordinates.

[11] The meaning of the interaction is here obtained as proposed by Kroonenberg (1983), Henrion and Andersson (1999), Kroonenberg (2008) and Kroonenberg and Röder (2008), and followed by several authors (see for example, Sinha et al. (2009), Singh et al. (2007), and Astel et al. (2010)) for the Tucker3 model in general. Note that, the sign of g_{pqr} is determined by the signs of the p elements of the first mode, q elements of the second mode and r elements of the third mode. It consists by multiplying the signs of the elements along all components and the core element.

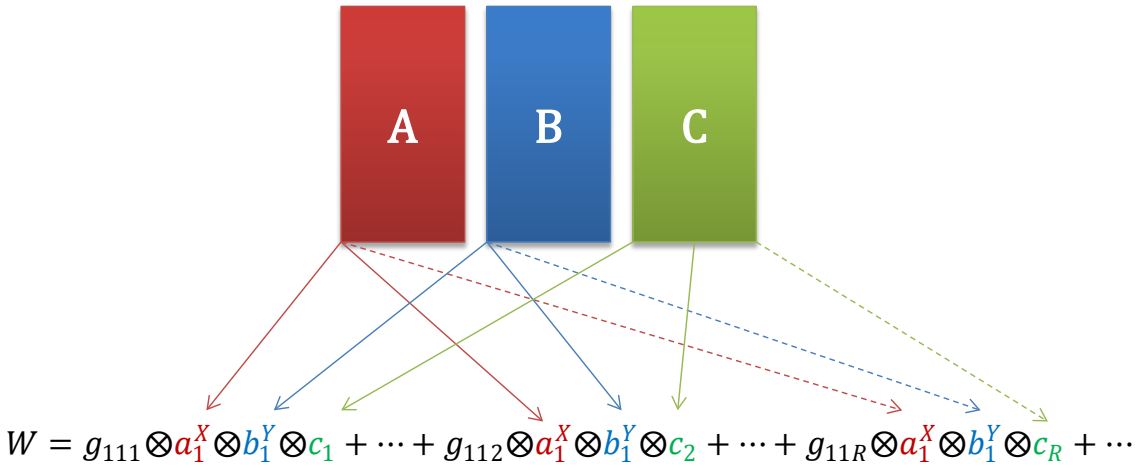


Figure 22. The CO-TUCKER core array scheme ($\mathbb{G}$). The core array $\mathbb{G}$ is implicitly given by the loading matrices. In the CO-TUCKER there are interactions between every component vector in every mode. An individual element of the core array, g_{pqr}, gives the importance of the specified combination of components.

Thus, the quantity,

$$\left(g_{pqr}\right)^2 \Big/ \sum_{pqr}\left(g_{pqr}\right)^2 \qquad (5.2.5.1)$$

represents the part of the data variability that is explained by the analysis and attributed to the p, q and r combination of components (recall that, the core elements are computed as $g_{pqr} = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} a_{ip}^X\, b_{jq}^Y c_{kr} w_{ijk}$; for details, see Section 5.2, Equation (5.2.2)).

As the squared elements of the core array are proportional to the variation explained by the combination of the factors corresponding to their indices, then when an element g_{pqr} is the largest core element, special attention in interpreting the model has to be given to the interaction between component p of mode one, component q of mode two and component r of mode three. So, the largest elements of the core will indicate what

the most important factors are in the model of the three-way data array (that is for CO-TUCKER, $\mathbb{W}$).

Similarly, when the objective is to know how much variability is captured by a specific component, simply add all the values of $\left(g_{pqr}\right)^2$ while keeping the index associated with the analyzed component fixed. That is, for example, to find the importance of component p of the first mode, simply calculate the quantity

$$\sum_{q=1}^{Q}\sum_{r=1}^{R}\left(g_{pqr}\right)^2 \Big/ \sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{r=1}^{R}\left(g_{pqr}\right)^2 \qquad (5.2.5.2)$$

However, in agreement to what arise for Tucker3 model, this is only meaningful if the factors in the component matrices A^X, B^Y and C are limited in some way. Otherwise the factors in A^X (for example) could be scaled by random scalars and the corresponding elements in the core would be down-scaled by the same scalars thereby not providing any meaningful basis for interpretation.

In order to make the weights in $\mathbb{G}$ comparable it is therefore generally applied to scale the factors to have unit length, i.e., the root of the sum of squared elements equals one, such that the core elements represents the importance of the respective factor combinations. Furthermore, by requiring that the component matrices are orthogonal, it can be shown that, when the model fits data ideally, the sum of the squared entries of the core equals the sum of the squared entries of $\mathbb{W}$.

As a final point, a core array is superdiagonal if it is a hypercube and only the elements with the same index, i.e., g_{sss}, are nonzero; if all superdiagonal elements are equal to 1, the core array is a superidentity array; finally, a three-way core array is slice diagonal if only the elements with two equal indices, i.e., g_{ssr}, are nonzero.

5.2.6 JOINT BIPLOTS

A large part of the analysis of multiway data concentrates on the links between the modes contained in the core array, that is, the g_{pqr} in the three-way case. The interpretation of these links can be difficult to understand when no clear meaning can be attached to the components themselves. After all, they represent a direction in the component space of maximal variance, which is not necessarily the direction of maximal interpretability. Not being able to assign substantive meaning to a component restricts the possibility of interpreting combinations of such components.

A possible solution for this is to construct a joint biplot of the components of two modes (the display modes) given a component of the third (the reference mode) (Kroonenberg, 2008, 1983). A joint biplot in three-mode analysis is like a standard biplot (Gabriel, 1971) and all the interpretational principles of standard biplots can be used. The specific features are that each joint biplot is constructed using a different slice of the core array. The slicing is done for each component of the reference mode. Each slice contains the strength of the links or weights for the components of the display modes. Thus, the coefficients in the associated component of the reference mode weight the entire joint biplot by their values (Kroonenberg, 2008).

To build a joint biplot after adjusting the CO-TUCKER method, it is necessary to obtain the matrix

$$\Delta(I \times J) = A^X \mathbb{G}_r (B^Y)^T = (A^X)_r^* (B^Y)_r^{*T} \quad (5.2.6.1)$$

For each core slice, G_r, a joint biplot for the component matrix $(A^X)^*(I \times P)$ and component matrix $(B^Y)^*(J \times Q)$ needs to be constructed such that the P columns of $(A^X)^*$ and the Q columns of $(B^Y)^*$ are close to each other as possible. Closeness is measured as the sum of all $(P \times Q)$ squared distances $\delta^2\left[(a^X)_i^*, (b^Y)_j^*\right]$, for all i and j. The construction of a joint biplot is as follows (Kroonenberg, 2008, 1983). The core slice $G_r(P \times Q)$ is decomposed via SVD into

$$G_r = U_r \Lambda_r V_r^T \quad (5.2.6.2)$$

and the orthonormal[12] left singular vectors U_r and the orthonormal right singular vectors V_r are combined with A^X and B^Y, respectively, and the diagonal matrix Λ_r with the singular values is divided between them in such a way that

$$(A^X)_r^* = \left(\frac{I}{J}\right)^{1/4} (A^X) U_r \Lambda_r^{1/2} \text{ and } (B^Y)_r^* = \left(\frac{J}{I}\right)^{1/4} (B^Y) V_r \Lambda_r^{1/2} \quad (5.2.6.3)$$

where the fourth-root fractions take care of different numbers of levels in the two component matrices. The columns of the adjusted component matrices, $(A^X)_r^*$ and $(B^Y)_r^*$,

[12] Two vectors are orthonormal if they are orthogonal and have length equal 1. A $I \times J$ matrix is (column-wise) orthonormal if $X^T X = I_J$, where I_J is a $J \times J$ diagonal matrix with ones and possibly some zeros on the diagonal.

are the joint biplot axes. As $(A^X)_r^*[(B^Y)_r^*]^T = \Delta_r$, each element δ_{ij}^r is equal to the inner product[13] $(a^X)_i^*\left[(b^Y)_j^*\right]^T$, and it provides the strength of the link between i and j as far it is contained in the rth core slice.

Accordingly, for $R = 1$, a biplot of matrix $(A^X)\mathbb{G}_1(B^Y)^T$ was needed; for $R = 2$ a biplot of matrix $(A^X)\mathbb{G}_2(B^Y)^T$ is required, and so on for all the components retained in the third mode.

Thus, as G_r is the part of the three-way array $\mathbb{G}$ that is associated with $R = r$, the principal component r of the third mode, then results of the three principal components (P, Q and R) can be simultaneously shown in a diagram that displays the component scores of the first and second mode associated with each of the components of the third mode. Thus, the joint biplot can be seen as a variant of the biplot of Gabriel (1971). The number of joint biplots that can be displayed depends on the fact that if the core slices, G_r, are squared or not. Thus, when G_r are not square, their rank is equal to $M = min(P, Q)$, and only M joint biplot axes can be displayed.

By simultaneously displaying the two modes in one plot, visual inferences can be made about their relationships. Explicitly:

[13] The inner product of two vectors produces a scalar: if y and x are two vectors, then their inner product in matrix notation is $y^T x$; it is equal to the sum of the products of the corresponding elements, $\sum y_i x_i$. In geometric terms, is the product of the lengths of the vectors times the cosine of the angle between them.

(1) the spacing and the order of the projections of the subjects on a variable correspond to the sizes of their inner products, and thus to the relative importance of that variable to the subjects;

(2) the interpretation of the subjects-variables relationships can be made directly, without involving components axes or their labels. This advantage gains extremely importance when a situation like the described previously (that is, when the scheme of labeling the components is ambiguous or unclear) is legitimacy in the data set under study;

(3) via the core slice G_r the coordinate axes of the joint biplots are scaled according their relative importance, so that visually a correct impression of the spread of the components is created.

In practice, the joint biplot for the CO-TUCKER is explicitly meant to investigate the subjects with respect to the variables, given a component of the third mode. Thus, a first decision has to be made: the selection of the display and the reference mode. This decision depends on the focus of the research and sometimes from *a priori* knowledge that probably exists (and obviously from the perspective of what could be more fruitful for analyses in the study).

In terms of projections, directions, and proximities, the joint biplots are interpreted as the classical biplot from the two-way array (since one dimension has been reduced). On the other hand, in generally, the analysis of this projected biplots can be explored and

interpreted in the background of the Tucker3 model. Specifically, for the present research, the interpretation will be provided in the framework developed by Varela et al. (2009) and Kroonenberg and Röder (2008). For instance, the biplot of the first two modes is projected onto the rth principal component of the third mode such that all levels of the first two modes appear in the biplot; then select, from matrix C (matrix of principal components of the third mode), the levels of the third mode with high weights in the rth components (whether positive or negative).

Suppose that the matrix C has a high positive value associated with the kth level of the third mode on the rth principal component. Additionally, suppose that the biplots of the first two modes are projected onto rth component of the third mode, which is dominated by the high positive value associated with the kth level of the third mode.

Thus,

- the proximities (i.e., how close the points in the biplot are located) among levels of the first and second modes (i.e., ith for the first mode and jth for the second mode) indicate that the three-way interaction among ith, jth and kth is positive.

In contrast,

- if the ith level is far from the jth level, this indicates that the three-way interaction associated with the combined ith, jth and kth levels has a negative effect.

On the contrary, if matrix C has high negative value associated with kth level of the third mode for the rth component, then:

- when the biplots of the first two modes are projected onto rth level of the third mode, proximities between levels of the first and second modes in the biplot indicate that the three-way interaction among ith, jth and kth is negative.

Therefore,

- if the ith level of the first mode is far from the jth level of the second mode at the kth level of the third mode, then the ijk level combination of the three modes has a positive three-way interaction; when the ith level of the first mode is near to the jth level of the second mode at the kth level of the third mode, the ijk level combination of the three modes has a negative three-way interaction.

In general,

- levels of one specific mode located at the centre of the joint biplot are considered to have an average performance (regarding to the three-way interaction) across other modes.

In order to illustrate the previously interpretation, an example of a $2 \times 3 \times 3$ model is provided in figure 23. The first, second and third modes represent sites, environmental variables and months, respectively.

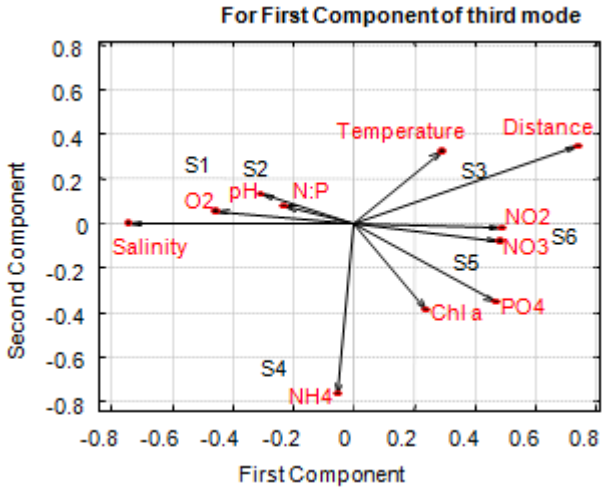


Figure 23. First joint biplot for a hypothetical example of environmental variables and sites for a $2 \times 3 \times 3$-model. The plot represents the first versus second component for first component of third mode (here months). S1 to S6 represents six sampling sites and Salinity, Temperature, pH, O_2, NO_3, NO_2, NH_4, PO_4, N:P ratio, Chl *a* and Distance the environmental variables.

The joint biplot showed the general pattern between sites and environmental variables given the first component of third mode. For the present example the third mode represented the sampling period of the study (months) (Fig. 24). For the first component, the month's mode was indeed relative constant (presented only positive scores) and allowed the examination of the general relationships between the environmental variables.

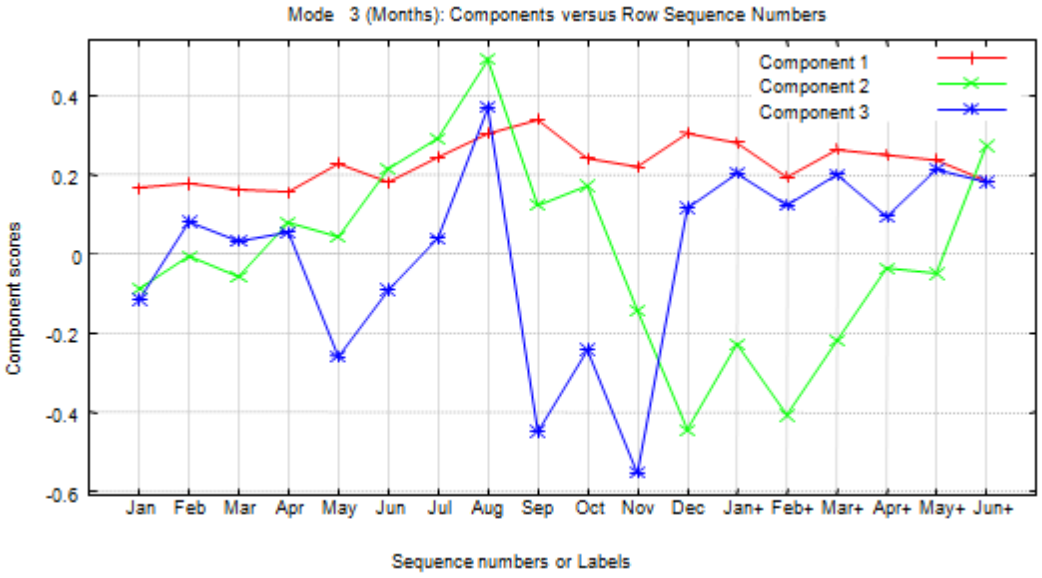


Figure 24. All-components plot for a hypothetical example of months for a 2 × 3 × 3-model. Horizontal axis displays the months and the vertical axis displays the first, second and third component coefficients (scores).

Thus, the general pattern is outlined as follows:

- S1 and S2 presented high salinity values, as well as high concentrations of O2, pH, and N:P ratio, but low concentrations of NO2, NO3, PO4, Chl *a*, and lowest NH4;
- S4 had high concentration of NH4 (followed by PO4 and Chl *a*), high values of salinity, and slightly above average O2, pH, and N:P ratio; additionally, S4 presented somewhat below average temperature, distance, NO2 and NO3;
- S5 and even more S6 presented the reverse pattern of S1 and S2;
- S3 was similar to S5, but its temperature and distance was higher, as well as concentration of PO4 and Chl *a* values are somewhat lower.

As mentioned, since months presented only positive scores (Fig. 24) in first component then the described pattern can be generalized for all sampling period.

Also the joint biplot for second and third month's component can be built. Note, however, that the second and third biplot can be seen as modulations of the general patterns. Thus the patterns in these biplots serve to enhance of diminish patterns observed in the general joint biplot.

5.2.7 INTERACTIVE BIPLOTS

In some studies it is useful to inspect scores of all combinations of levels of two modes on the components of the third mode. For instance, for longitudinal data the scores of each subject-time combination on the variable components can be used to inspect the progress of a score of an individual on a latent variable or variable component over time. These component scores turn out to be a very successful summary of the relationships involved.

Following the basic form developed by Kroonenberg (2008, 1996, 1983), for the CO-TUCKER the component scores will be defined as,

$$w_{ijk} \approx \sum_{p=1}^{P} a_{ip}^{X} f_{pjk} + e_{ijk} \qquad (5.2.7.1)$$

where

$$f_{pjk} \approx \sum_{q=1}^{Q} \sum_{r=1}^{R} b_{jq}^{Y} c_{kr} g_{pqr} \qquad (5.2.7.2)$$

Thus, a component f_{pjk} can be thought of as component score of individual k at occasion j on component p of the first mode. The symbol ≈ indicates that an approximation with a limited number of components is used (note that, the number of P components is conditioned by the number of components that have been retained) (Veldscholte et al., 1998). The plots f_{pjk} against the $f_{p'jk}$ are called interactive plots (also called nested-mode biplots) since (here) the second and third mode are coded interactively. In general, these plots consist in the construction of a biplot in which the row markers are fully crossed combinations of two modes, so that one mode is nested within the other. These plots can be considered as an alternative to joint biplots.

Sometimes it is not very useful to inspect the interactive plots of the scores of different components against one another, as is customary for component loadings. Instead, it is often more useful to inspect the component scores per component against the sequence numbers of the second or third mode. Thus, it is useful to organize data as subjects × conditions in rows and variables in columns. Additionally, when the third mode presents a natural order, the points with the same j for each k can be connected in interactive plots, and the so-called trajectories can be obtained. These trajectories are commonly presented in STATIS (Lavit et al., 1994) (and in its family methods). If variables are centered per subject-occurrence condition, then the origin of the plot represents the profile of the subject with average values on each variable on all occurrences, and the patterns in the plot represent the deviations from the average subject's profile. A remarkable point of the subject-condition coefficients on the combination-mode components is that per variable component they are the inner products between the

variables and the occurrences and thus express the closeness of the elements from the two modes in the joint biplot (Kroonenberg, 2008, 1983).

5.2.8 UNDERSTANDING INTERPRETATIONS ABOUT CORE ARRAY AND JOINT BIPLOTS IN THE CO-TUCKER TECHNIQUE

There are several ways to interpret the core array:

- strength of the link between components of different modes or the weight of the components combination;
- percentages of explained variation;
- three-way interaction measures.

In several of these interpretations it is crucial to know how the components matrices are scaled.

Given the basic form of the Tucker models, the core array contains the linkage between the components from different modes. In particular, the core array indicates the weight of a combination of components, one from each reduced mode. For instance, the element g_{111} of the CO-TUCKER core array indicates the strength of the link between the first components of the three modes, $\left(a_{i1}^{X} b_{j1}^{Y} c_{k1}\right)$, and g_{221} indicates the strength of the link between the second components of the first and the second mode in combination with the first component of the third mode $\left(a_{i2}^{X} b_{j2}^{Y} c_{k1}\right)$. This means that when the data are

reconstructed on the basis of the model (i.e., the structural image is computed), $\widehat{w}_{ijk} = \sum_{pqr}(g_{pqr})\left(a_{ip}^{X} b_{jq}^{Y} c_{kr}\right)$ and g_{pqr} is the weight of the term $\left(a_{ip}^{X} b_{jq}^{Y} c_{kr}\right)$.

Each squared coefficient g_{pqr}^2 indicates the explained variability. Therefore, in the CO-TUCKER $SS(FIT) = \sum_{pqr}(g_{pqr}^2)$. Thus, the fitted sum of squares can be completely partitioned by the elements of the core array. Dividing the g_{pqr}^2 by the total sum of squares, the proportion of the variability explained is obtained by each of the combination of components, for which the term relative fit will be used, and the relative residual sum of squares is similarly defined.

This interpretation of the core array can be used to assess whether additional components in any one mode introduce sizeable new combination of components, or whether the gain in explained variability is spread over a larger number of additional elements of the core array. In the latter case, one might doubt whether it is really useful to add the component.

An interesting side effect of this partitioning of the overall fitted sum of squares is that by adding the g_{pqr}^2 over two of their indices it is possible to get the explained variability of the components themselves; that is, for the component a_p^X the explained variability $SS(FIT)_p$ is $\sum_{qr}(g_{pqr}^2)$. By diving them by the total sum of squares, again the explained variability per component (or standardized components weights) is obtained.

Thus, the core array represents a partitioning of the overall fitted sum of squares into small units through which the (possibly) complex relationships between the

components can be analyzed. In other words, the squared entries of the core array indicate that part of data variance which is modeled by the joint action of component p of the object mode, component q of the variable mode and component r of the condition mode. This fact clarifies the role of the core matrix elements and is crucial for the interpretation of results (Henrion, 1994).

Regarding the content of the modes of a general core array, this can be seen as "miniature data box" (Kroonenberg, 2008). Instead of real data of subjects on variables in several conditions, the "miniature data box" contains elements which represent a combination of latent entities. This means, that each core element g_{pqr} is the score of an idealized subject p on a latent variable q in a prototype condition r. It depends very much on the applications as to whether the idealized subject interpretation is useful in practice. The naming of components of all modes seems to be indispensable in such arrangement, because core elements link the components of the modes, and therefore these must have substantive interpretation. When the scheme of labeling the components is ambiguous (or unclear), the idealized-entity interpretation loses some of its power. In such cases, other interpretations should be preferred. The solution is to discard the idea of trying to find nameable components for all modes that is finding a component interpretation. It seems better to attempt for subspace interpretation by examining joint biplots of two display modes given one component of the remaining reference mode.

5.2.8.1 COMBINING COMPONENT COEFFICIENTS AND CORE ARRAY ELEMENTS IN A CO-TUCKER

Interpretational complexities take place in core arrays because the same components appear in several combinations with components of other modes. This results from the quadrilinear form of each term of the model, $g_{pqr}\left(a_{ip}^{X}b_{jq}^{Y}c_{kr}\right)$, which complicates the assessment of the influence of a combination of components. Thus, the final effect of the components not only depends on whether g_{pqr} is large (or not) but also on the combinations of signs of all four parts of the term $g_{pqr}\left(a_{ip}^{X}b_{jq}^{Y}c_{kr}\right)$. Note that the sign of g_{pqr} is determined by the signs of the p elements of the first mode, q elements of the second mode and r elements of the third mode (for details, see Section 5.2, Equation (5.2.2), that is,

$$g_{pqr} = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} a_{ip}^{X}\, b_{jq}^{Y} c_{kr} w_{ijk}).$$

It is therefore essential to follow the path of the signs (Fig. 25). For instance, suppose that g_{pqr} has a positive sign ($(+)g_{pqr}$) and that the product of pth, qth and rth components of the first, second, and third modes (pth component × qth component × rth component), respectively, is also positive; then the overall effect of the term $g_{pqr}\left(a_{ip}b_{jq}c_{kr}\right)$, is positive. Symbolically, it is:

$$[(+)core] \times [(+)P \times (+)Q \times (+)R] = +$$

A positive contribution of this term can be the result of the four different sign combinations for the elements of the three modes:

$$[(+)P, (+)Q, (+)R]$$

$$[(+)P, (-)Q, (-)R]$$

$$[(-)P, (+)Q, (-)R]$$

$$[(-)P, (-)Q, (+)R]$$

in which a " + " (or " − ") on the pth, qth and rth place in (p, q, r) refers to positive (or negative) coefficients on the pth component of the first mode, qth component of the second mode, and rth component of the third mode respectively.

An analogous formulation is that for a positive contribution, one needs positive coefficients on p to go together with coefficients of the same sign of q and r:

$$[(+)P, (+)Q, (+)R]$$

$$[(+)P, (-)Q, (-)R]$$

and negative coefficients on p to go together with coefficients of opposite signs on q and r:

$$[(-)P, (+)Q, (-)R]$$

$$[(-)P, (-)Q, (+)R]$$

On the other hand, a negative g_{pqr} $((-)g_{pqr})$ with any of the combinations:

$$[(+)P, (-)Q, (+)R]$$

$$[(+)P, (+)Q, (-)R]$$

$$[(-)P, (+)Q, (+)R]$$

$$[(-)P, (-)Q, (-)R]$$

results in a positive contribution to the structural image.

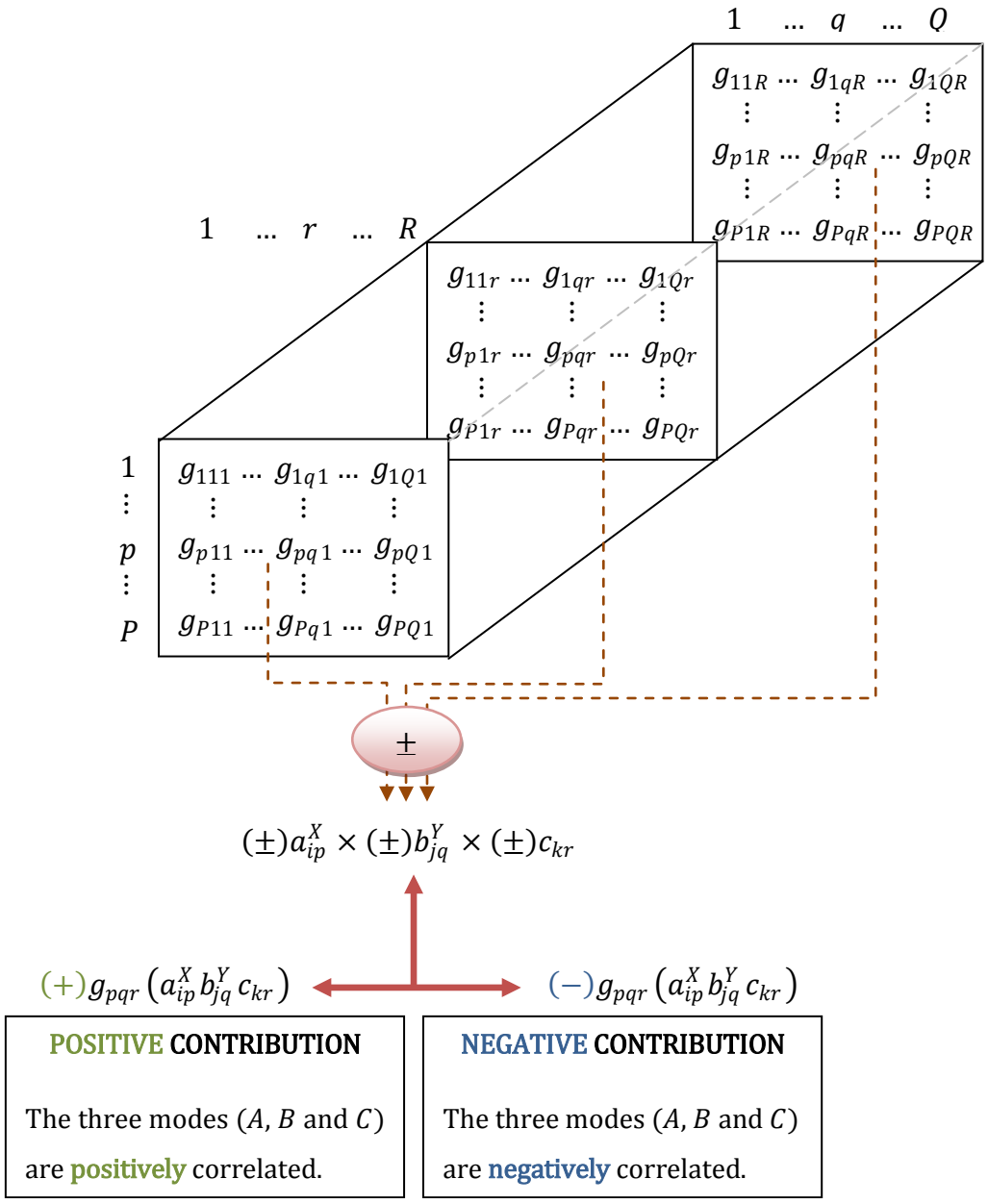


Figure 25. Graphical representation of sign interpretational scheme. Note that, $g_{pqr} = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} a_{ip}^{X}\, b_{jq}^{Y} c_{kr} w_{ijk}$.

In order to explain the meaning of the final effect of the components (which depends on the dimension of g_{pqr} and also on the combinations of signs of all four parts of the term $g_{pqr}\left(a_{ip}^{X} b_{jq}^{Y} c_{kr}\right)$), an example will be provided.

Suppose that high values of the environmental parameter i (high positive score on component p) were observed for high density of species j (high positive score on component q) at a site k (high positive score on component r).

Thus all components (a_{ip}^{X}, b_{jq}^{Y} and c_{kr}) will have high positive coefficients on their own components (high positive scores on p, q and r, respectively). The product of the three component coefficients, $a_{ip}^{X} \times b_{jq}^{Y} \times c_{kr}$, is positive and contributes with a positive value to the structural image $\widehat{w}_{ijk}$ of w_{ijk}, provided the weight of this (p, q, r) combination is also positive. If this weight or element of the core array is comparatively large compared to the other elements of the core array, the entire term $g_{pqr}\left(a_{ip}^{X} b_{jq}^{Y} c_{kr}\right)$ contributes heavily to the structural image of w_{ijk}.

To take this a bit further, first suppose that the components matched to the first environmental parameter, the first species and the first site, that is, the data value w_{111}. In addition, assuming that the p, q and r components were all referring to the fist component ($p = q = r = 1$). Then $g_{111}(a_{11}^{X} b_{11}^{Y} c_{11})$ is the first term in the structural image.

For instance, for a $3 \times 3 \times 2$ model and considering:

$$g_{111} = +13$$

$$a_{11}^{X} = +0.4$$

$$b_{11}^{Y} = +1.5$$

$$c_{11} = +0.5$$

the structural image for w_{111} is built up the following way:

$$\begin{aligned}\widehat{w}_{111} = {} & g_{111}a^X_{11}b^Y_{11}c_{11} + [g_{211}a^X_{12}b^Y_{11}c_{11} + g_{311}a^X_{13}b^Y_{11}c_{11} + g_{121}a^X_{11}b^Y_{12}c_{11} + \\ & + g_{221}a^X_{12}b^Y_{12}c_{11} + g_{321}a^X_{13}b^Y_{12}c_{11} + g_{131}a^X_{11}b^Y_{13}c_{11} + g_{231}a^X_{12}b^Y_{13}c_{11} \\ & + g_{331}a^X_{13}b^Y_{13}c_{11} + g_{112}a^X_{11}b^Y_{11}c_{12} + g_{212}a^X_{12}b^Y_{11}c_{12} + g_{312}a^X_{13}b^Y_{11}c_{12} \\ & + g_{122}a^X_{11}b^Y_{12}c_{12} + g_{222}a^X_{12}b^Y_{12}c_{12} + g_{322}a^X_{13}b^Y_{12}c_{12} + g_{132}a^X_{11}b^Y_{13}c_{12} \\ & + g_{232}a^X_{12}b^Y_{13}c_{12} + g_{332}a^X_{13}b^Y_{12}c_{12}] \\ = {} & 13 \times 0.4 \times 1.5 \times 0.5 + [\text{the others terms}] \\ = {} & 3.9 + [\text{the others terms}]\end{aligned}$$

Because all the coefficients and the corresponding core element are positive (+) the first term has a positive (+) contribution of 3.9 to the structural image of w_{111}.

A further positive contribution to $\widehat{w}_{111}$could from the fact that variable $j = 1$ has a large negative loading on species component $q = 2$ (say, for example, $b_{12} = -1.5$), condition $k = 1$ has a large negative loading on sites component $r = 2$ (say, for example, $c_{12} = -0.5$).

Then the term:

$$a^X_{11}b^Y_{12}c_{12} = (+0.5) \times (-1.5) \times (-0.5) = +0.3$$

is positive, and depending on the value g_{122} the contribution to the structural image is positive (+) or negative (−).

Thus, the final effect of the components on the structural image of a data point depends not only on the sizes but also on the combinations of the signs of all four parts of the quadrilinear term $g_{pqr}\left(a_{ip}^{X}b_{jq}^{Y}c_{kr}\right)$.

In fact, this intrinsic complexity in the combinations of positive and negative loadings of different components and core array elements obviously requires concern regarding interpretation. Moreover, part of the problem arises because the components are continuous rather than discrete entities, unlike interactions between levels of factors in analysis of variance and categories in contingency tables.

A considerable simplification of the process arise when in the data sets, certain components have only positive coefficients. Thus, the number of combinations is substantial reduced and the procedure is further straightforward. Also, when core elements are small (near zero), they do not need interpretation, and the evaluation of the results are also simplified.

All this mental juggling requires quite a bit of ability to get a properly worded statement. Besides (and further then ability), it is necessary to analyze if the means and names provided for all components in all modes make sense for the data under study (which are not always to devise and depends on the type of data).

A good policy to simplify the interpretation is to make conditional statements by only make statements about elements that have, for instance, positive coefficients on a

component. The core array then represents only the interaction between the coefficients of the other modes, given the positive coefficients on the third.

Besides the discussion of the core array, per se, this will also be used to construct the joint biplots. They turn out to be very useful of such approach.

However, all these reported difficulties easily become a motivation for the use of such models in the statistical analysis of (complex) data. The ability to go much further in consideration of the data under study is undoubtedly an appeal.

5.3 CO-TUCKER: ADAVANTAGES AND DRAWBACKS

One point that should be noted is the fact that due to the least squares approach used in the Tucker3 model to estimate the parameters, then it is possible to indicate how well the variability of each subject (i), variable (j) and condition (k) is represented by the fitted three-mode model. This may be an advantage of CO-TUCKER over STATICO, since this is based on maximizing the similarity of all the series of co-structures, not allowing to understand if it adequately fits individual parts of data. Alternatively, such an investigation can demonstrate that certain levels (i, j, k) of a mode $(A^X(I \times P), B^Y(J \times Q), C(K \times R))$ are dominating the solution more than is acceptable, given their role in the investigation.

On the other hand, great care must be taken in the way of organizing the k-tables in a series of k-tables (Fig. 14). For a three-way array, there are three ways to slice the data cube into a series of tables. However, only two are really interesting: the option to put one table per each variable is mostly uninteresting and inconsistent with the objective of most investigations. Therefore, the choice between putting a table for each k condition or a table for each k subject plays an important role. It is obvious that this choice is dependent on the objective in question. However, this possibility in CO-TUCKER should be used with care, as this flexibility might be obtained at the expense of losing some of the structures in the data set. Indeed, if the organization is to have a table by condition (i.e., I subjects/objects $\times J$ variables $\times K$ conditions), then executing the CO-TUCKER, information on the subjects/objects will be "lost" (analogous to an organization type I conditions $\times J$ variables $\times K$ subjects/objects, where the information that is "lost" is related to the conditions). This is in fact a drawback of CO-TUCKER approach.

Basically, most multi-table methods look for some resemblances between tables' data (i.e., matrices) and not for differences. If there are many differences between the two table data sets, then these multi-tables methods will point out that there is no common structure and further explorations have to be made to explain them with other techniques. Moreover, as the first axes of these methods will show the main accordance between tables' data, a look at the higher dimensions or at residuals can give information on the differences between both tables. Hence, probably this is one of the main goals of CO-TUCKER since (via Tucker3 model) it enables the selection of different number of components for each mode and because of the least squares approach used to estimate the parameters, it is possible to indicate how well the variability of each subject, variable, and condition is represented by the fitted three-mode model.

In conclusion, the availability of methods able to the analysis data sets with a complex organization, like pairs of data cubes, is important because it allows to take into account this organization and to analyze the data sets globally. It is obvious that adding complexity in the exploration of data cubes relationships could increase difficulty in interpretations. However, this is an indispensable step to understand the data content and what they might mean. Therefore, the method presented here can be seen as an additional contribution (as well as an alternative) to understand the problems behind the data sets.

Finally, it is important to refer that the method name does not follow the terminology of Thiolouse. This author takes into account that it is first placed the abbreviation for the second method in use (for instance, STATICO, "STATIS and CO-inertia", is based on Co-

Inertia Analysis and then a PTA, and COSTATIS, “CO-inertia and STATIS” is based on PTA plus Co-Inertia Analysis).

For this work it is considered that the abbreviation should reflect the sequence of methods in use (making it more intuitive). Hence, CO-TUCKER means “CO-inertia and TUCKER", since first use the Co-Inertia Analysis k times (to compute the sequence of k cross-covariance tables), and then a Tucker model (namely, the Tucker3 model) to analyze this new k-table.

5.4 CO-TUCKER: AN APPLICATION TO STUDY A COPEPOD COMMUNITY

In the next section an application will be shown illustrating many of the aspects of the CO-TUCKER.

The focus will be on the three-way aspects while details concerning the data survey design can be found in the several works mentioned across the text.

5.4.1. PROBLEM

The Mondego River estuary, located on the west coast of Portugal (40º 08'N, 8º 50'W), has an area of 8.6 km^2 and a volume of 0.0075 km^3 (Fig. 26). At about 5.5 km from the sea, the river branches into two arms (north and south), which converge again near the mouth. Tides in this system are semi-diurnal.

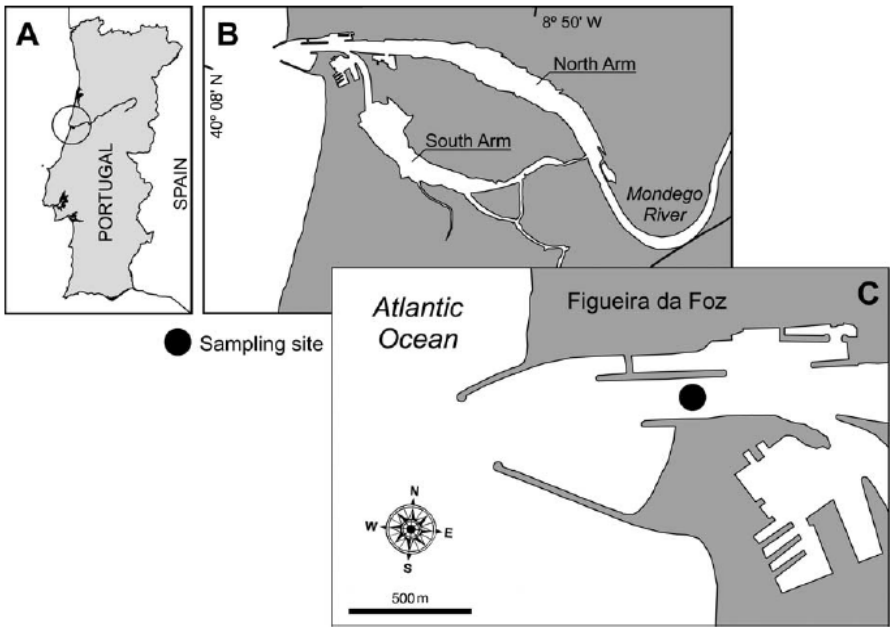


Figure 26. Map of the Mondego estuary and location of sampling site (Marques et al., 2009).

In order to investigate copepod assemblages' distribution patterns and determine the main hydrological and physical factors structuring it at different temporal scales, copepod and hydrographic surveys were conducted seasonally during a one-year period, at a fixed station in Mondego estuary. Hourly samples were collected, during a diel cycle from sub-surface and 1 m above the bottom, at spring and neap tides.

Taxonomic and quantitative analyses were performed on sub-samples, a minimum of 500 individuals were counted. At each sampling event the following environmental parameters were measured, water temperature (T), salinity (S), dissolved oxygen (DO), and pH which were recorded *in situ* with appropriate sensors (WTW) at both depths (DEPTH). Also, water sample was filtered for determination of chlorophyll *a* (Chl a), suspended particulate matter (SPM) and particulate organic matter (POM).

For a complete description about hydrological conditions recorded during the study period see Marques et al. (2009).

In particular, for the present application, only the spring season was considered.

The aim of this example is to present an application to real data of the proposed CO-TUCKER method (which can be applied for analysis of three-way data sets) and to compare its performance with STATICO technique (Thioulouse et al., 2004; Simier et al., 1999). The two methods are applied in order to investigate copepod assemblages' distribution patterns.

The results for CO-TUCKER method are interpreted and simultaneously the performance with the STATICO method is compared.

5.4.2. DATA

The copepods species density and environmental parameters for spring season, lunar phase and depth were combined to generate two series of tables (Fig. 27): one for 8 environmental parameters and another for 15 species densities (the most common species were selected in order to decrease the number of zeros in the analyses, see Table 1).

Table 1. List of species, habitat and abbreviation (abbrev) of copepod taxa used in the study.

Species	Habitat	Abbrev
Acartia clause	Marine	ACCL
Acartia tonsa	Estuarine	ACTO
Calanus helgolandicus	Marine	CAHE
Centropages chierchiae	Marine	CECH
Isias clavipes	Marine	IS
Clausocalanus arcuicornis	Marine	CLAR
Pseudocalanus elongatus	Marine	PSEL
Copidodiaptomus numidicus	Freshwater	COPNU
Paracalanus parvus	Marine	PAPA
Temora longicornis	Marine	TELO
Acanthocyclops robustus	Freshwater	ACRO
Oithona plumifera	Marine	OIPL
Monstrilla grandis	Marine	MOGA
Corycaeus anglicus	Marine	COR
Sapphirinida sp.		SAPH

Each pair of tables shares the same sampling season (spring, in rows), 2 lunar conditions (neap and spring), 2 depths (sub-surface and bottom), the diel cycle (daylight and dark period) and 4 tidal conditions (ebb, low tide, flood and high tide). Therefore, each series of table (biological and environmental) was composed of 4 matrices. Thus, the data form a 25 × 15 × 4 data block of hourly samples (during a diel cycle from sub-surface and 1 m above the bottom, at neap and spring tides) by species densities by spring season and a 25 × 8 × 4 data block of hourly samples (during a diel cycle from sub-surface and 1 m above the bottom, at neap and spring tides) by environmental variables by spring season.

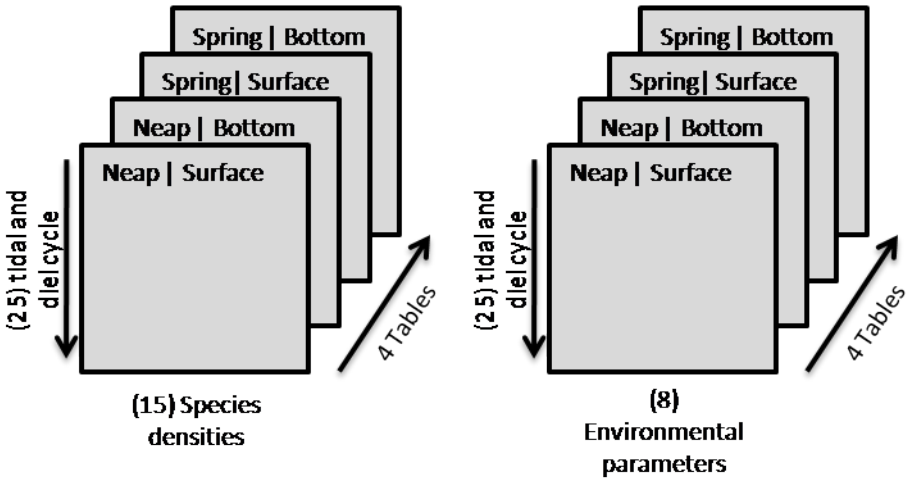


Figure 27. Tables arrangement for copepods species density and environmental parameters was combined for each lunar phase and depth. Neap surface and bottom will be coded as NP(S) and NP(B) and, similarly, spring surface and bottom will be coded as SP(S) and SP(B).

5.4.3 METHODS: CO-TUCKER

The CO-TUCKER model was used to the analysis of two data cubes. As previously was presented in the Chapter 5 (Section 5.2) it is a two-main step procedure. First, a Co-Inertia analysis *k* times was used to compute the sequence of *k* cross-covariance tables,

and then Tucker3 (Tucker, 1966) analysis to analyze this new k-table. Note that, prior to calculations of Co-Inertia analysis species abundance was changed to $log(x+1)$ (Legendre and Legendre, 1979) (to minimize the dominant effect of exceptional catches) and environmental data were normalized to homogenize the table. Therefore, for CO-TUCKER model no pre-processing method was used before the data analysis. Moreover, the data will be treated as so-called three-way profile data (for details of the method, see Chapter 5, Section 5.2).

Each cross-covariance table $X_k(I \times J_k)$ contains the environmental parameters ($I = 8$), characteristic of the species identified and considered for research ($J = 15$) measured in a certain spring tide-depth condition k ($K = 4$). In the subsequent discussion, the components of the different modes will be noted as follows. Environmental parameters mode (P): $P1, P2, \ldots, PI$, species mode (Q): $Q1, Q2, \ldots, QJ$, and spring occurrence mode (R): $R1, R2, \ldots, RK$.

Because the prime interest of this study is focused on the environmental parameters and species relationship at four different sampling situations, and whether there are individual differences in this respect, the results will be presented with plots for each component, but also with joint plots. Such plots (the so-called joint biplots) display the sampling tide-depth situations and environment-species relationship jointly, have the properties of standard biplots (Kroonenberg, 2008, 1994; Kroonenberg and Röder, 2008) and are built using the core array elements. The joint biplot for the CO-TUCKER model is explicitly meant to investigate the subjects (rows) with respect to variables (columns) (in this study, environmental parameters and species densities, respectively),

given a component of the third mode (here the four sampling tide-depth situations). Besides this, when the core array are not square (such as in this case), their rank is equal to $M = min\ (P, Q)$ and only M joint biplots axes can be displayed (here $M = 2$).

Regarding the interpretation of the joint biplots obtained from the CO-TUCKER analysis it follows the explanation presented in Chapter 5, Sections 5.2.6 and 5.2.8. Thus, the biplot of the first two modes (first mode = environmental parameters and second mode = species) is projected onto the rth principal component (say, the first principal component) of the third mode (tide-depth conditions) such that all the levels of the parameters and species modes appear in the biplot.

First, from the matrix of principal components of tide-depth conditions (third mode), the levels with more weights in the rth component (positive or negative) are selected.

Thus,

- when a positive and high value associated to the kth level of tide-depth conditions is selected, then proximities (i.e., how close the points in the biplot are placed) among levels of environment and species modes indicate that the three-way interaction among levels of first mode (environment), second (species) and third mode (tide-depth conditions) is positive.

On the contrary,

- if a level of environment mode (first mode) is far from the level of species mode (second mode), this indicates that the three-way interaction associated with the combined levels of each mode has a negative effect.

Similarly, if the matrix of principal components of tide-depth conditions (third mode) has a high negative value associated with the kth level (for the rth component), then,

- proximities among levels of parameters and species modes in the biplot indicate that the three-way interaction among first, second and third mode levels (environmental parameters, species and tide-depth conditions, respectively) is negative.

Therefore,

- if a level of first mode (environmental parameters) is far apart from a level of the second mode (species) at the kth level of third mode (tide-depth conditions), thus leads to a positive three-way interaction.

Conversely,

- if a level of first mode (environmental parameters) is close to a level of the second mode (species) at the kth level of third mode (tide-depth conditions), thus describes a negative three-way interaction.

As a general rule, levels of one specific mode located at the centre of the joint biplot are considered to have an average performance (regarding to the three-way interaction) across other modes.

On the other hand, and regarding the meaning of the interaction between environmental parameters, species and tide-depth conditions, it follows the rationalization proposed in Chapter 5, Sections 5.2.5. It consists by multiplying the signs of the elements along all components and the core element.

The results of the CO-TUCKER model using graphical outputs obtained on the mentioned three-way data set are presented. The general scheme of CO-TUCKER is showed in figure 28.

Calculations and graphs were done using the 3WayPack software (http://three-mode.leidenuniv.nl) (Kroonenberg, 2003).

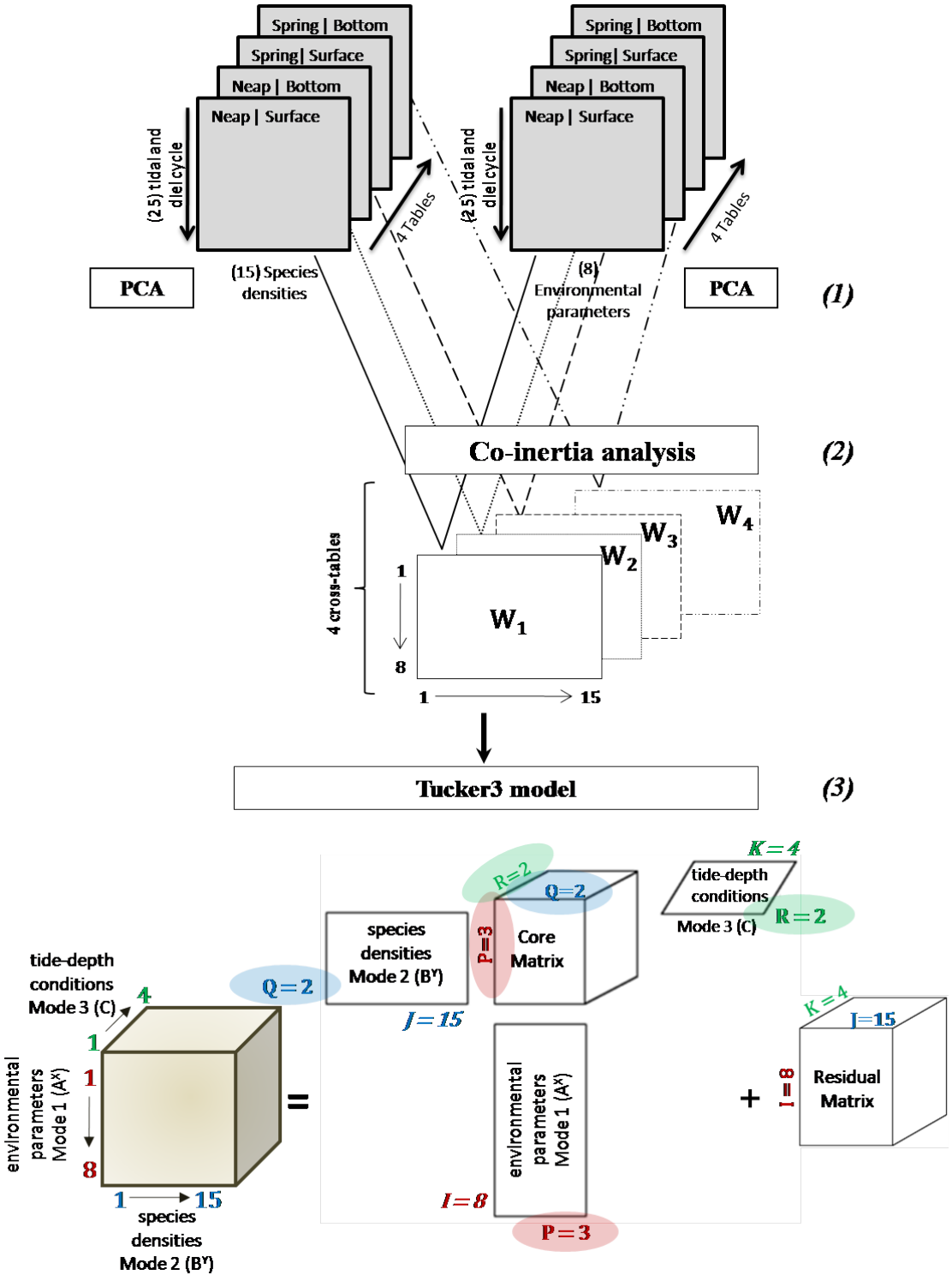


Figure 28. Statistical method scheme for CO-TUCKER method: (1) Basic analyses (PCA for species abundance and environmental tables) are performed on each table; (2) Co-inertia analyses allow linkage of the pairs of PCA-PCA, producing a sequence of 4 cross-tables; (3) Tucker3 model is finally used to analyze this sequence. The data structure is a sequence of four paired ecological tables. Species densities (dimension 25 × 15) and environmental parameters (dimension 25 × 8) are tables in the pair. W_k is the cross-table at 4 occurences of tide-depth conditions.

5.4.4 RESULTS AND DISCUSSION

5.4.4.1 CO-TUCKER ON THE SPECIES-ENVIRONMENT DATA STRUCTURE

The CO-TUCKER model was applied to the sequence of k cross-covariance tables. All possible CO-TUCKER models with different number of components (starting from $1 \times 1 \times 1$ to $5 \times 5 \times 4$) in each mode were evaluated (Table 2). Also, the explained data variance for each combination versus model complexity plots (Fig. 29) were drawn in order to observe, how the added number of components result in the explained variance of data (for details see Chapter 5, Section 5.2.3).

Generally, the optimal complexity of the CO-TUCKER model is the one, which requires the smallest number of components, but still capturing a high fraction of data variance. A systematic and safe approach for finding the optimal model dimensionality is required. It lies between a high fit to data on one hand and a low number of factors representing the systematic variation in the data on the other hand. A model with too many components may lead to over-fitting, whereas, low dimensionality may not allow projection of data in space, thus making it too difficult to overview and interpret the data (Singh et al., 2007).

Table 2. The critical value(1) for $DifFit$-criterion is equal to 0.042. So, $DifFit$ ratios for models with more components than 8 should not be considered.

Sum of Components	Model Size(2)	Difference in Proportional Fit	$DifFit$ Ratio(3)	Proportional $SS(FIT)$
3	1×1×1	0.518	3.331	0.5180
5	2×1×2	0.156	1.207	0.6735
6	2×2×2	0.129	3.068	0.8022
7	**3×2×2**	**0.042**	**1.056**	**0.8442**
8	3×2×3	0.018	(4)	0.8627
9	4×2×3	0.040	1.873	0.9025
10	4×3×3	0.021	(4)	0.9232
11	4×3×4	0.021	1.129	0.9444
12	5×3×4	0.019	2.519	0.9632
13	5×4×4	0.007	2.819	0.9706
14	5×5×4	0.003	(4)	0.9733

Notes:

(1) $Critical\ Value = 1/[S(Min) - 3]$, $S(Min) = MIN[8, (15 \times 4)] + MIN[15, (4 \times 8)] + MIN[4, (8 \times 15)]$, and I, J, K the numbers of levels of the environmental variables, species and tide-depth mode, respectively.

(2) $P \leq QR, Q \leq PR$ and $R \leq PQ$ must hold for all possible solutions, which eliminates redundant solutions.

(3) $DifFit$ ratio computation steps:

Step 1- Consider a model with the same total number of components $S = P + Q + R$;

Step 2- Select the model that has the highest $SS(FIT)$ (or the smallest $SS(FIT)$);

Step 3- Compare classes with different S:

(i) Compute $dif_s = proportional\ SS(FIT)_s - proportional\ SS(FIT)_{s-1}$;

(ii) Consider only the differences that are sequentially highest;

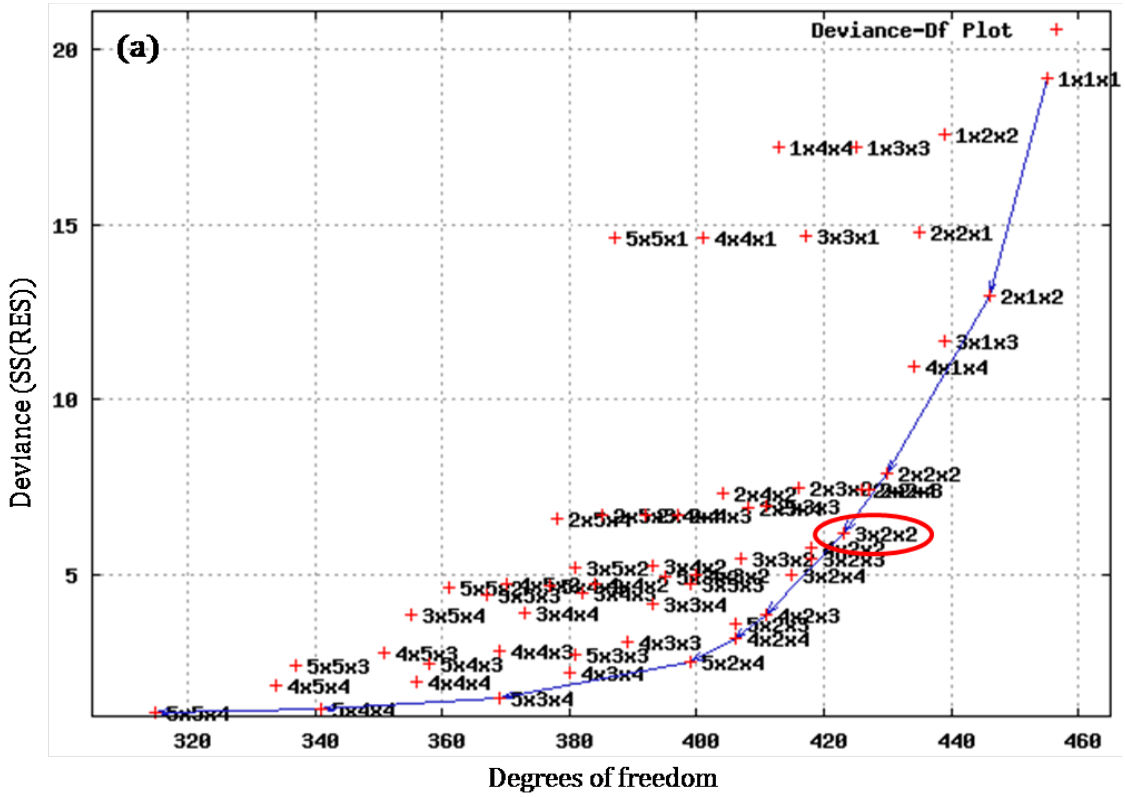
(iii) Compute the salience value, i.e., $b_s = \frac{dif_s}{dif_{s*}}$ (where dif_{s*} has the next highest value after dif_s);

(iv) The $DifFit$-criterion ends designating that the chosen model is one that has the highest salience value.

(4) Cases where the differences are not sequentially highest.

These models with range complexity of (1,1,1) and (5,5,4) explained 51.8% and 97.33% of total variability, respectively.

Balancing the *st*-criterion (Ceulemans and Kiers, 2006), which suggests to choose the model that has the convex hull at the point that has the smallest angle (Fig. 29a), the residual sum of squares (Fig. 29b), the overall fitted and the critical value for $DifFit$ (Table 2), the decomposition model with complexity $3 \times 2 \times 2$ was settled. Note that, the $3 \times 2 \times 2$ product can be considered as an indicator for the number of possible interactions and, consequently, for the complexity of the model (Barbieri et al., 2002).



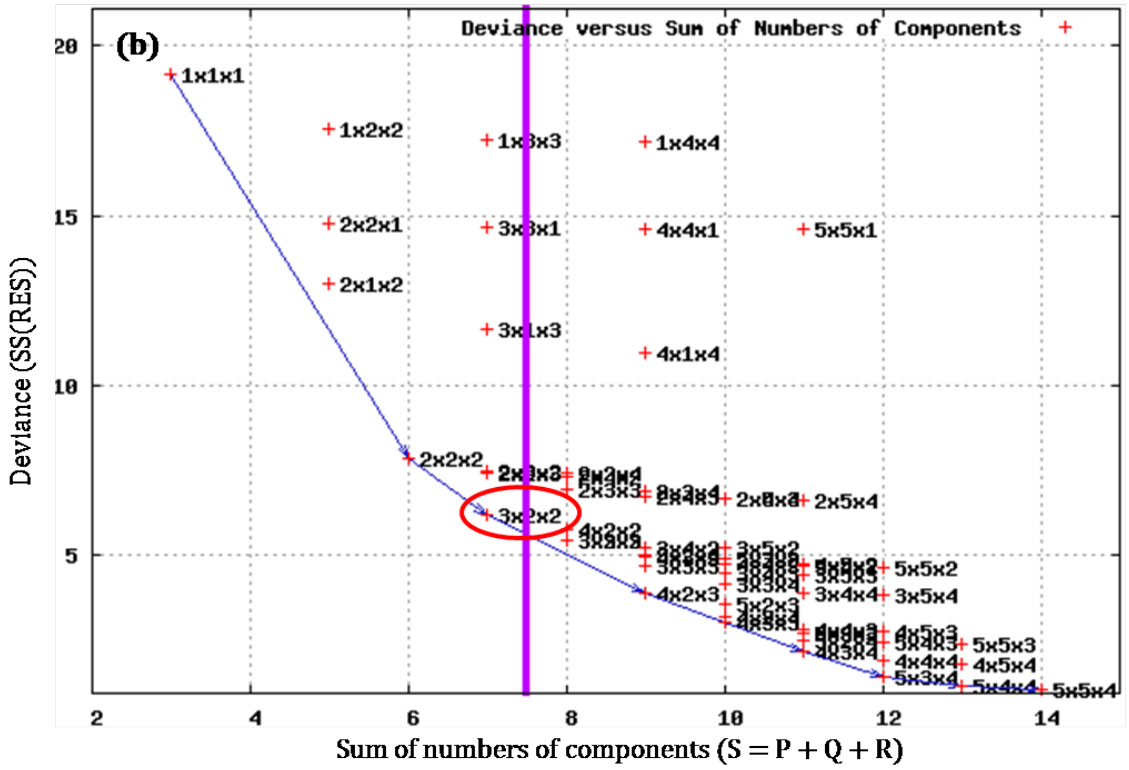


Figure 29. Deviance plots. In the plot $P \times Q \times R$ refers to a CO-TUCKER model with P components for the environment parameters, Q components for the species and R components for the spring samples. **(a)** Deviance (d) vs. degrees of freedom (df) plot for the CO-TUCKER model; **(b)** Deviance (d) vs. sum of numbers of components (S) plot (or three-way scree plot) for CO-TUCKER model. The $DifFit$ values of models to the right of the vertical (purple) line fall below the critical level, $DifFit = 0.042$. The critical value for $DifFit = 1/[S(Min) - 3]$, where $S(Min) = MIN(I, JK) + MIN(J, KI) + MIN(K, IJ)$, and I,J,K the numbers of levels of the 1st, 2nd, and 3rd mode, respectively (Timmerman and Kiers, 2000).

Hence, the choice was a model with 3 components for the environment mode (P), 2 components for the species mode (Q) and 2 components for the spring tide-depth samples mode (R). This model explains 84.42% of data variance (and account a total sum of squares of 39.76, i.e., $SS(TOT) = 39.765$)

It should be noted that, relevant gains in the fitting performances at the increase of complexity could be appreciated (see Table 2). However its interpretation becomes more complex, which would not result in a more comprehensive interpretation. To

clarify, it should be mentioned that, data variance combines all the values in a data set to produce a measure of spread. The percentage of explained data variance indicates a part of total information related to a particular research object carried out in a spread data.

Furthermore, in addition to the overall fit, the fit of the components for each of the modes was examined as well as that of their combinations (Table 3).

Table 3. Sums of squares accounted for by components (based on the core array $\mathbb{G}$). The total sum of squares (or variability) accounted for by the core array is equal to the standardized $SS(FIT)$.

MODE		SUM	PROPORTION VARIATION ACCOUNTED FOR EACH COMPONENT		
1	ENVIRONMENTAL PARAMETRS ($P = 3$)	0.844	0.544	0.255	0.045
2	SPECIES ($Q = 2$)	0.844	0.666	0.179	
3	TIDE-DEPTH ($R = 2$)	0.844	0.621	0.223	
Total sum of squares accounted for		**0.844**			

Note that, the proportion of variation accounted for each component is equal to the standardized component weights. Those standardization is obtained by division through the variability of the three-way interaction (that is, $SS(TOT) = 39.765$) and thus each weight represents the proportion of explained variation. The sum of the weights is equal to the standardized $SS(FIT)$ ($SS(FIT) = 0.844$).

To assess the quality of the fit of a mode, it is useful to observe at their residual sum of squares ($SS(RES)$) in conjunction with their total sum of squares ($SS(TOT)$), in

particular, to investigate per level the relative residual sum of squares, i.e., per level $Relative\ SS(RES) = SS(RES)/SS(TOT)$ (for details see Chapter 5, Section 5.2.4).

Hence:

(1) if a level has a high $Relative\ SS(RES)$, a limited amount of the data is fitted by the model ($SS(RES) > SS(TOT)$);

(2) if the total sum of squares $\big(SS(TOT)\big)$ is small, it does not matter that a level is not fitted well;

(3) if, per level, $SS(TOT)$ is large, there is a large amount of variability that could not be modeled.

All this information could be presented via the fit/residual ratio per level, that is, $SS(FIT)/SS(RES)$, from which the relative performance of an element can be gauged. In particular, large values indicate that the level contains more fit than error information, and vice-versa (for details see Chapter 5, Section 5.2.4).

The $SS(RES)$ and the $SS(FIT)$, per level, as well as their relationships can be shown by a sums-of-squares plot. Figures 30 and 31 show the sums-of-squares plot for the environmental parameters and species in the $3 \times 2 \times 2$ model, respectively.

In each plot:

(1) the red line indicates the average fit. It is drawn from the point (0,0) through the point with the average $SS(FIT)$ and the average $SS(RES)$. Points above the line fit worse compared to the overall model fit and the points below the line fit

better than the overall model. The further away from the line the worse (or better) the fit;

(2) the negatively sloping dashed line (and all lines that could be drawn parallel to it) represents points with an equal $SS(TOT)$, and the further away a line is from origin, the larger total sum of squares;

(3) point on lines fanning out from the origin have the same fit/residual ratio ($SS(FIT)/SS(RES)$).

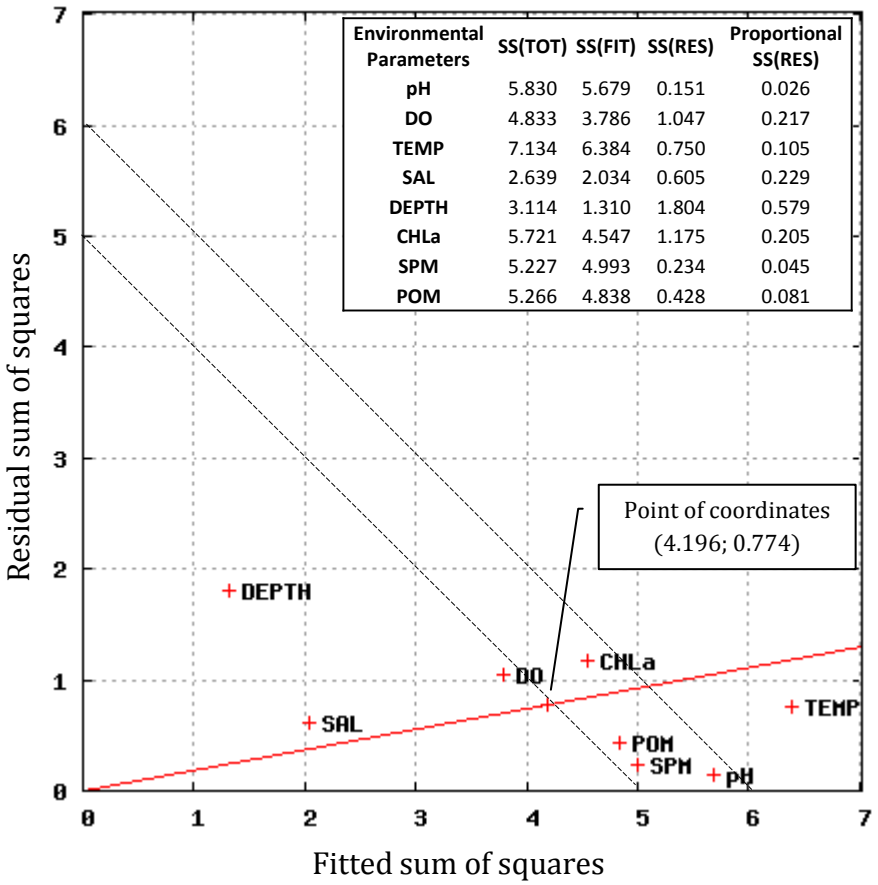


Figure 30. Sums-of-squares plot (or Level-Fit plot) for the levels of mode 1 (environmental parameters). Lines at an angle of -45 degrees are *loci* of points with equal $SS(TOT)$. Fit The line from the point (0,0) through the point with coordinates (4.196; 0.774) separates the ILL-FITTING points (above the line) from the WELL-FITTING points (below the line). T-water temperature, S-salinity, DO-dissolved oxygen, CHLa-chlorophyll *a*, SPM-suspended particulate matter and POM-particulate organic matter.

For environmental standpoint (Fig. 30):

- temperature is the variable with the largest total sum of squares (7.134), and salinity has the smallest one (2.639);
- the best-fitting variables are pH ($SS(FIT)/SS(RES) = 37.61$), SPM ($SS(FIT)/SS(RES) = 21.34$), POM ($SS(FIT)/SS(RES) = 11.30$) and temperature ($SS(FIT)/SS(RES) = 8.51$);
- depth, salinity, dissolved oxygen and chlorophyll *a* are ill represented in the solution. Thus, only very partial statements can be made about the influence of these parameters at the four sampling spring situations.
- pH, SPM, POM and chlorophyll *a* have almost equal $SS(TOT)$, but the first three fit much better than the last one;
- depth is almost similar (low) in $SS(FIT)$ and $SS(RES)$, which means that for all samples depth was an influence more or less equally at bottom or surface.

For species standpoint (Fig. 31):

- ACCL and ACTO are the species with the largest $SS(TOT)$ (18.266 and 14.740, respectively), while the remainder species have extremely low values. The unique species that can reach a value farther from zero is the TELO species (although it is also low, i.e., 3.496);
- the best-fitting species is ACCL (relative fit of 10.88);
- the lack of fit of the majority of species is noticeable (they are located in the origin of the plot). Thus, only very partial statements can be made about the pattern of these species at the four sampling spring situations.

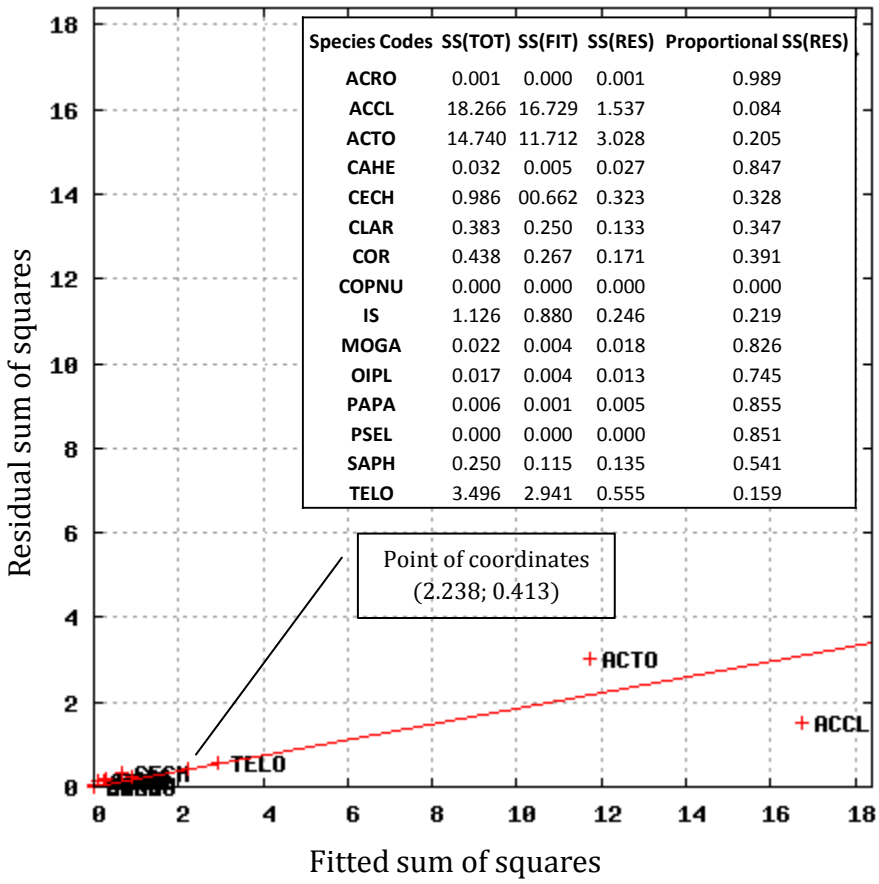


Figure 31. Sums-of-squares plot (or Level-Fit plot) for the levels of mode 2 (species). Lines at an angle of -45 degrees are *loci* of points with equal $SS(TOT)$. The line from the point (0,0) through the point with coordinates (2.238; 0.413) separates the ILL-FITTING points (above the line) from the WELL-FITTING points (below the line). Species are coded as Table 1.

The loading plot of environmental parameters (Mode 1) showed that the first three components of the mode environmental parameters explained 54.4%, 25.5% and 4.5% of total variability (Figs. 32, 33, 34 and Table 4); the loading plot of species (Mode 2) showed that the first two components of the mode species explained 66.6% and 17.9% of total variability (Fig. 35 and Table 5); and the loading plot of tide-depth conditions (Mode 3) showed that the first two components of the mode species explained 62.1% and 22.3% of total variability (Fig. 36 and Table 6).

Table 4. Principal Coordinates for component loadings of Mode 1. Environmental variables codes as Figure 30.

ENVIRONMENTAL PARAMETERS	P1 Explained variation of 54.4%	P2 Explained variation of 25.5%	P3 Explained variation of 4.5%
pH	-1.041	-0.070	-0.231
DO	-0.864	0.115	-0.045
TEMP	1.090	-0.278	-0.136
SAL	-0.306	-0.230	0.512
DEPTH	-0.395	0.324	-0.054
CHLa	0.815	-0.500	-0.024
SPM	0.518	0.856	-0.056
POM	0.391	0.895	0.136

The environmental information is described in terms of different loadings of each parameter on the three components of the first mode (Figs. 32, 33 and 34). A mixed (both positive and negative) pattern of loadings on environmental parameters under the three components of the first mode shows a differential behavior of the measured parameters in the study.

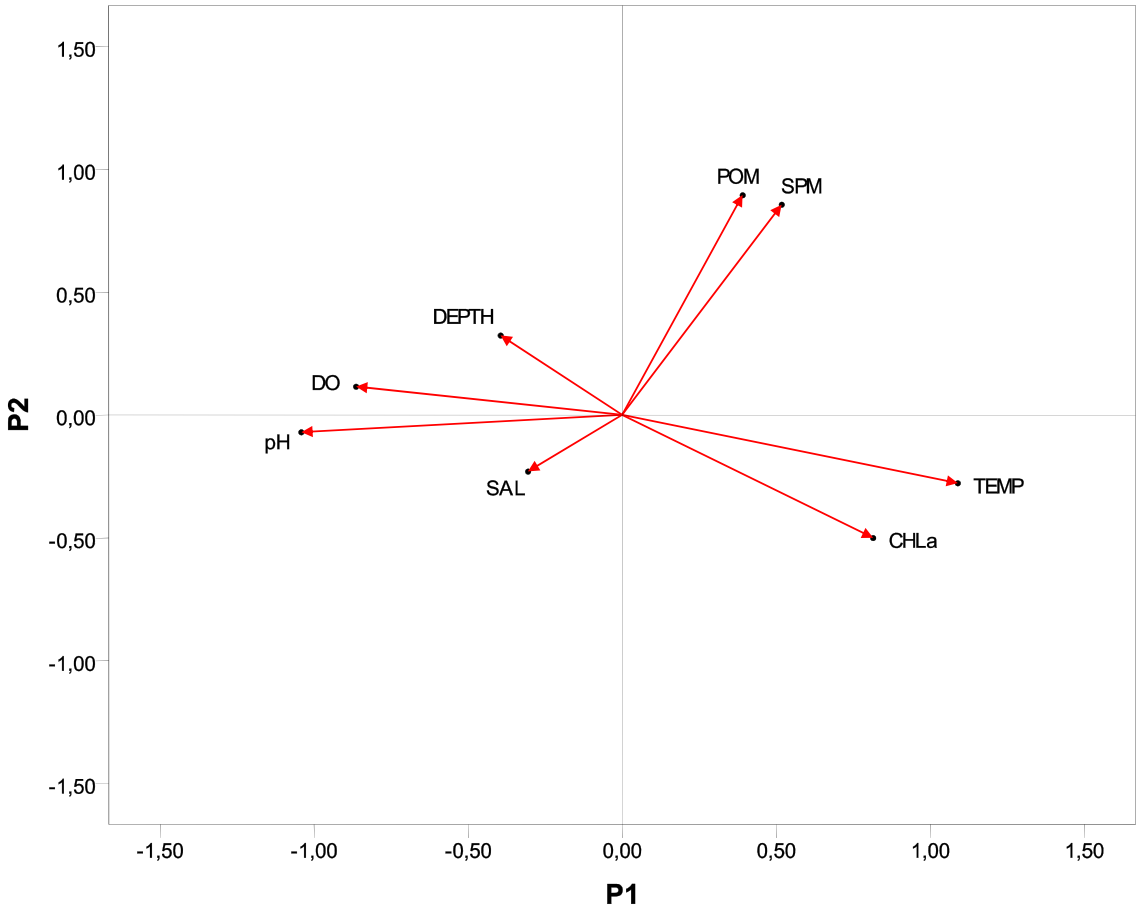


Figure 32. Loading plots (principal coordinates) for mode 1 (environmental parameters). Component *P*1 (54.4%) vs. Component *P*2 (25.5%). Environmental variables codes as Figure 30.

Temperature and chlorophyll *a* under first component (*P*1), POM and SPM under second component (*P*2), and salinity under the third component (*P*3) of the parameters mode (mode 1) exhibited high positive loadings, whereas, dissolved oxygen and pH in *P*1, chlorophyll *a* in *P*2, and pH and temperature in *P*3 component show the highest negative loading.

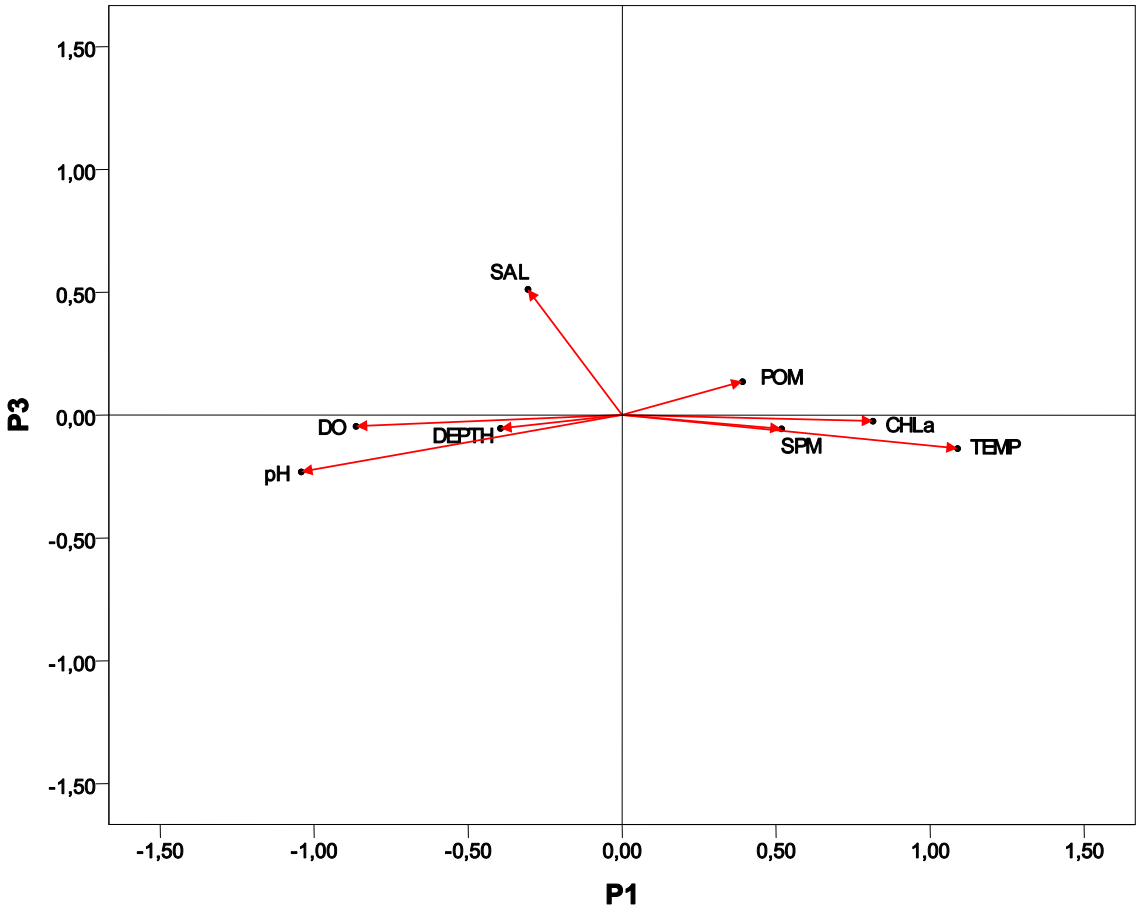


Figure 33. Loading plots (principal coordinates) for mode 1 (environmental parameters). Component *P*1 (54.4%) vs. Component *P*3 (4.5%). Environmental variables codes as Figure 30.

It can be seen that temperature (followed by chlorophyll *a*) is well separated from the rest of the parameters along the first component (*P*1) (Fig. 32 and Table 4), while POM and SPM attaining higher loadings along second component (*P*2) (Fig. 33 and Table 4), as well as, salinity along third component (*P*3) (Fig. 34 and Table 4).

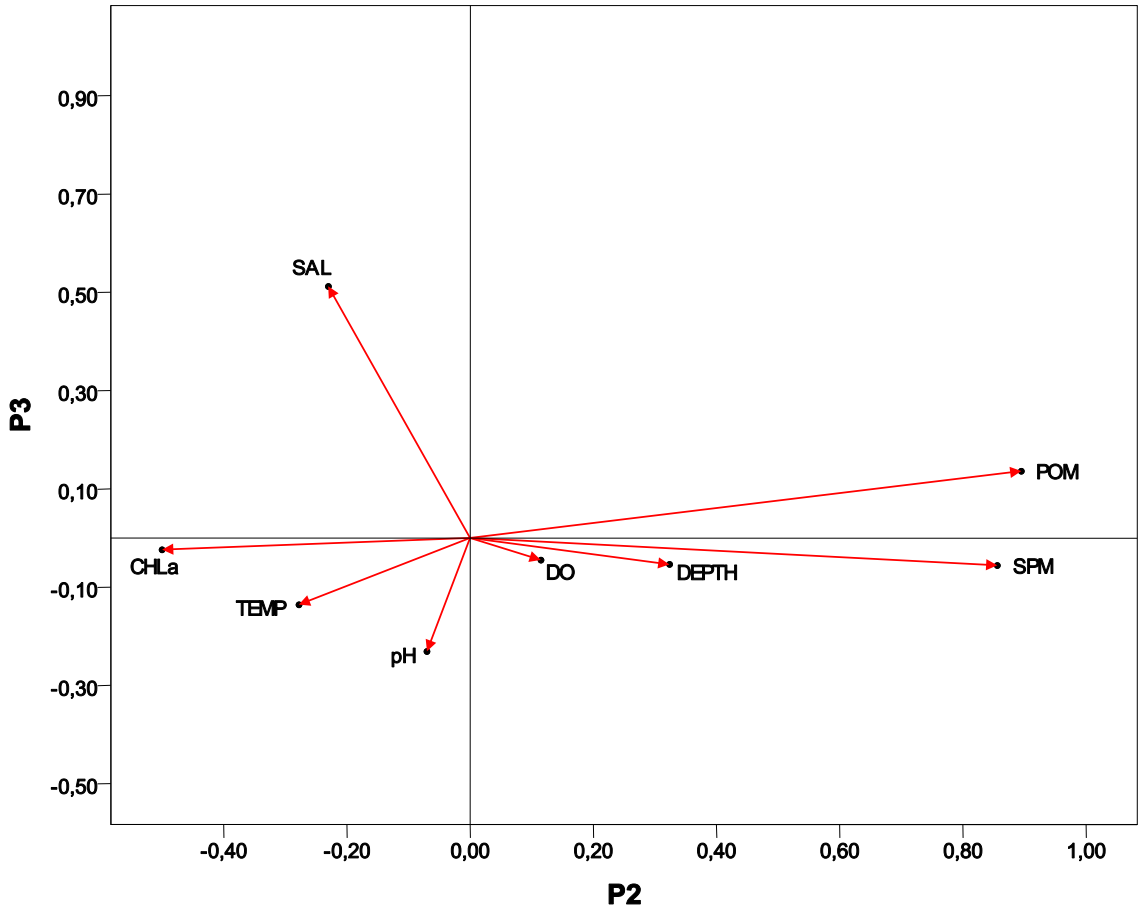


Figure 34. Loading plots (principal coordinates) for mode 1 (environmental parameters). Component $P2$ (25.5%) vs. Component $P3$ (4.5%). Environmental variables codes as Figure 30.

Among the species components (Q) (Fig. 35 and Table 5), the first component ($Q1$) shows a contrast between ACTO species, having high positive loading, and ACCL species (followed by TELO species) having high negative loadings, while the loadings of all the fifteen species in second component ($Q2$) are mostly negative. However, the larger contribute is marked by estuarine species, that is, ACTO species. This loading plot clearly segregated (by means of $Q1$ component) species into two distinguished groups, that is, marine and estuarine species. The remaining species have scores of nearly zero on the first component, indicating that the characteristic pattern mainly of ACTO and ACCL species is not shared by them.

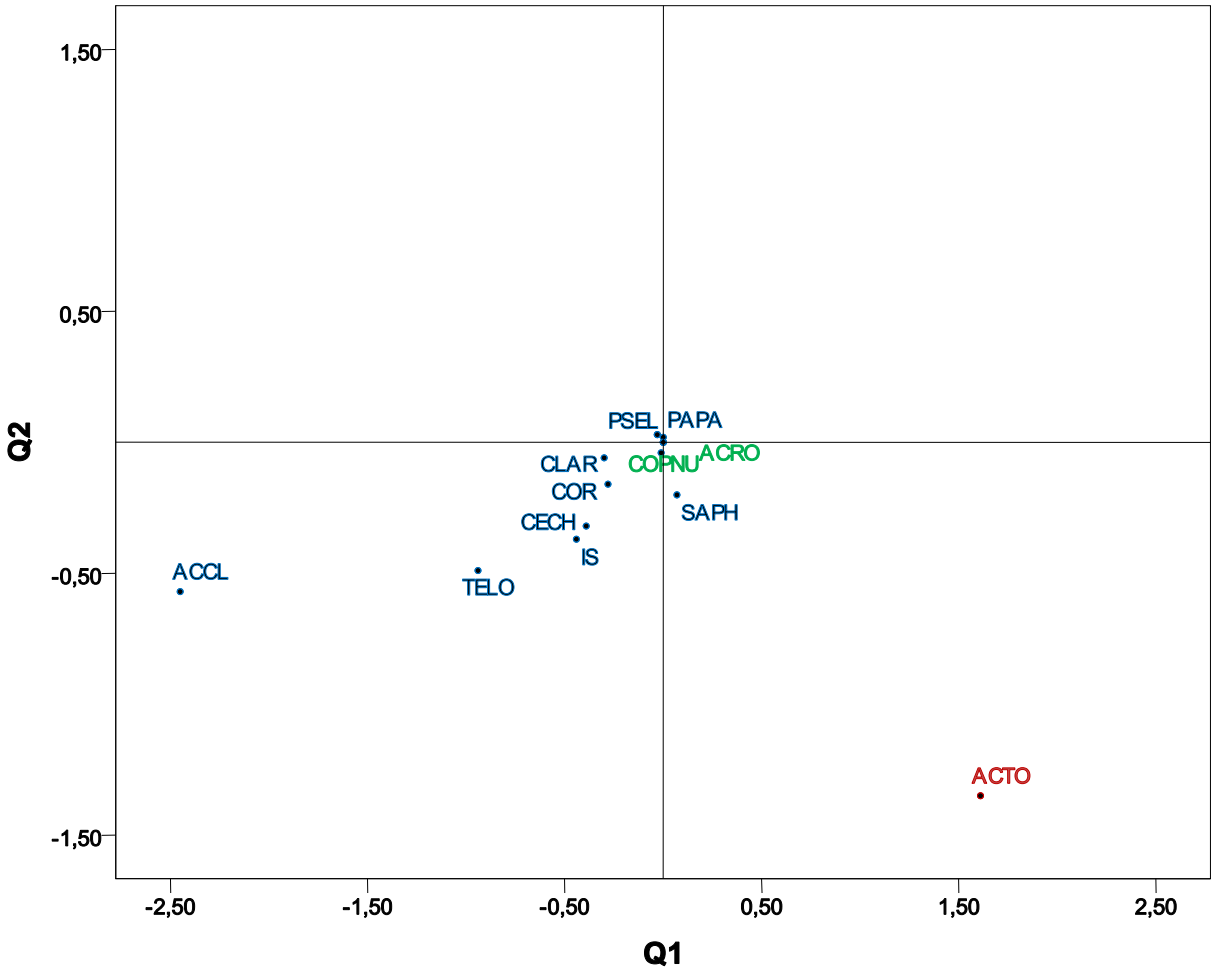


Figure 35. Loading plot (principal coordinates) for mode 2 (species). Component *Q*1 (66.6%) vs. Component *Q*2 (17.9%). Species are coded as Table 1. (●) Marine species, (●) Freshwater species and (●) Estuarine species. For clarity, some overlapping labels were suppressed.

Table 5. Principal Coordinates for component loadings of Mode 2. Species are coded as Table 1. (M) Marine species, (F) Freshwater species and (E) Estuarine species.

SPECIES CODES	HABITAT	Q1 Explained variation of 66.6%	Q2 Explained variation of 17.9%
ACRO	F	-0.002	-0.001
ACCL	M	-2.447	-0.567
ACTO	E	1.614	-1.347
CAHE	M	-0.013	-0.041
CECH	M	-0.388	-0.315
CLAR	M	-0.302	-0.058
COR	M	-0.275	-0.159
COPNU	F	0.000	0.000
IS	M	-0.440	-0.372
MOGA	M	-0.029	0.026
OIPL	M	-0.029	0.029
PAPA	M	0.002	0.018
PSEL	M	-0.004	-0.001
SAPH	M	0.069	-0.196
TELO	M	-0.935	-0.485

Similarly, the information related to the sampling conditions is described through the loadings of each tide-depth condition on the two components ($R1$ and $R2$) of third mode (Fig. 36 and Table 6). Further, it can be seen that the loadings of all the four sampling conditions in first component ($R1$) are positive, while second component ($R2$) segregated tide-depth conditions into two clearly distinguished groups. One is related to neap tides (NP(S) and NP(B)), while the second to the spring tides (SP(S) and SP(B)).

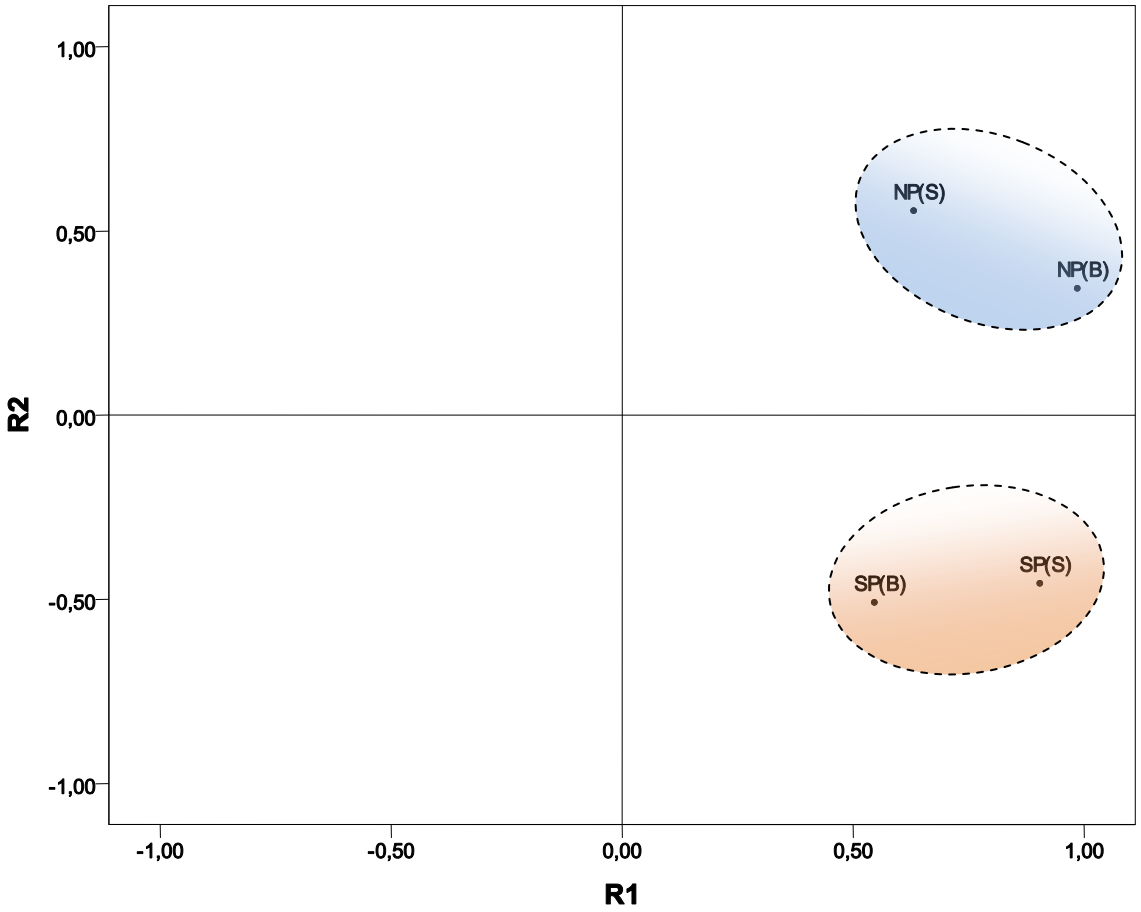


Figure 36. Loading plot (principal coordinates) for mode 3 (tide-depth conditions). Component *R*1 (62.1%) vs. Component *R*2 (22.3%). Tide-depth codes as Figure 27.

Table 6. Principal Coordinates for component loadings of Mode 3. Tide-depth codes as Figure 27.

TIDE-DEPTH CONDITIONS	R1 Explained variation of 62.1%	R2 Explained variation of 22.3%
NP(S)	0.631	0.555
NP(B)	0.985	0.344
SP(S)	0.904	-0.456
SP(B)	0.546	-0.508

To interpret the relationships between the elements in different modes, which reflect the interactions between them, the so-called core array $\mathbb{G}$ is needed (Table 7). It contains the amount of explained variability by the combination of components and has the same dimensionality as the complexity of the model, i.e. $3 \times 2 \times 2$. Thus, the core element g_{pqr} reflects the extent of the interaction between $p = 1,2,3$, $q = 1,2$ and $r = 1,2$.

Note that, the sign of g_{pqr} is determined by the signs of the p elements of the first mode (environmental parameters), q elements of the second mode (species) and r elements of the third mode (tide-depth conditions). Thus, for the $3 \times 2 \times 2$-model ($p = 1,2,3$, $q = 1,2$ and $r = 1,2$), the sign of g_{pqr} is obtained as follows (for details see Chapter 5, Section 5.2, Equation 5.2.2):

$$g_{pqr} = \sum_{i=1}^{8} \sum_{j=1}^{15} \sum_{k=1}^{4} a_{ip}^{X}\, b_{jq}^{Y} c_{kr} w_{ijk}$$

Besides the discussion of the core array, per se, this will also be used to construct the joint biplots. Therefore, for clarity and better understanding, all the important interactions achieved by core matrix elements will be supported by the detailed interpretation of the suitable joint biplot.

Table 7. Frontal slices of core array $\mathbb{G}$. Note that, down are environmental parameters components and across are species components. In bold are the most important core elements and that will be subject to interpretation. Only combinations of factors that have high weights should be used for interpretation.

		Unstandardised Core Slice		**Proportion Explained Variation**	
TIDE-DEPTH component 1					
		SPECIES component		SPECIES component	
		1	2	1	2
ENVIRONMENTAL	1	**4.514**	-0.151	**0.512**	0.001
PARAMETERS	2	0.006	**2.033**	0.000	**0.104**
Component	3	-0.323	-0.249	0.003	0.002
TIDE-DEPTH component 2					
		SPECIES component		SPECIES component	
		1	2	1	2
ENVIRONMENTAL	1	0.043	**1.114**	0.000	**0.031**
PARAMETERS	2	**2.443**	0.155	**0.150**	0.001
Component	3	0.127	**1.271**	0.000	**0.041**

The core matrix $\mathbb{G}$ indicates that the combination of components that extracts the largest variability is the combination of the first component of each mode $g_{111} = 4.514$ (Table 7 and Fig. 37). This combination of components accounted for 51.2% of the variability explained by the analysis and 43.2% of the total variability explained by the model (that is, 51.2% of 84.4%), i.e.,

$$\frac{(g_{111})^2}{\sum_p \sum_q \sum_r (g_{pqr})^2} = \frac{4.514^2}{39.765} = 0.512$$

and

$$0.512 \times 0.844 \; = \; 0.432.$$

Besides g_{111}, other four prominent elements were selected from the core array $\mathbb{G}$ (Table 7 and Fig. 37), namely g_{212} (15% of the variance), g_{221} (10.4% of the variance), g_{322} (4.1% of the variance) and g_{122} (3.1% of the variance). The remaining core elements show rapidly decreasing importance: more than half of them have importance smaller than 1.0%. Thus, the first three loading vectors in the first mode, two loading vectors in the second mode and two loading vectors in the third mode should be considered in the interpretation.

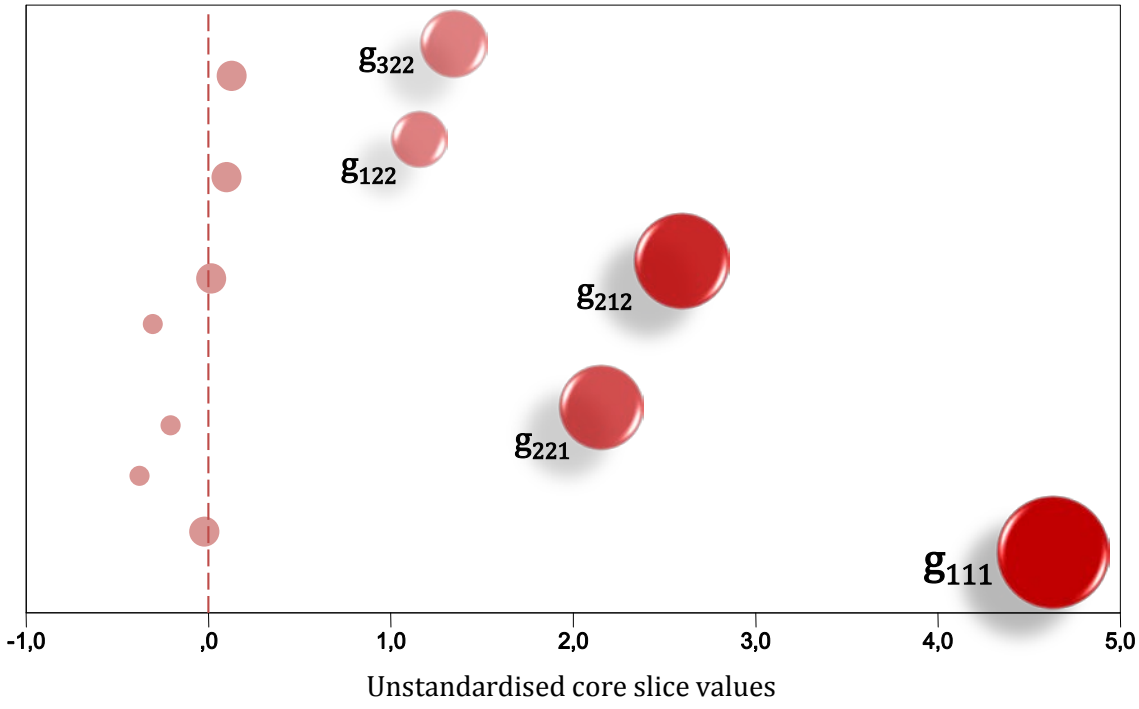


Figure 37. Core array ($\mathbb{G}$) plot. The five prominent core elements ($g_{111}, g_{212}, g_{221}, g_{322}, g_{122}$) are highlighted. The remaining core elements show rapidly decreasing importance.

- **g_{111} core element interpretation**

The first important core element g_{111} (explaining 52.4% of the variability, Table 7 and Fig. 37) with positive sign ($g_{111} = (+)4.514$) indicates interaction among components $P1$, $Q1$ and $R1$. Since, all loadings on the conditions (tide-depth conditions) mode ($R1$)

are ranked according their positive values, the sign of the environmental parameters mode ($P1$), and species mode ($Q1$) determine the sign of the first core element.

Hence, this element can be explained considering four calculations and the first joint biplot (Fig. 38).

Regarding the first calculation, it takes into account the high positive loadings of $P1$ (temperature and chlorophyll *a*), high negative loadings of $Q1$ (mainly contributed by ACCL and followed by TELO) and positive loadings of $R1$ (all tide-depth conditions) and the sign of the core element g_{111}, which is "+", resulting in the "−" sign of their product, that is:

$$P1(+) \times Q1(-) \times R1(+) \times core(+) = (-)$$

The second takes into account the high positive loadings of temperature and chlorophyll *a*, along $P1$, the high positive loading of ACTO along $Q1$ and positive sign of sampling tide-depth conditions along $R1$ and the positive sign of the core element g_{111} resulting in the "+" sign of their product, that is:

$$P1(+) \times Q1(+) \times R1(+) \times core(+) = (+)$$

The third calculation step takes into account the high negative loadings of $P1$ (pH and dissolved oxygen), high positive loadings of $Q1$ (clearly highlighted by ACTO) and positive loadings of R1 and the sign of the core element resulting in the "−" sign of their product, that is:

$$P1(-) \times Q1(+) \times R1(+) \times core(+) = (-)$$

Finally, the fourth takes into account the high negative loadings of pH and dissolved oxygen along $P1$, high negative loadings of ACCL and TELO species along $Q1$ and positive sign of sampling tide-depth conditions along $R1$ and the sign of the core element resulting in the "+" sign of their product, that is:

$$P1(-) \times Q1(-) \times R1(+) \times core(+) = (+)$$

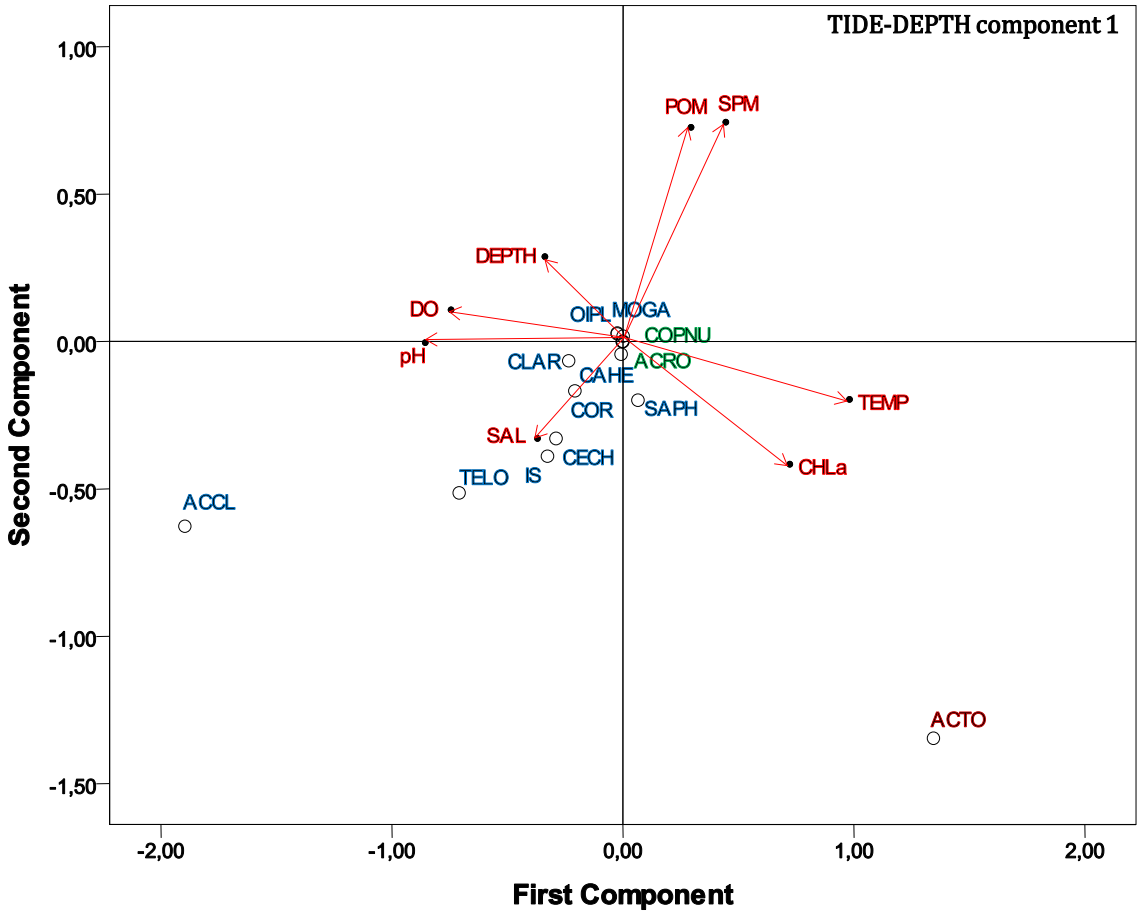


Figure 38. First vs. second component of joint biplots (symmetric scaling of displayed modes) of Mode 1 (ENVIRONMENTAL PARAMETERS, 51.6% of the variance) and Mode 2 (SPECIES, 10.6% of the variance) for TIDE-DEPTH component 1. Species are coded as Table 1. (●) Marine species, (●) Freshwater species and (●) Estuarine species. For clarity, some overlapping labels were suppressed. Environmental parameters codes as Figure 30.

The first joint biplot (Fig. 38) is projected onto the first component of tide-depth condition ($R1$) and it explains 62.2% of the variance. This biplot depicts the interaction environment × species at the neap (sub-surface, NP(S) and bottom, NP(B)) and spring (sub-surface, SP(S) and bottom, SP(B)) tides (since all conditions have positive sign along $R1$). It can thus be pointed out that, in general, for all tide-depth conditions, marine species (mainly for ACCL and TELO) have a positive interaction with pH, dissolved oxygen and salinity (i.e., abundances were higher for higher values of pH, dissolved oxygen and salinity) but a negative interaction with POM, SPM, temperature and chlorophyll *a*; ACTO species (estuarine affinity) has positive interaction with temperature, chlorophyll *a* and salinity (although to a lesser extent) and negative with POM, SPM, dissolved oxygen, pH and depth (that is, ACTO species were lesser abundant when POM, SPM, dissolved oxygen, pH and depth increased).

In general, pH, dissolved oxygen, POM, SPM, temperature and chlorophyll *a* reflects more environment-species interaction variability associated with the first component of the third mode (longer arrows from the centre of the biplot), than depth and salinity (shorter arrows). The remainder species, located at the centre of the biplot, had low interactions with all environmental parameters during all tide-depth conditions.

- **g_{212} core element interpretation**

The core element $g_{212} = (+)2.443$ (which explains 15% of the variability, Table 4) expresses the interactions among components $P2$, $Q1$ and $R2$. Since, $P2$, $Q1$ and $R2$ have a mixed (positive and negative) pattern of loadings, this element can be explained by taking into account the different combinations of these components based on the sign

(positive and negative) of loadings and, simultaneous by the inspection of the second joint biplot (Fig. 39).

Hence, it is possible to considerer positive loadings of all three components ($P2(+)$, $Q1(+)$ and $R2(+)$) or positive loadings of $P2(+)$ with negative loadings of both $Q1(-)$ and $R2(-)$; or positive loadings of $Q1(+)$ with negative loadings of $P2(-)$ and $R2(-)$; or positive loadings of $R2(+)$ with negative loadings of $P2(-)$ and $Q1(-)$. Note that, all these combinations lead to positive interactions (since the core element g_{212} is positive).

That is,

$$\left.\begin{array}{l} \left.\begin{array}{l} P2(+)Q1(+)R2(+) \\ P2(-)Q1(-)R2(+) \end{array}\right\} neap_{tide} \Longrightarrow (+) \\ \qquad\qquad or \\ \left.\begin{array}{l} P2(+)Q1(-)R2(-) \\ P2(-)Q1(+)R2(-) \end{array}\right\} spring_{tide} \Rightarrow (+) \end{array}\right\} \; g_{212}(+) \Rightarrow positive\ interaction$$

Also, it is possible to considerer negative loadings of all three components ($P2(-)$, $Q1(-)$ and $R2(-)$) or negative loadings of $P2(-)$ with positive loadings of both $Q1(+)$ and $R2(+)$; or negative loadings of $Q1(-)$ with positive loadings of $P2(+)$ and $R2(+)$; or negative loadings of $R2(-)$ with positive loadings of $P2(+)$ and $Q1(+)$. In an analogous way, all these combinations lead to negative interactions (since the core element g_{212} is positive).

That is,

$$\left.\begin{array}{l} \left.\begin{array}{l} P2(+)Q1(-)R2(+) \\ P2(-)Q1(+)R2(+) \end{array}\right\} neap_{tide} \Rightarrow (-) \\ \qquad\qquad or \\ \left.\begin{array}{l} P2(+)Q1(+)R2(-) \\ P2(-)Q1(-)R2(-) \end{array}\right\} spring_{tide} \Rightarrow (-) \end{array}\right\} \; g_{212}(+) \Rightarrow negative\ interaction$$

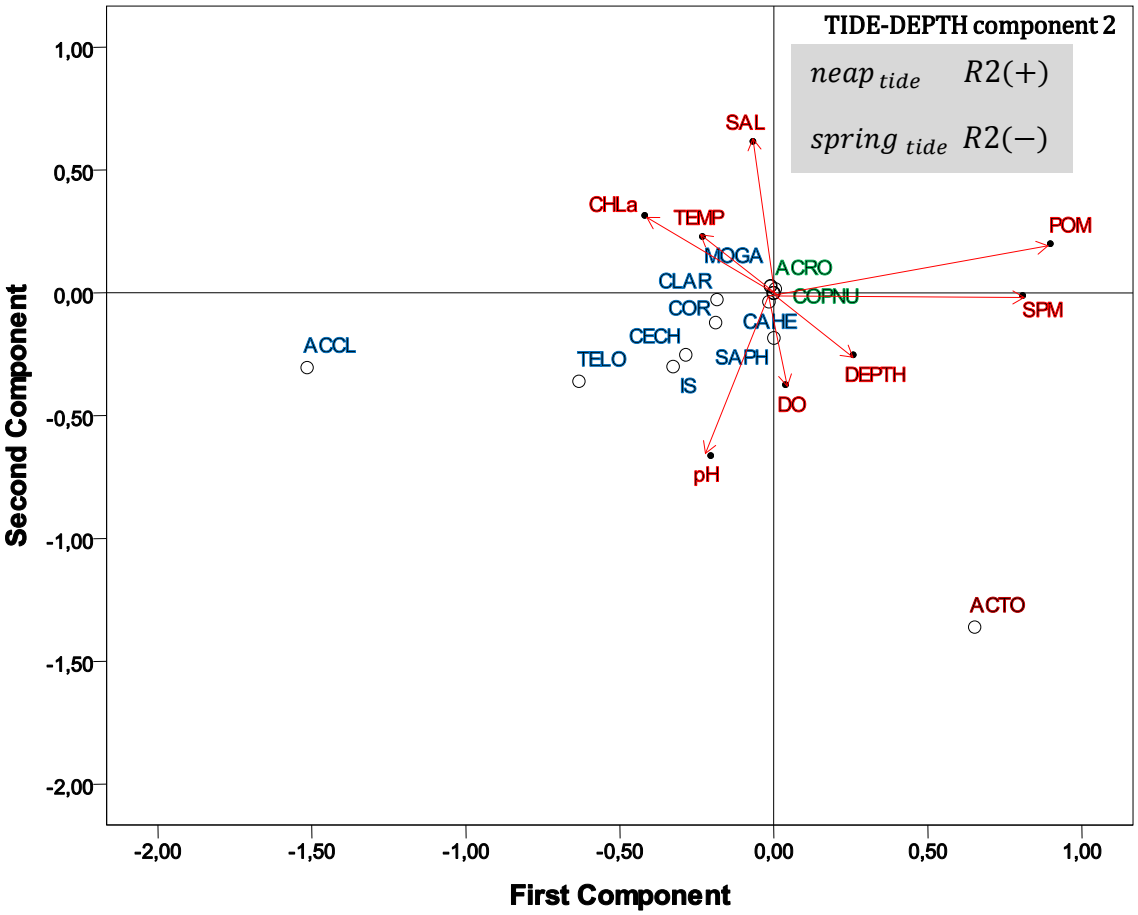


Figure 39. First vs. second component of joint biplots (symmetric scaling of displayed modes) of Mode 1 (ENVIRONMENTAL PARAMETERS, 15.3% of the data variance) and Mode 2 (SPECIES, 7% of the data variance) for TIDE-DEPTH component 2, which is dominated by the contrast of NP(S) and NP(B) vs. SP(S) and SP(B) (see Fig. 36 and Table 6). Species are coded as Table 1. (●) Marine species, (●) Freshwater species and (●) Estuarine species. For clarity, some overlapping labels were suppressed. Environmental parameters codes as Figure 30.

The second join biplot (Fig. 39) illustrates the environmental × species projected onto the second component of tide-depth conditions ($R2$) and explains 22.3% of the variance. Since the second component of the tide-depth contrasts conditions NP(S) (0.555, see Table 6) and NP(B) (0.344, see Table 6) vs. SP(S) (−0.4056, see Table 6) and SP(B) (−0.508, see Table 6), then species with positive interaction in NP(S) or NP(B) should have negative interaction in SP(S) or SP(B) and vice-versa (Fig. 39). Thus, during neap tides (positive sign along $R2$), ACTO species has a positive interaction (i.e., were more abundant) with respect to environmental parameters pH, dissolved oxygen, depth, POM and SPM, but negative interaction (i.e., were less abundant) with salinity, chlorophyll *a* and temperature, whereas ACCL and TELO have a positive interaction with pH, dissolved oxygen, chlorophyll *a* and temperature, but an inverse interaction with depth, POM, SPM and salinity. It is important to note that estuarine species as well as marine species shared the same negative interaction with salinity (that is, at neap tide, as higher the salinity the lower the abundance).

Regarding spring tides, with negative values on the component $R2$, relationships between the species and environmental parameters are reversed. Thus ACTO species interacted positively with salinity, chlorophyll *a* and temperature and negatively with pH, dissolved oxygen, depth, POM and SPM. Hence, ACTO species increased when salinity, chlorophyll *a* and temperature were high and decreased for high values of pH, dissolved oxygen, depth, POM and SPM. Similarly, ACCL and TELO have a positive interaction with respect to depth, POM, SPM and salinity, whereas with pH, dissolved oxygen, chlorophyll *a* and temperature the interaction was in a negative way. Once more, the interactions with salinity play an important role, since during spring tides

(sub-surface and bottom) both estuarine and marine species shared the same positive three-way interaction.

Figure 39 show that not all environmental parameters made the same contribution to the three-way interaction variability (the lengths of the arrows are not similar). Hence, dissolved oxygen, depth and temperature presented small contribution to explain the interactions, when compared to the remainder parameters, which had a sizeable contribution. On the other hand, the majority of the species, including freshwater populations, presented low interactions with all environmental parameters during spring and neap tides (Fig. 39).

- **g_{221} core element interpretation**

The third prominent core element ($g_{221} = (+)2.033$) describes the interaction among components $P2$, $Q2$ and $R1$ and explains 10.4% of variability (Table 7 and Fig. 37). As mentioned, $P2$ is mainly contributed by POM and SPM with positive loadings and by chlorophyll *a* and temperature with negative loadings (Figs. 32 and 34, and Table 4). The sampling tide-depth conditions along $R1$ are ranked according to their positive values, while species densities ($Q2$) are for the most part ranked according to their negative values, with emphasis for the larger contribute of ACTO species (Figs. 36 and 35, respectively). Since, $Q2$ represents an entire assemblage copepods community (pointed out by ACTO species) and hence, could be recognized as a baseline pattern and $R1$ concerns all sampling conditions, a relatively simple assessment of core elements sign and components loadings signs supported by the assessment of the first joint biplot (Fig. 38) is sufficient in this case.

Hence, the first sign combination takes into account the high positive signs of POM and SPM (along $P2$), the negative sign of species densities along $Q2$, the positive sign of all sampling conditions along $R1$ and the sign of the core element g_{221}, which is "+", resulting in the "−" sign of their product, that is:

$$P2(+) \times Q2(-) \times R1(+) \times core(+) = (-)$$

The second calculation takes into account the negative sign of chlorophyll *a* and temperature and, the same as above, signs of remaining parts of equation resulting in the "+" sign of the product, that is:

$$P2(-) \times Q2(-) \times R1(+) \times core(+) = (+)$$

Thus, for all tide-depth conditions, POM and SPM interacted negatively with the generality of copepods community, while chlorophyll *a* and temperature presented a positive interaction mainly for estuarine species (ACTO) (Fig. 38).

- **g_{322} core element interpretation**

Interaction among $P3$, $Q2$ and $R2$ components is expressed by core element $g_{322} = (+)1.271$ (Table 7 and Fig. 37). To facilitate the interpretation of this three-way interaction, the joint biplot that displays species and tide-depth conditions projected onto third component of the environment mode, was built (Fig. 40). Therefore, this plot represents species response patterns relative to the tide-depth conditions associated with a particular combination of environmental parameters ($P3$).

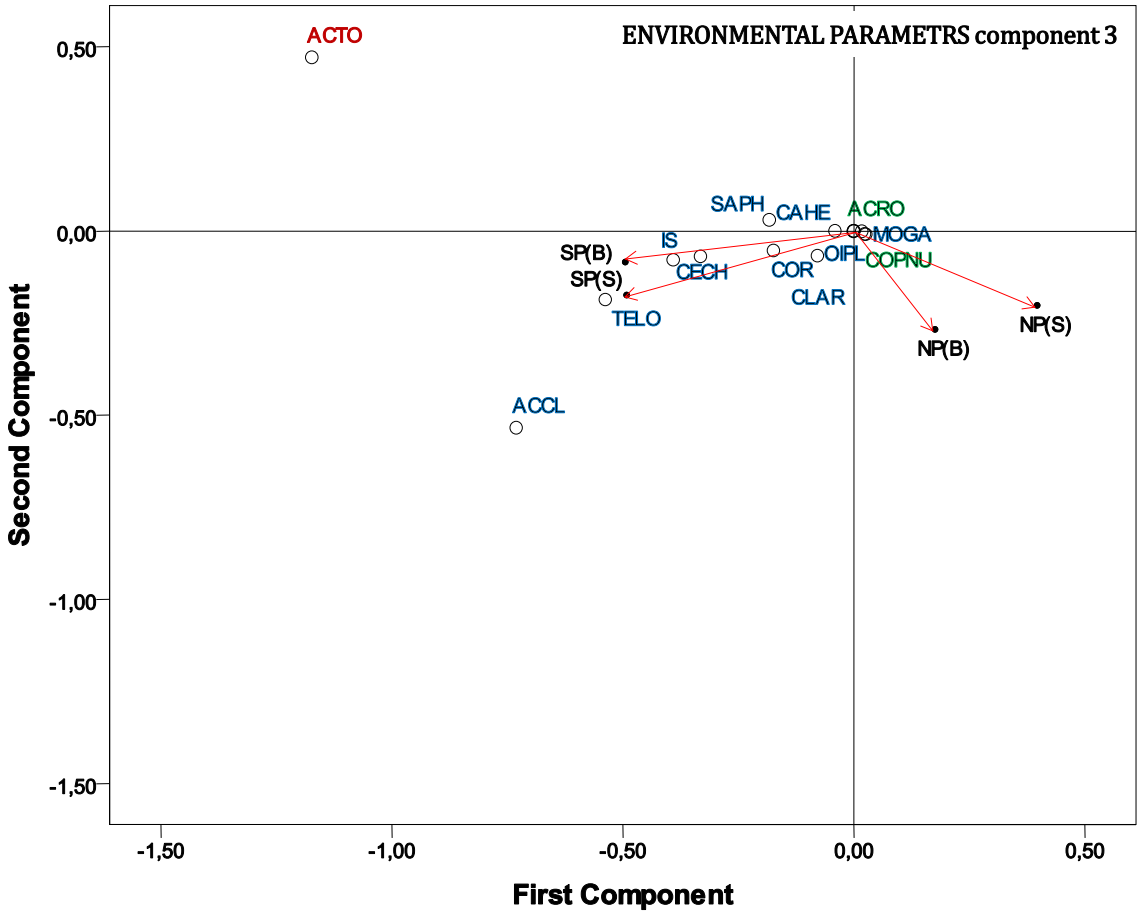


Figure 40. First vs. second component of joint biplots (symmetric scaling of displayed modes) of Mode 3 (TIDE-DEPTH, 4.3% of the data variance) and Mode 2 (SPECIES, 0.2% of the data variance) for PARAMETRS component 3, which is dominated by the contrast of the environment pattern characterized by salinity and POM vs. the environment pattern characterized by pH, dissolved oxygen, temperature, depth, chlorophyll *a* and SPM (see Figs. 33 and 34, and Table 4). Species are coded as Table 1. (●) Marine species, (●) Freshwater species and (●) Estuarine species. For clarity, some overlapping labels were suppressed. Tide-depth codes as Figure 27.

Thus, *P*3 is mainly contributed by high positive loadings of salinity (and POM however with lower importance) and (a weak) negative loadings of pH, depth, dissolved oxygen, chlorophyll *a*, SPM and temperature. Similarly as above along *Q*2, almost loadings have negative sign, and most contribute is made by ACTO species. Component *R*2, as described earlier, discriminated neap from spring tides. Since, for *Q*2 only negative loadings are relevant and, to retain the positive sign of this core element, positive and

negative loadings of the components $P3$ and $R2$ or negative loadings of $P3$ and positive loadings of $R2$ or vice-versa may be considered.

Therefore, four calculations are required to explain the g_{322} element and hence, to explain the inherent interaction that describes 4.1% of the variability.

The first one takes into account the high positive sign of salinity (followed by POM) along $P3$, the negative sign of species (pointed out by ACTO species) along $Q2$ and the positive sign of sampling tide-depth conditions (neap tides) along $R2$ and the positive sign of the core element, resulting in the "$-$" sign of their product, that is:

$$P3(+) \times Q2(-) \times R2(+) \times core(+) = (-)$$

The second takes into account the same as above (high positive sign of salinity along $P3$ and the negative sign of species along $Q2$), combined with the negative sign of spring tides along $R2$ and the sign of the core element resulting in the "$+$" sign of their product, that is:

$$P3(+) \times Q2(-) \times R2(-) \times core(+) = (+)$$

As a result, in general, the response pattern described by these combinations and depicted in figure 40 show that for saline waters the copepods community (highlighted by ACCL, TELO and ACTO species) had low density during neap tides, whereas during spring tides copepods densities increase. Only ACCL species presented an exceptional pattern at bottom during neap tide (NP(B)). However, this exception is not noticeable,

since NP(B) has ACCL density somewhat superior than average as well the second component axis only explains 0.2% of total variability.

The third calculation step takes into account the negative signs along $P3$ and $Q2$, and positive sign of tide-depth samples along $R2$ and the sign of the core element, which is "+", resulting in the "+" sign of their product, that is:

$$P3(-) \times Q2(-) \times R2(+) \times core(+) = (+)$$

On the other hand, the last combination of component loadings leads to consider the negative signs of all the three components ($P3(-)$, $Q2(-)$, $R2(-)$) and the positive sign of the core element, resulting in the "−" sign of their product, that is:

$$P3(-) \times Q2(-) \times R2(-) \times core(+) = (-)$$

In this case it can be pointed out that, in general, for lower salinities the relationships between the species and conditions are reversed. So, on the whole, species densities increase during neap tides but decreases in spring tides when salinity was lower (Fig. 40).

All conditions contribute in the same manner to the three-way interaction variability (the lengths of the vectors are similar). Again all other species reveled to have low interactions with the four tide-depth conditions.

- **g_{122} core element interpretation**

The last important core element $g_{122} = (+)1.114$ (explaining 3.1% of variability, Table 7 and Fig. 37) indicates $P1$, $Q2$ and $R2$ interaction. All components are already described (Figs. 32, 33, 34, 35, 36 and Fig. 39). Thus, since the core element has a positive sign and considering the negative signs along $P1$ (mainly for pH and dissolved oxygen) and $Q2$ (species densities) and the positive loadings along $R2$, the interaction results in a "+" sign, that is:

$$P1(-) \times Q2(-) \times R2(+) \times core(+) = (+)$$

Similarly as above, but now considering negative loadings along $R2$, the product calculation results in a "−" sign, that is:

$$P1(-) \times Q2(-) \times R2(-) \times core(+) = (-)$$

Finally, when $R2$ and $P1$ loadings are positive and $Q2$ is negative, product results in a negative interaction, that is:

$$P1(+) \times Q2(-) \times R2(+) \times core(+) = (-)$$

In a similar way, but for negative loadings of $R2$, their product results in appositive interaction, that is:

$$P1(+) \times Q2(-) \times R2(-) \times core(+) = (+)$$

Thus, during neap tides (positive sign along *R*2), the generality of species presented higher densities with respect to pH and dissolved oxygen, while estuarine species had inferior densities with respect chlorophyll *a* and temperature. On the contrary, during spring tides the interpretation of the response patterns of species by means of environmental parameters projected onto the second component of tide-depth conditions (*R*2) will be exactly the opposite of the interpretation previously given (i.e., for neap tides) (Fig. 39).

5.4.4.2 COMPARISON WITH STATICO

The STATICO method proceeds in three stages (Fig. 41): (1) the first stage consists in analyzing each table by a one-table method (normed PCA of the environmental variables and centered PCA of the species data); (2) after that, each pair of tables is linked by the Co-Inertia analysis (Dolédec and Chessel, 1994) which provides an average image of the co-structure (species-variables cross-table); (3) Partial Triadic Analysis (PTA) (Jaffrenou, 1978; Thioulouse and Chessel, 1987; Blanc et al., 1998) is finally used to analyze this sequence. PTA is a three step procedure, namely the interstructure, the compromise and the intrastructure (or trajectories) analyses (for details of the method, see Chapter 3, Section 3.3; calculations and graphs were done using ADE-4 software, available at http://pbil.univ-lyon1.fr/ADE-4 (Thioulouse et al., 1997)).

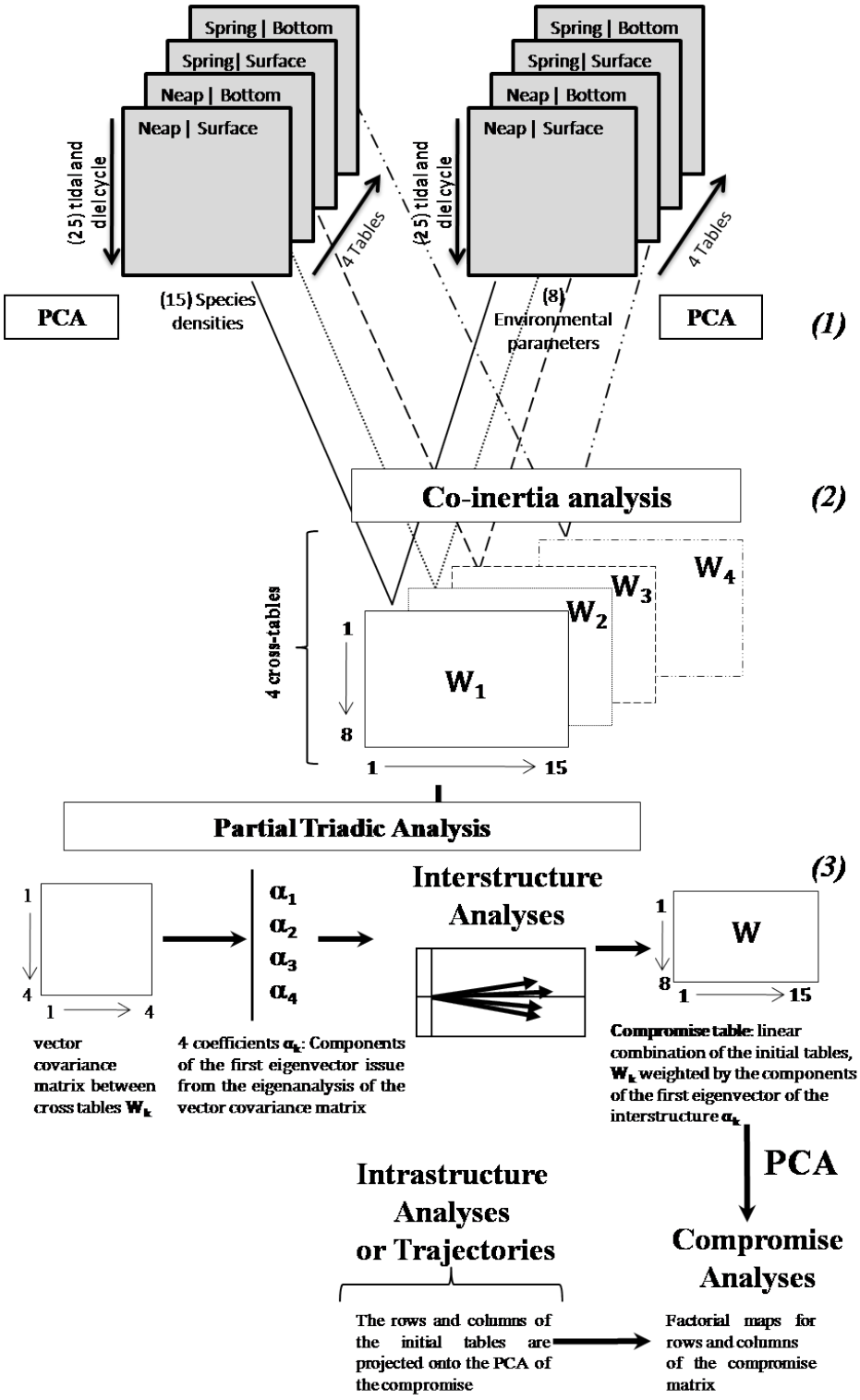
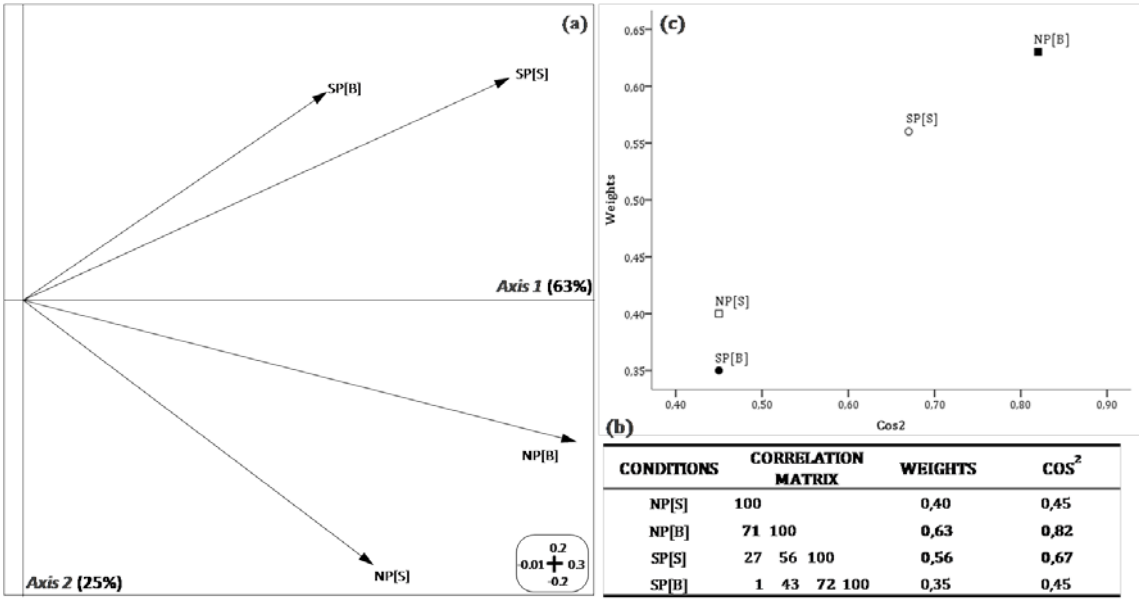


Figure 41. Statistical method scheme for STATICO technique: (1) Basic analyses (PCA for species abundance and environmental tables) are performed on each table; (2) Co-inertia analyses allow linkage of the pairs of PCA-PCA, producing a sequence of 4 cross-tables; (3) PTA is finally used to analyze this sequence. The data structure is a sequence of four paired ecological tables. Species densities (dimension 25 × 15) and environmental parameters (dimension 25 × 8) are tables in the pair. W_k is the cross-table at 4 occurences of tide-depth conditions.

The solution with three components for first mode and two components for second and three modes accounted for 84.4% of the variability, which compares reasonably well with the STATICO solution for the tide-depth conditions of 88% (Fig. 42) and for the species-environment compromise of 82% (Fig. 43).



(b)

CONDITIONS	CORRELATION MATRIX				WEIGHTS	COS^2
NP[S]	100				0,40	0,45
NP[B]	71	100			0,63	0,82
SP[S]	27	56	100		0,56	0,67
SP[B]	1	43	72	100	0,35	0,45

Figure 42. Results of the STATICO method. (a) Interstructure plot showing the four tide-depth conditions and the importance of the corresponding tables in the definition on the compromise (coordinate of the points on the first axis); (b) Typological value indices table: weights represent the weights of tables in the compromise and cos^2 is the square cosine between each table and the approximated compromise; (c) Typological values plot of the four tables (cos^2 vs. weights). Neap surface and bottom are coded as NP(S) and NP(B) and, similarly, spring surface and bottom are coded as SP(S) and SP(B).

In general, all the previously conclusions were also achieved by the STATICO method. In the first instance and, in general, it is possible to observe that the pattern observed for the three modes, parameters, species and tide-depth conditions resembles the pattern of those modes obtained by the STATICO method (see Figs. 42, 43 and 44). However, for parameters and species the sequence of factors obtained in both methods is reversed. The first quadrant of CO-TUCKER corresponds to the four quadrant of STATICO

compromises (Fig. 47). Similar situation was reported by Stanimirova et al. (2004). On the other hand, regarding tide-depth conditions loadings plot (Fig. 36), it is clearly observed that the same pattern (in the same order) are reported by STATICO interstructure plot along axis 1 (Fig. 42). Note that, both methods explained similar proportion of total data variance, i.e., 63% and 62% for STATICO interstructure and tide-depth loadings plot, respectively.

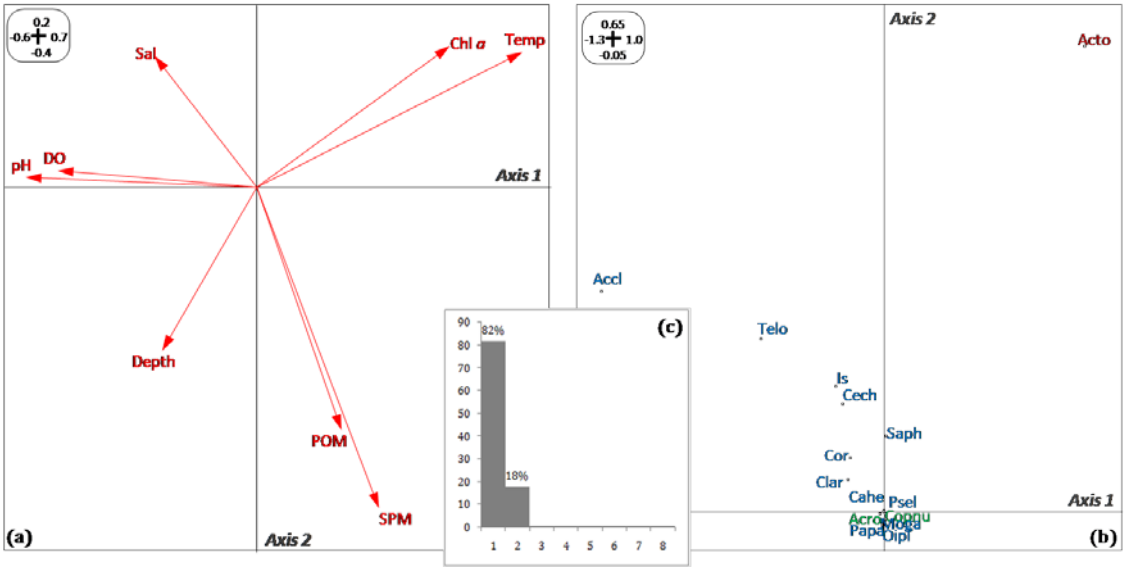
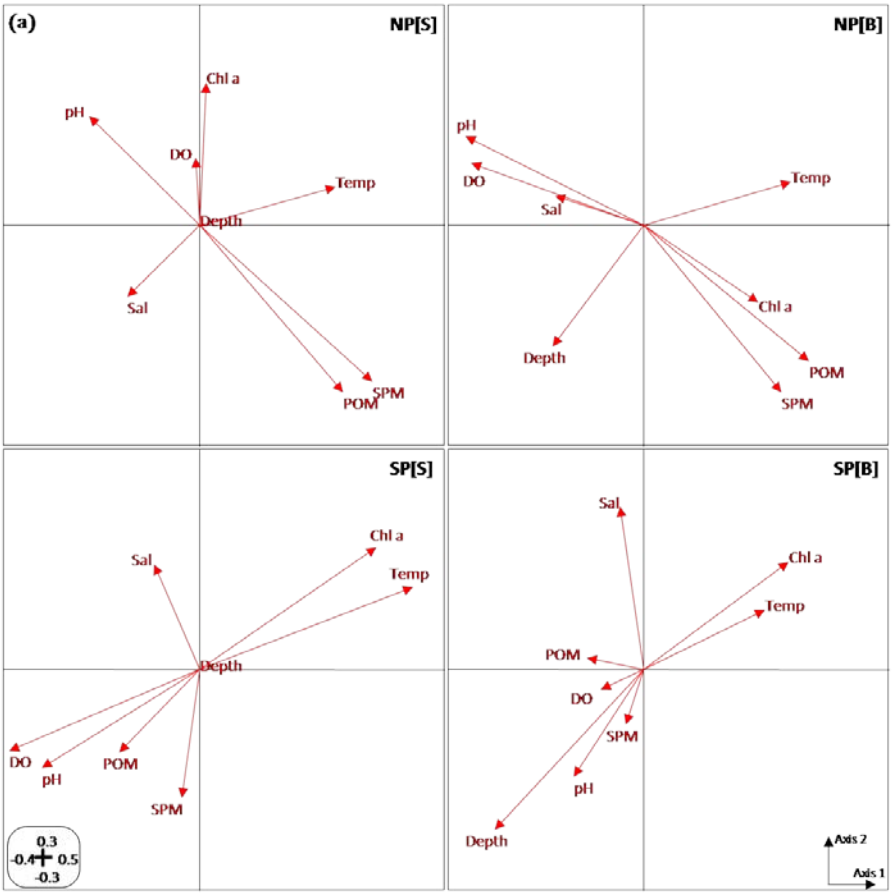


Figure 43. Results of the STATICO method. (a) Compromise analysis principal axes map for environmental parameters. Environmental parameters codes as Figure 30; (b) Compromise analysis principal axes map for copepods species. Species are coded as Table 1. (●) Marine species, (●) Freshwater species and (●) Estuarine species; (c) Explaining variance diagram of the compromise analysis.

Taking into account the results obtained by the interpretation of the first core element (i.e., g_{111}, which extracts the largest variability explained by the analysis, that is, 51.2%; see Fig. 38 and Table 7), as well as the results achieved by the STATICO compromise (Fig. 43) and trajectories (Fig. 48) plots, some considerations should be made.

Therefore, regarding the position of the species in the compromise (Fig. 43b), they are distributed by their salinity affinity (and the correlated factors), with estuarine (ACTO) species on the right side and marine species on the left side. Actuality, the environment variables that separates ACTO from ACCL and TELO are temperature and chlorophyll *a* (at right side), and dissolved oxygen and pH (followed by the depth) (at left side). This was perfectly achieved by the first joint biplot (Fig. 38). So, the estuarine populations were associated with warm and less saline water, while the marine assemblage was characterized by higher values of salinity, pH and dissolved oxygen, related with deeper and colder water (Figs. 43 and 38).



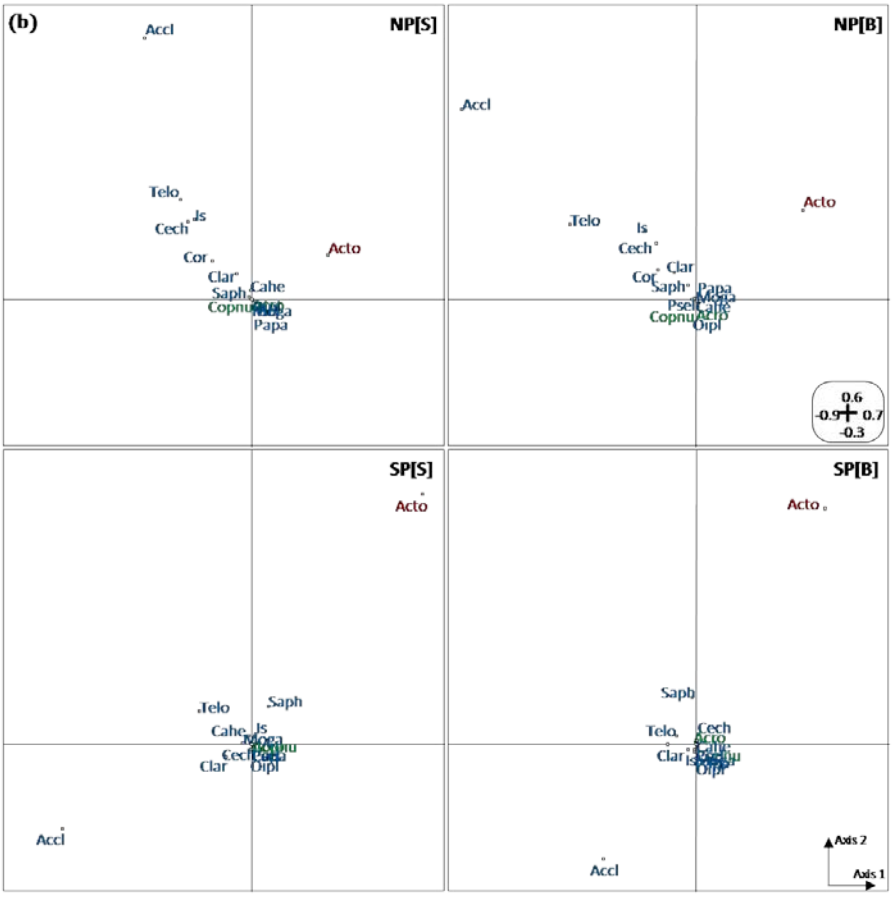


Figure 44. Results of the STATICO method. (a) Environmental intrastructure plot at each condition (neap surface, bottom and spring surface, bottom). (b) Species intrastructure plot at the same conditions. Environmental parameters codes as Figure 30. Species are coded as Table 1. (●) Marine species, (●) Freshwater species and (●) Estuarine species. Tide-depth codes as Figure 42.

On the other hand, with respect to trajectories plots (Fig. 44) the patterns observed also resembled CO-TUCKER first joint biplot. For instance, pH and dissolved oxygen also appeared in association to ACCL (the more abundant marine species) and in opposition to ACTO (the more abundant estuarine species). In addition, the inverse association between ACCL and temperature and chlorophyll *a*, presented a similar pattern (mainly for spring tides and less for neap tide, sub-surface, which is not atypical since this is the condition with minor cos^2; see Fig. 42).

However, looking in more detail some of the results, figure 38 also showed that pH, dissolved oxygen, POM, SPM, temperature and chlorophyll *a* reflects more environment-species interaction variability associated with the first component of the tide-depth mode, than depth and salinity. This is an important result of CO-TUCKER model, since it demonstrates that the method captures not what it is obvious but what differentiates the species assemblages. For instance, it is evident that marine species have positive interaction with high saline waters, while estuarine species have a preference of low saline waters. On the other hand, the positive/negative interactions detected with the high-potential parameters could be more relevant to explain these copepods community patterns distribution.

By observing the second joint biplot (Fig. 40) and the second more important interaction (g_{212}) captured by the CO-TUCKER model (which explains 15% of the variability, Table 7), it is important to note that, estuarine species as well as marine species shared the same interaction with salinity (negative during neap tides and positive during spring tides). Thus result was not clearly captured by STATICO plots (Figs. 43 and 44). The reason for this is probably within the purpose of the method, which is to find a consensus configuration and/or compare configurations. Thus, the method highlights what occurs in average and in a structured way (with a good performance and because it is inherent to its purpose), but does not put in evidence the corresponding parameters that more interact (in a non-consensual way) with the species assemblages. So, CO-TUCKER model assumes an important role, since it achieves results that emphasize what is not trivial and could be essential to the comprehension of ecosystem function.

Similarity as CO-TUCKER model, freshwater species (COPNU, ACRO) presented scores of nearly zero on the two components (meaning that, in general, they do not play an important role in the species-environment relationship pattern).

STATICO also enables to plot the projection of the samples in respect to the diel and tidal cycle, of each original table on the compromise axes, in terms of species abundance and environmental factors structures (Fig. 45).

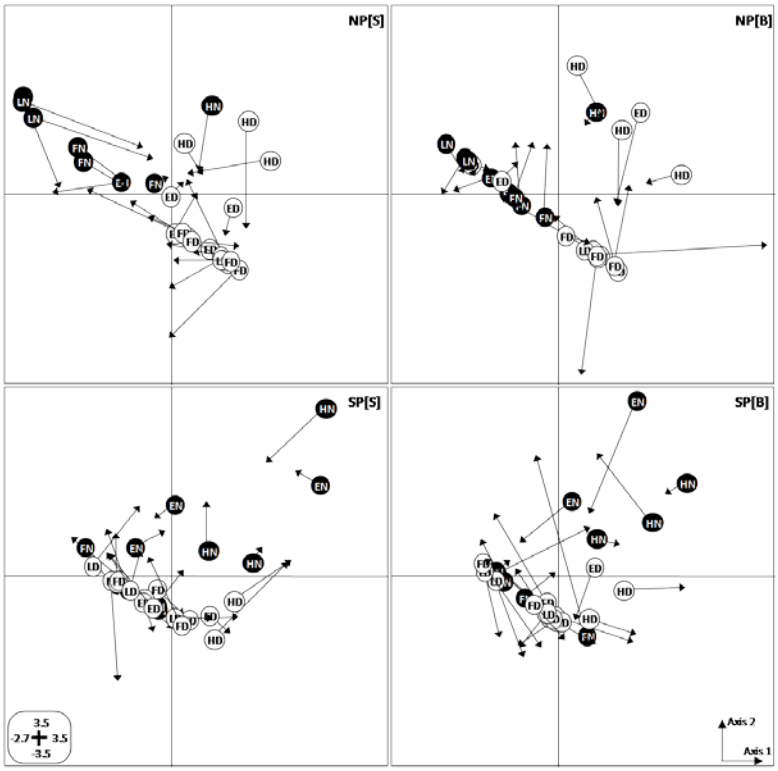


Figure 45. Trajectories factor plots of the STATICO analysis: projection of the samples in respect to tidal and diel cycle on the first factorial plan of the compromise analysis, for each tide-depth condition. Each sample is represented by two points: one is the projection of the row of the species table (circle, "•" or "o": origin of arrows), and the other is the projection of the row of the environmental table ("→", end of arrows). The length of the connecting line reveals the disagreement or the consensus between the two profiles (species–environment), i.e., the length of the line is proportional to the divergence between the datasets. When the datasets agree very strongly (that is, when variables strongly influence the species' distribution), the arrows will be short. Likewise, a long arrow demonstrates a locally weak relationship between the environment and copepod features for that case. L-Low tide, H-High tide, E-Ebb tide, F-Flood tide. Black and white circles represent samples of night and day, respectively. The scale of the graph is given in the rounded box.

Hence, it is possible to discuss the correlation between species distribution and environmental factors. This diagram can take advantage over joint biplots because highlights species-environment relationships by means of diel and tidal cycle. However, it is advantageous only when strong structures are under study. Otherwise the two-way correlation (and the identification of each species and/or environmental parameters that reveals the disagreement or the consensus) provided by figure 45 can only be general, since loss information occurs by "condensing" the two profiles (species–environment) in that way (that is, circles and arrows). Moreover by this diagram is not possible to quantify the intensity of the correlation. In this sense, the CO-TUCKER approach has an advantage, since from the information given by the core matrix that intensity is measured.

5.4.6 MOST IMPORTANT RESULTS OF THE COMPARASION

Some differences between the results of STATICO and the CO-TUCKER approach can be expected due to the different objectives of the methods. CO-TUCKER is a decomposition model, which fit the original data as well as possible, whereas STATICO reveals object distributions on the compromise plot. In this study STATICO and CO-TUCKER methods gave interpretable results, presented similar results and, in general, captured the same information about species-environment relationship.

STATICO is dedicated to visualize the agreement between species-environment relationships in each condition (here tide-depth) and between them. CO-TUCKER model allows easier visualization of the data structure highlighting each source of variability, in

particular capturing species assemblages' distribution which is common in ecological studies. It provides a better insight into data structure, since it take into account the three-dimensional structure of the data. Moreover CO-TUCKER model reduce the noise, and shows which of the original variables are correlated and which of them are the most significant for a certain environmental problem description. So, one of the major advantages of modeling ecological data sets with CO-TUCKER model is the gain in interpretability.

Nevertheless, Tucker models do not integrate the two data sets (species and environment sets) which makes more difficult to interpret the species-relationship by each condition. On the other hand, STATICO study the relationships between species and their environment through a pair of ecological tables but it cannot be generalized to more than three-way data and it is only exploratory tool since it cannot be used in a modeling context as can be done with multi-way models. In addition, STATICO captures what works in a structured way (by means of species-environment relationships), i.e., it shows what acts by consensus, without explaining the three-way interaction between the entities under study (here species × environment × tide-depth conditions).

Thus, CO-TUCKER is a combined approach of the two previous methods which takes advantage of both. Thus, the multivariate modeling of species data collected at four sampling tide-depth conditions provided by CO-TUCKER indicated that it is possible to obtain considerable information about a specific environment which could not be directly derived from the usual multi-tables methods.

CO-TUCKER could be more attractive for complex data sets analysis since it allows for choosing a different number of components per matrix. This could be very important because in many studies relevant information, or non-trivial, it is beyond the first two principal axes. This study is an example of this. The g_{322} element (explaining 4.1% of data variance) of the core matrix, although not the most important, is within the five most important interactions, enabling the devising environmental conditions for which species and tides interact positively or negatively.

Most of the multi-tables methods look for some resemblances between tables (or matrices) and not for differences. If there are mainly differences between data sets, then these methods will point out that there is no common structure and some explorations have to be made to explain them with other techniques. Moreover, as the first axes of these methods will show the main accordance between tables, a look at higher dimensions of three-way models can give information on the differences between both tables, that is, species-environment tables.

The decision whether the use of STATICO or of CO-TUCKER is entirely connected to the research objective as well as the how much the data structure is strong.

Given the two set of solutions offered by the two approaches (STATICO and CO-TUCKER), a fundamental question is how the solutions of the models are comparable in practice and whether one representation holds an advantage over the other. Clearly, there is no unequivocal answer to this question as it is data dependent. The important in

the selection of one or another multi-way model is to assess correctly whether any given solution is technically acceptable and can be interpreted.

However, the use of CO-TUCKER approach may lead to more interesting results as it combines the Co-Inertia Analysis with the core weights provided by the Tucker3 model and yield better understanding of patterns of variability associated to experimental factors.

Finally, it is important to note that the technique has been illustrated with species density and environmental parameters data. However other types of three-way data can be handled in a similar manner.

CONCLUSIONS

CONCLUSIONS

The comprehensive review of the literature in the field of ecology enabled:

1. Understanding the impact of descriptive multivariate methods for analysis of data organized in a two-way matrix, in a data cube, in two matrices and/or two cubes, concluding that the Correspondence Analysis and the Canonical Correspondence Analysis techniques are used in a high percentage of published articles. The use of STATIS family methods is much inferior, although growth was detected during the study period. Models of the Dutch school, despite capturing information that can be much more relevant, are practically unknown to ecologists.

2. In the literature, there is no procedure to describe not only the stable part of the relationship between two data sets (observed at different times or conditions), but also the changes in the structures.

A general framework for analyzing three-way multivariate data from ecological studies was proposed, in which it was shown that:

3. In many three-way data sets (objects × variables × conditions) the relevant information (or non-trivial) is beyond the first two principal components.

4. The use of multi-table methods to investigate the consensus configurations (as in the case of Partial Triadic Analysis and STATICO) provide, in many cases, “only”

the trivial information that is expressed through the first two principal components.

5. The use of such multi-tables methods only shows the common structure, and when this common structure is nonexistent some additional examinations should be performed.

6. The use of three-way models (for three-way data sets) gains an advantage, since it allows choosing a different number of components of each mode.

7. A new method of multivariate analysis to analyze two sets of three-way tables was developed, which, following the terminology of Thiolouse was called CO-TUCKER. The name CO-TUCKER means “CO-inertia and TUCKER", since first use Co-Inertia Analysis k times to compute the sequence of k cross-covariance tables, and then Tucker3 model to analyze this new k-table.

8. The model solves problems when the objective is to highlight the differences and not the common configurations within the data. Unlike other related models, CO-TUCKER model proves to have some intrinsic and unique properties due to its relationship with the STATICO and the Tucker3 model.

9. The fact that the new proposed model requires the use of calculation algorithms means that changes in the algorithm can provide modifications to the model.

10. With the new proposal was reached that:

    Because of parsimony and correspondence between the nature of the data and the model results of high structural quality, as well as more interpretable,

were achieved. In addition, through the core matrix and the joint biplots, these results capture better and / or more real predictions. It also highlights information about the differences between data tables and provides a better understanding of the patterns of variability associated with temporal and spatial changes of complex data sets.

11. The contribution of the group of experts in Marine Ecology, which facilitated the interpretation of results from analysis of the real data that was used, showed that the interaction between statisticians and ecologists not only enriches one or the other, but also the Science.

REFERENCES

REFERENCES

ASTEL, A., SIMEONOV, V., BAUER, H. & PUXBAUM, H. 2010. Multidimensional modeling of aerosol monitoring data. *Environmental Pollution,* 158**,** 3201-3208.

BARBIERI, P., ADAMI, G., PISELLI, P., GEMITI, F. & REISENHOFER, E. 2002. A three-way principal factor analysis for assessing the time variability of freshwaters related to a municipal water supply. *Chemometrics and Intelligent Laboratory Systems,* 62**,** 89–100.

BARBIERI, P., ANDERSSON, C. A., D.L., M., PREDONZANI, S., ADAMI, G. & REISENHOFER, E. 1999. Modeling bio-geochemical interactions in the surface waters of the Gulf of Trieste by three-way principal component analysis (PCA). *Analytica Chimica Acta,* 398**,** 227–235.

BENZÉCRI, J. P. 1973. L'analyse des données. *Dunod, Paris*.

BLANC, L., CHESSEL, D. & DOLÉDEC, S. 1998. Etude de la stabilité temporelle des structures spatiales par analyses d'une série de tableaux faunistiques totalement appariés. *Bulletin Français de la Pêche et de la Pisciculture,* 348**,** 1-21.

BRO, R. 1998. *Multi-way Analysis in the Food Industry. Models, Algorithms, and Applications.* PhD, Royal Veterinary and Agricultural University, Denmark.

BRO, R. & ANDERSSON, C. A. 1998a. Improving the speed of multi-way algorithms: Part I. Tucker3. *Chemometrics and Intelligent Laboratory Systems,* 42**,** 93-103.

BRO, R. & ANDERSSON, C. A. 1998b. Improving the speed of multiway algorithms. *Chemometrics and Intelligent Laboratory Systems,* 42**,** 105-113.

BRO, R. & SMILDE, A. 2003. Centering and scaling in component analysis. *Journal of Chemometrics,* 17**,** 16-33.

CARASSOU, L., BORGNE, R. L., ROLLAND, E. & PONTON, D. 2010. Spatial and temporal distribution of zooplankton related to the environmental conditions in the coral reef lagoon of New Caledonia, Southwest Pacific. *Marine Pollution Bulletin,* 61**,** 367-374.

CARASSOU, L. & PONTON, D. 2007. Spatio-temporal structure of pelagic larval and juvenile fish assemblages in coastal areas of New Caledonia, southwest Pacific. *Marine Biology,* 150**,** 697-711.

CARLIER, A., LAVIT, C., PAGÉS, M., PERNIN, M.-O. & TURLOT, J.-C. 1989. A comparative review of methods which handle a set of indexed data tables. *In:* COPPI, R. & BOLASCO, S. (eds.) *Multiway data analysis.* Amsterdam: Elsevier.

CEULEMANS, E. & KIERS, H. A. L. 2006. Selecting among three-mode principal component models of different types and complexities: A numerical convex hull based method. *British Journal of Mathematical and Statistical Psychology,* 59**,** 133-150.

CHESSEL, D. & HANAFI, M. 1996. Analyse de la co-inertie de K nuages de points. *Revue de statisque appliquée,* 44**,** 35-60.

COVAIN, R., DRAY, S., FISCH-MULLER, S. & MONTOYA-BURGOS, J. I. 2008. Assessing phylogenetic dependence of morphological traits using co-inertia prior to investigate character evolution in Loricariinae catfishes. *Molecular Phylogenetics and Evolution,* 46**,** 986-1002.

DECÄENS, T., JIMÉNEZ, J.-J. & ROSSI, J.-P. 2009. A null-model analysis of the spatio-temporal distribution of earthworm species assemblages in Colombian grasslands. *Journal of Tropical Ecology,* 25**,** 415-427.

DEMPSTER, A. P., LAIRD, N. M. & RUBIN, D. B. 1977. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B,* 39**,** 1-38.

DIGBY, P. G. N. & KEMPTON, R. A. 1987. *Multivariate analysis of ecological communities,* London, Chapman and Hall.

DOLÉDEC, S. & CHESSEL, D. 1994. Co-inertia analysis: an alternative method for studying species-environment relationships. *Freshwater Biology,* 31**,** 277-294.

DRAY, S., CHESSEL, D. & THIOULOUSE, J. 2003. Co-inertia analysis and the linking of ecological data tables. *Ecology,* 84**,** 3078-3089.

DUFOUR, A. B. 2009. Coinertia Analysis. *Fiche TD avec le logiciel R: course6* [Online], 6. Available: http://pbil.univ-lyon1.fr/R/pdf/course6.pdf.

ERNOULT, A., FREIRÉ-DIAZ, S., LANGLOIS, E. & ALARD, D. 2006. Are similar landscapes the result of similar histories? *Landscape Ecology,* 21**,** 631-639.

ESCOFIER, B. & PAGÈS, J. 1984. L' analyse factorielle multiple. *Cahiers du bureau universitaire de researche opérationnelle,* 42.

ESCOFIER, B. & PAGÈS, J. 1994. Multiple factor analysis (AFMULT package). *Computational Statistics and Data Analysis,* 18**,** 121-140.

ESTIENNE, F., MATTHIJS, N., MASSART, D. L., RICOUX, P. & LEIBOVICI, D. 2001. Multi-way modelling of high-dimensionality electroencephalographic data. *Chemometrics and Intelligent Laboratory Systems,* 58**,** 59-72.

FRANQUET, E., DOLÉDEC, S. & CHESSEL, D. 1995. Using multivariate analyses for separating spatial and temporal effects within species-environment relationships. *Hydrobiologia,* 300**,** 425-431.

GABRIEL, K. R. 1971. The biplot-graphic display of matrices with application to principal component analysis. *Biometrika,* 58**,** 453-467.

GAERTNER, J.-C. 2000. Seasonal organization patterns of demersal assemblages in the Gulf of Lions (north-western Mediterranean Sea). *Journal of the Marine Biological Association of United Kingdom,* 80**,** 777-783.

GALINDO, M. P. 1985. *Contribuciones a la Respresentación Simultánea de Datos Multidimensionales.* Ph D, Salamanca University.

GALINDO, M. P. 1986. Una alternativa de representación simultánea: HJ-Biplot. *Qüestiio,* 10**,** 13-23.

GALINDO, M. P. & CUADRAS, C. M. 1986. Una extension del método Biplot y su relación con otras técnicas. *Publicaciones de Bioestadística y Biomatemática. Universidad de Barcelona.,* 17.

GAUCH, H. G. J. 1982. Noise Reduction by Eigenvector Ordinations. *Ecology,* 63**,** 1643-1649.

GIUSSANI, B., MONTICELLI, D., GAMBILLARA, R., POZZI, A. & DOSSI, C. 2008. Three-way principal component analysis of chemical data from Lake Como watershed. *Microchemical Journal,* 88**,** 160-166.

GONÇALVES, A. M. M., PARDAL, M., MARQUES, S. C., MENDES, S., FERNÁNDEZ-GÓMEZ, M. J., GALINDO-VILLARDÓN, M. P. & AZEITEIRO, U. M. 2011a. Diel vertical behavior of Copepoda community (naupliar, copepodites and adults) at the boundary of a temperate estuary and coastal waters. *Estuarine, Coastal and Shelf Science***,** in press.

GONÇALVES, A. M. M., PARDAL, M., MARQUES, S. C., MENDES, S., FERNÁNDEZ-GÓMEZ, M. J., GALINDO-VILLARDÓN, M. P. & AZEITEIRO, U. M. 2011b. Response to Climatic variability of Copepoda life history stages in a southern European temperate estuary. *Zoological Studies***,** in press.

GREENACRE, M. J. 2010. Correspondence analysis of raw data. *Ecology,* 91**,** 958-963.

HENRION, R. 1994. N-way principal component analysis. Theory, algorithms and applications. *Chemometrics and Intelligent Laboratory Systems,* 25**,** 1-23.

HENRION, R. & ANDERSSON, C. A. 1999. A new criterion for simple-structure transformations of core arrays in N-way principal components analysis. *Chemometrics and Intelligent Laboratory Systems,* 47**,** 189-204.

HILL, M. O. 1974. Correspondence analysis: a neglected multivariate method. *Journal of the Royal Statistical Society,* 23**,** 340-354.

HOTELLING, H. 1933. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology,* 24**,** 498-520.

HOTELLING, H. 1936. Simplified calculation of principal components. *Psychometrika,* 1**,** 27-35.

JAFFRENOU, P. A. 1978. *Sur l'analyse des familles finies de variables vectorielles. Bases algébriques et application à la description statistique.* Thèse de 3° cycle, Sciences et Technique du Languedoc, Montpellier-II.

JOHNSON, K. W. & ALTMAN, N. S. 1999. Canonical correspondence analysis as an approximation to Gaussian ordination. *Environmetrics,* 10**,** 39-52.

JONGMAN, R. H., BRAAK, C. J. F. T. & TONGEREN, O. F. R. V. 1987. *Data analysis in community and landscape ecology,* Wageningen, PUDOC.

JUAN, A. D., RUTAN, S. C., TAULER, R. & MASSART, D. L. 1998. Comparison between the direct trilinear decomposition and the multivariate curve resolution-alternating least squares methods for the resolution of three-way data sets. *Chemometrics and Intelligent Laboratory Systems,* 40**,** 19-32.

KIERS, H. A. L. 1988. Comparison of “Anglo-Saxon” and “French” Three-Mode Methods. *Statistique et Analyse des Données,* 13**,** 14-32.

KIERS, H. A. L. & DER KINDEREN, A. 2003. A fast method for choosing the numbers of components in Tucker3 analysis. *British Journal of Mathematical and Statistical Psychology,* 56**,** 119-125.

KROONENBERG, P. M. 1983. *Three-mode Principal Component Analysis,* Leiden, DSWO Press.

KROONENBERG, P. M. 1989. The analysis of multiple tables in factorial ecology. III. Three-mode principal component analysis: ‘‘analyse triadique complète’’. *Acta Oecologica, Oecologia Generalis,* 10**,** 245–256.

KROONENBERG, P. M. 1992. Three-mode component models: a survey of the literature. *Statistica Applicata,* 4**,** 619-633.

KROONENBERG, P. M. 1994. The TUCKALS line: A suite of programs for three-way data analysis. *Computational Statistics & Data Analysis,* 18**,** 73-96.

KROONENBERG, P. M. 1996. 3WAYPACK User’s Manual. A package of three-way programs. 2.0 ed.

KROONENBERG, P. M. 2003. 3WayPack.

KROONENBERG, P. M. 2008. *Applied Multiway Data Analysis*, Wiley.

KROONENBERG, P. M. & BROUWER, P. 1985. User's guide to TUCKAL3. 4.0 ed. Leiden, Pays-Bas: Department of Education. Leiden University.

KROONENBERG, P. M. & DE LEEUW, J. 1980. Principal component analysis of three-mode data by means of alternating least-squares algorithms. *Psychometrika,* 45**,** 69-97.

KROONENBERG, P. M. & RÖDER, I. 2008. Situational Dependence of Emotions and Coping Strategies in Children with Asthma: A Three-Mode Analysis. *New Trends in Psychometrics*, 1-8.

KROONENBERG, P. M. & TEN BERGE, J. M. F. 2011. The equivalence of Tucker3 and Parafac models with two components. *Chemometrics and Intelligent Laboratory Systems,* 106**,** 21-26.

L' HERMIER DES PLANTES, H. 1976. *Structuration des Tableaux a Trois Indices de la Statistique: Theorie et Application d´une Méthode d´Analyse Conjointe.* PhD, Université des Sciences et Techniques du Languedoc.

LAIR, N. 2005. Abiotic vs. biotic factors: lessons drawn from rotifers in the Middle Loire, a meandering river monitored from 1995 to 2002, during low flow periods. *Hydrobiologia,* 546**,** 457-472.

LAVIT, C. 1988. *Analyse conjointe de tableaux quantitatifs,* Paris, Masson.

LAVIT, C., ESCOUFIER, Y., SABATIER, R. & TRAISSAC, P. 1994. The ACT (STATIS) method. *Computational Statistics & Data Analysis,* 18**,** 97-119.

LEGENDRE, L. & LEGENDRE, P. 1979. *Ecologie numerique. 1. Le traitement multiple des donnees ecologiques,* Paris, Masson et Le presses de l'université du Québec.

LEGENDRE, P. & LEGENDRE, L. 1998. *Numerical ecology,* Amsterdam, Elsevier.

MARQUES, S. C., AZEITEIRO, U. M., MARTINHO, F., VIEGAS, I. & PARDAL, M. 2009. Evaluation of estuarine mesozooplankton dynamics at a fine temporal scale: the role of seasonal, lunar and diel cycles. *Journal of Plankton Research,* 31**,** 1249-1263.

MARTÍNEZ-RUIZ, C. & FERNÁNDEZ-SANTOS, B. 2005. Natural revegetation on topsoiled mining-spoils according to the exposure. *Acta Oecologica,* 28**,** 231-238.

MARTÍNEZ-RUIZ, C., FERNÁNDEZ-SANTOS, B. & GÓMEZ-GUTIÉRREZ, J. M. 2001. Effects of substrate coarseness and exposure on plant succession in uranium-mining wastes. *Plant Ecology,* 155**,** 79–89.

MECHELEN, I. V. & SMILDE, A. K. 2011. Comparability problems in the analysis of multiway data. *Chemometrics and Intelligent Laboratory Systems,* 106**,** 2-11.

MENDES, S., FERNÁNDEZ-GÓMEZ, M. J., GALINDO-VILLARDÓN, M. P., MORGADO, F., MARANHÃO, P., AZEITEIRO, U. & BACELAR-NICOLAU, P. 2009a. Bacterioplankton dynamics in the Berlengas Archipelago (West coast of Portugal) using the HJ-biplot method. *Arquipélago. Life and Marine Sciences,* 26**,** 25-35.

MENDES, S., FERNÁNDEZ-GÓMEZ, M. J., PEREIRA, M. J., AZEITEIRO, U. & GALINDO-VILLARDÓN, M. P. 2010. The efficiency of Partial Triadic Analysis method: an ecological application. *Biometrical Letters,* 47**,** 83-106.

MENDES, S., FERNÁNDEZ-GÓMEZ, M. J., PEREIRA, M. J., AZEITEIRO, U. M. & GALINDO-VILLARDÓN, M. P. 2011a. An empirical comparison of Canonical Correspondence Analysis and STATICO in the identification of spatio-temporal ecological relationships. *Journal of Applied Statistics***,** in press.

MENDES, S., FERNÁNDEZ-GÓMEZ, M. J., RESENDE, P., PEREIRA, M. J., GALINDO-VILLARDÓN, M. P. & AZEITEIRO, U. M. 2009b. Spatio-temporal structure of diatom assemblages in a temperate estuary. A STATICO analysis. *Estuarine, Coastal and Shelf Science,* 84**,** 637-664.

MENDES, S., MARQUES, S. C., AZEITEIRO, U. M., FERNÁNDEZ-GOMÉZ, M. J., GALINDO-VILLARDÓN, M. P., MARANHÃO, P., MORGADO, F. & LEANDRO, S. M. 2011b. Zooplankton distribution in a Marine Protected Area: the Berlengas Natural Reserve (Western coast of Portugal). *Fresenius Environmental Bulletin,* 20**,** 496-505.

ORLÓCI, L. 1975. *Multivariate analysis in vegetation research,* The Hague, W. Junk B.V.

PAATERO, P. & ANDERSSON, C. A. 1999. Further improvements of the speed of the Tucker3 three-way algorithm. *Chemometrics and Intelligent Laboratory Systems,* 47**,** 17-20.

PALMER, M. W. 1993. Putting Things in Even Better Order: The Advantages of Canonical Correspondence Analysis. *Ecology,* 74**,** 2215-2230.

PARDO, R., VEGA, M., DEBÁN, L., CAZURRO, C. & CARRETERO, C. 2008. Modelling of chemical fractionation patterns of metals in soils by two-way and three-way principal component analysis *Analytica Chimica Acta,* 606**,** 26-36.

PAVOINE, S. & BAILLY, X. 2007. New analysis for consistency among markers in the study of genetic diversity: development and application to the description of bacterial diversity. *BMC Evolutionary Biology,* 7**,** 156-172.

PERÉ-TREPAT, E., GINEBREDA, A. & TAULER, R. 2007. Comparasion of different multiway methods for the analysis of geographical metal distributions in fish, sediments and river waters in Catalonia. *Chemometrics and Intelligent Laboratory Systems,* 88**,** 69-83.

POTTIER, J., MARRS, R. H. & BÉDÉCARRATS, A. 2007. Integrating ecological features of species in spatial pattern analysis of a plant community. *Journal of Vegetation Science,* 18**,** 221-228.

RESENDE, P., AZEITEIRO, U. & PEREIRA, M. J. 2005. Diatom ecological preferences in a shallow temperate estuary (Ria de Aveiro, Western Portugal). *Hydrobiologia,* 544**,** 77-88.

RIVAS-GONZALO, J. C., GUTIERREZ, Y., POLANCO, A. M., HEBRERO, E., VICENTE, J. L., GALINDO, P. & SANTOS -BUELGA, C. 1993. Biplot Analysis Applied to Enological Parameters in the Geographical Classification of Young Red Wines. *American Society for Enology and Viticulture,* 44**,** 302.

RIZZI, A. & VICHI, M. 1995. Representation, synthesis, variability and data preprocessing of three-way data set. *Computational Statistics & Data Analysis,* 19**,** 203-222.

ROLLAND, A., BERTRAND, F., MAUMY, M. & JACQUET, S. 2009. Assessing phytoplankton structure and spatio-temporal dynamics in a freshwater ecosystem using a powerful multiway statistical analysis. *Water Research,* 43**,** 3155-3168.

ROSSI, J.-P. 2006. The spatiotemporal pattern of a tropical earthworm species assemblage and its relation with soil structure. *Pedobiologia,* 47**,** 497-503.

SANTOS, C., MUÑOZ, S. S., GUTIÉRREZ, Y., HERRERO, E., VICENTE, J. L., GALINDO, P. & RIVAS, J. C. 1991. Characterization of young red wines by application of HJ-Biplot analysis to anthocyanin profiles. *Journal of Agricultural and Food Chemistry,* 39**,** 1086–1090.

SIMIER, M., BLANC, L., PELLEGRIN, F. & NANDRIS, D. 1999. Approche simultanée de K couples de tableaux: application à l'étude des relations pathologie végétale – environnement. *Revue de statisque appliquée,* 47**,** 31-46.

SIMIER, M., LAURENT, C., ECOUTIN, J.-M. & ALBARET, J.-J. 2006. The Gambia River estuary: A reference point for estuarine fish assemblages studies in West Africa. *Estuarine, Coastal and Shelf Science,* 69**,** 615-628.

SINGH, P. K., MALIK, A., SINHA, S., MOHAN, D. & SINGH, K. V. 2007. Exploring groundwater hydrochemestry of alluvial aquifers using multi-way modeling. *Analytica Chimica Acta,* 596**,** 171-182.

SINHA, S., BASANT, A. & MALIK, A. 2009. Multivariate modeling of chromium-induced oxidative stress and biochemical changes in plants of *Pistia stratiotes* L. *Ecotoxicology,* 18**,** 555-566.

SMILDE, A., BRO, R. & GELADI, P. 2004. *Multi-way Analysis. Applications in the Chemical Sciences*, John Wiley & Sons, Ltd.

SMITH, E. P. 2002. Ecological statistics. *In:* EL-SHAARAWI, A. H. & PIEGORSCH, W. W. (eds.) *Encyclopedia of Environmetrics.* Chichester: John Wiley & Sons, Ltd.

STANIMIROVA, I., WALCZAK, B., MASSART, D. L., SIMEONOV, V., SABY, C. A. & CRESCENZO, E. D. 2004. STATIS, a three-way method for data analysis. Application to environmental data. *Chemometrics and Intelligent Laboratory Systems,* 73**,** 219–233.

STANIMIROVA, I., ZEHL, K., MASSART, D. L., VANDER HEYDEN, Y. & EINAX, J. W. 2006. Chemometric analysis of soil pollution data using the Tucker N-way method. *Analytical and Bioanalytical Chemistry,* 385**,** 771–779.

TAULER, R. 1995. Multivariate curve resolution applied to second order data. *Chemometrics and Intelligent Laboratory Systems,* 30**,** 133-146.

TEN BERGE, J. M. F., DE LEEUW, J. & KROONENBERG, P. M. 1987. Some additional results on principal components analysis of three-mode data by means of alternating least squares algorithms. 52**,** 183-191.

TER BRAAK, C. J. F. 1985. Correspondence analysis of incidence and abundance data: properties in terms of a unimodal response model. *Biometrics,* 41**,** 859-873.

TER BRAAK, C. J. F. 1986. Canonical Correspondence Analysis: A New Eigenvector Technique for Multivariate Direct Gradient Analysis. *Ecology,* 67**,** 1167-1179.

TER BRAAK, C. J. F. 1987. The analysis of vegetation-environment relationships by canonical correspondence analysis. *Vegetatio,* 69**,** 69-77.

TER BRAAK, C. J. F. & LOOMAN, C. W. N. 1986. Weighted averaging, logistic regression and the Gaussian response model. *Plant Ecology,* 65**,** 3-11.

TER BRAAK, C. J. F. & PRENTICE, I. C. 1988. A Theory of Gradient Analysis. *Advances in Ecological Research,* 18**,** 271-317.

TER BRAAK, C. J. F. & SMILAUER, P. 2002. *CANOCO Reference manual and CanoDraw for Windows User's guide: Software for Canonical Community Ordination (version 4.5),* Ithaca, NY, Microcomputer Power.

TER BRAAK, C. J. F. & VERDONSCHOT, P. F. M. 1995. Canonical correspondence analysis and related multivariate methods in aquatic ecology. *Aquatic Sciences,* 57**,** 255-289.

THIOULOUSE, J. 2010. Simultaneous analysis of a sequence of paired ecological tables: a comparison of several methods. *Annals of Applied Statistics,* (in press)**,** 1-29.

THIOULOUSE, J. & CHESSEL, D. 1987. Les analyses multitableaux en écologie factorielle. I. De la typologie d'état à la typologie de fonctionnement par l'analyse triadique. *Acta Oecologica, Oecologia Generalis,* 8**,** 463-480.

THIOULOUSE, J., CHESSEL, D., DOLÉDEC, S. & OLIVIER, J. M. 1997. ADE-4: a multivariate analysis and graphical display software. *Statistics and Computing,* 7**,** 75-83.

THIOULOUSE, J., SIMIER, M. & CHESSEL, D. 2004. Simultaneous analysis of a sequence of paired ecological tables. *Ecology,* 85**,** 272–283.

TIMMERMAN, M. E. & KIERS, H. A. L. 2000. Three-mode principal components analysis: Choosing the numbers of components and sensitivity to local optima. *British Journal of Mathematical and Statistical Psychology,* 53**,** 1-16.

TUCKER, L. R. 1963. *Problems in measuring change,* Madison, University of Wisconsin Press.

TUCKER, L. R. 1966. Some mathematical notes on 3-mode factor analysis. *Psychometrika,* 31**,** 279–311.

TWATWA, N. M., VAN DER LINGEN, C. D., DRAPEAU, L., MOLONEY, C. L. & FIELD, J. G. 2005. Characterising and comparing the spawning habitats of anchovy *Engraulis*

encrasicolus and sardine *Sardinops sagax* in the southern Benguela upwelling ecosystem. *African Journal of Marine Science,* 27**,** 487-499.

VARELA, M., CROSSA, J., ARUN, K. J., PAUL, L. C. & MANES, Y. 2009. Generalizing the Sites Regression Model to Three-Way Interaction Including Multi-Atributes. *Crop Science,* 49**,** 2043-2057.

VELDSCHOLTE, C. M., KROONENBERG, P. M. & ANTONIDES, G. 1998. Three-mode analysis of perceptions of economic activities in Eastern and Western Europe. *Journal of Economic Psychology,* 19**,** 321-351.

VERDONSCHOT, P. F. M. & TER BRAAK, C. J. F. 1994. An experimental manipulation of oligochaete communities in mesocosms with chlorprifos or nutrient additions: multivariate analysis with Monte Carlo permutation tests. *Hydrobiologia,* 278**,** 251-266.

VIVIEN, M. & SABATIER, R. 2004. A generalization of STATIS-ACT strategy: DO-ACT for two multiblocks tables. *Computational Statistics & Data Analysis,* 46**,** 155-171.

APPENDIX

PUBLISHED ARTICLES BY THE AUTHOR

PAPER I

The study of bacterioplankton dynamics in the Berlengas Archipelago (West coast of Portugal) by applying the HJ-Biplot method

Mendes S, Fernández-Gómez MJ, Galindo-Villardón MP, Morgado F,

Maranhão P, Azeiteiro U, Bacelar-Nicolau, P

Arquipélago Life and Marine Sciences 26: 25-35, 2009 (#)

(#) Additional information:

- Publ. Ponta Delgada: University of the Azores
- ISSN: 0873-4704
- **Indexed in**: Aquatic Sciences and Fisheries Abstracts (ASFA), Biological Abstracts, BIOSIS Previews, Web of Science, Zoological Record. Directory of Open Access Journals (DOAJ), Latindex (meets 28 of the 33 required criteria).

The study of bacterioplankton dynamics in the Berlengas Archipelago (West coast of Portugal) by applying the HJ-biplot method

SUSANA MENDES [1,2], M.J. FERNÁNDEZ-GÓMEZ [2], M.P. GALINDO-VILLARDÓN [2], F. MORGADO [3], P. MARANHÃO [1,4], U. AZEITEIRO [4,5] & P. BACELAR-NICOLAU [4,5]

Mendes, S., M.J. Fernández-Gómez, M.P. Galindo-Villardón, F. Morgado, P. Maranhão, U. Azeiteiro & P. Bacelar-Nicolau 2009. The study of bacterioplankton dynamics in the Berlengas Archipelago (West coast of Portugal) by applying the HJ-biplot method. *Arquipélago*. Life and Marine Sciences 26: 25-35.

The relationship between bacterioplankton and environmental forcing in the Berlengas Archipelago (Western Coast of Portugal) were studied between February 2006 and February 2007 in two sampling stations: Berlenga and Canal, using an HJ-biplot. The HJ-biplot showed a simultaneous display of the three main metabolic groups of bacteria involved in carbon cycling (aerobic heterotrophic bacteria, sulphate-reducing bacteria and nitrate-reducing bacteria) and environmental parameters, in low dimensions. Our results indicated that bacterial dynamics are mainly affected by temporal gradients (seasonal gradients with a clear winter versus summer opposition), and less by the spatial structure (Berlenga and Canal). The yearly variation in the abundance of aerobic heterotrophic bacteria were positively correlated with those in chlorophyll *a* concentration, whereas ammonium concentration and temperature decreased with increasing phosphates and nitrites concentration. The relationship between aerobic heterotrophic bacteria, chlorophyll *a* and ammonium reveals that phytoplankton is an important source of organic substrates for bacteria.

Key words: Berlengas Natural Reserve, aerobic heterotrophic bacteria, sulphate-reducing bacteria, nitrate-reducing bacteria

Susana Mendes (e-mail: smendes@estm.ipleiria.pt); [1]GIRM-Research Group on Marine Resources, Polytechnic Institute of Leiria, School of Tourism and Maritime Technology, Campus 4, PT-2520-641 Peniche, Portugal; [2]University of Salamanca, Department of Statistics, ES-37007 Salamanca, Spain; [3]CESAM-Centre for Environmental and Marine Studies and Department of Biology, University of Aveiro, PT-3810-193 Aveiro, Portugal; [4]IMAR - Institute of Marine Research, University of Coimbra, Department of Zoology, PT-3004-517 Coimbra, Portugal; [5]Open University, Department of Sciences and Technology, PT-1269-001 Lisboa, Portugal.

The Berlengas Natural Reserve (BNR) is an archipelago formed by 3 groups of islands (Berlenga, Estelas and Farilhões) located 5.7 miles from Cabo Carvoeiro (Peniche). It was created in 1981, aiming to preserve a rich natural heritage and to ensure sustainable development of human activities in the area. The Reserve has a total area of 9560 hectares, 9456 of which are marine. The fauna is characteristic, very rich, diverse and peculiar (Rodrigues et al. 2008).

Despite its known biodiversity, the few existing marine scientific studies were carried out only after the establishment of the BNR. Azeiteiro et al. (1997) and Pardal & Azeiteiro (2001) studied the macrozooplankton distribution and several studies addressed the macrobenthos communities (Almeida 1996; Bengala et al. 1997a, 1997b, 1997c; Metelo 1999; Neto 1999; Neto et al. 1999).

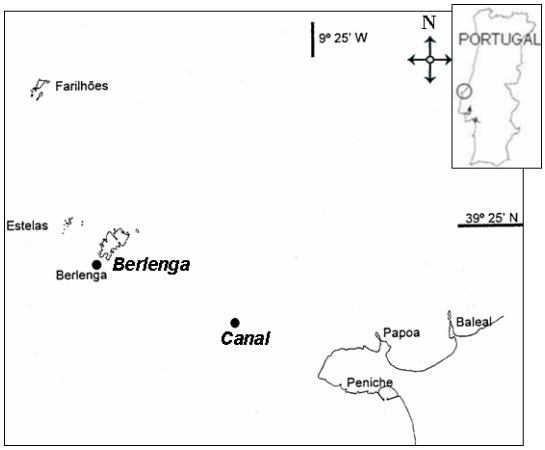


Fig. 1. Berlengas Archipelago with its 3 groups of islands and the location of the sampling stations (●).

Microbes are a key component of the structure and function of aquatic ecosystems, through their involvement in the various nutrient biogeochemical cycles namely in coastal areas (Wetz et al. 2008), and in the NW Iberian margin upwelling system (Barbosa et al. 2001).

Until now, no studies have been carried out to assess this component of the ecosystem in the BNR. Because of bacterial major role in planktonic microbial food webs (they comprise an important share of plankton biomass) and their activities impact on ecosystem metabolism and function (Gasol & Duarte 2000) this preliminary study in the BNR has a clear ecological relevance.

The present study aimed to investigate the annual dynamics (February 2006 – February 2007) of three main metabolic groups of heterotrophic bacteria: aerobic heterotrophic bacteria (AHB), sulphate-reducing bacteria (SRB) and nitrate-reducing bacteria (NRB), and to identify the key factors (among various physical and chemical parameters) that control the annual dynamics of the bacterioplankton in the Berlengas archipelago area, by an HJ-biplot (Galindo 1986; Galindo & Cuadras 1986). The HJ-biplot analysis is an extension of the classical biplots introduced by Gabriel in 1971 (Gabriel 1971, 1972; Gabriel & Odoroff 1990). The usefulness of HJ-biplot has been showed in many fields; for example in Biology (Martínez-Ruiz et al. 2001; Martínez-Ruiz & Fernández-Santos 2005; Cabrera et al. 2006; Martínez-Ruiz et al. 2007); in Geology (García-Talegón et al. 1999) and Medicine (Pedraz et al. 1985), etc.

MATERIALS AND METHODS

STUDY SITE AND FIELD PROGRAM

The BNR is an archipelago formed by 3 groups of islands (Berlenga, Estelas and Farilhões) located 5.7 miles from Peniche (Portugal). The sea floor consists primarily of irregular hard bottom substrate (i.e. rocks covered with sessile biota, including a variety of algae, sponges, hydrozoans, anemones, crustaceans, sea urchins and tunicates).

The two sampling sites used were Berlenga (inside BNR, 39º24.70' N, 9º31.22' W) and Canal (39º24.56' N, 9º26.70' W) (Fig. 1). Water samples were collected monthly from subsurface water (0.2 to 0.4 m depth) between February 2006 and February.

DETERMINATION OF PHYSICAL AND CHEMICAL PARAMETERS

Samples were analysed *in situ* for salinity and temperature (WTW MultiLine P4 portable meter). Water samples for chemical analyses and chlorophyll *a* quantification were collected and immediately stored in the dark at low temperature (4^{o} C), until further processing was possible. In the laboratory these water samples were filtered through GF/C filters (1.2 µm pore diameter), for quantification of photosynthetic pigments. Filtrates were used for determination of nutrient contents. Samples were also analysed in the laboratory (in triplicate) for their content in silica, nitrate, nitrite, ammonium, phosphate and chlorophyll *a* (Strickland & Parsons 1972).

SAMPLING PROCEDURE AND ENUMERATION OF VIABLE BACTERIA

Water samples (1 L) for the enumeration of viable bacteria were collected in parallel to those obtained for physical and chemical analyses. Samples were collected in sterile glass flasks, filled to capacity, sealed with gas-tight rubber stoppers and immediately placed on ice until processing in the laboratory (3 h later). Numbers of viable AHB, SRB and NRB were estimated by MPN method (Bacelar-Nicolau et al. 2003). Eight replicate ten-fold dilution series of water samples were prepared in appropriate selective liquid media (below) in multiwell plates (Bacelar-Nicolau et al. 2003). These were incubated at room temperature, in the dark, for 2 weeks (aerobes) or 4 weeks (anaerobes). The presence of bacteria was scored positive on the basis of turbidity or precipitate development, and confirmed by microscopic observation. Growth of SRB and NRB required strict anaerobic procedures that were used at all times (Strickland & Parsons 1972). YPG, VMN and the saltwater Widdel and Back media were used, respectively for AHB, SRB and NRB (Bacelar-Nicolau et al. 2003). In all media the concentration of NaCl was adjusted to 30 $g.L^{-1}$, and pH was adjusted to 7.

STATISTICAL ANALYSIS

Data were organized in a matrix: eight environmental variables and the abundance of the three metabolic groups of bacteria studied (AHB, SRB and NRB) at the two sites over the 12 months forming a data matrix of size 24 x 11 (the data were normalized by columns to homogenize the table).

The HJ-biplot is an exploratory data analysis method that looks for hidden patterns in the data matrix. The method is in some ways similar to correspondence analysis but is not restricted to frequency data. The HJ-biplot is a joint representation, in a low dimensional vector space (usually a plan), of the rows and columns of X**,** using markers (points/vectors), for its rows and for its columns. The markers are obtained from the usual singular value decomposition (SVD) of the data matrix. The rules for the interpretation of the HJ-biplot are a combination of the rules used in other multidimensional scaling techniques, correspondence analysis, factor analysis and classical biplots: 1) the distances among row markers are interpreted as an inverse function of similarities, in such a way that closer markers (individuals) are more similar. This property allows for the identification of clusters of individuals with similar profiles; 2) the lengths of the column markers (vectors) approximate the standard deviation of the variables; 3) the cosines of the angles among the column vectors approximate the correlations among variables in such a way that small acute angles are associated with variables with high positive correlations; obtuse angles near to the straight angle are associated with variables with high negative correlations and right angles are associated with non-correlated variables. In the same way, the cosines of the angles among the variable markers and the axes (Principal Components) approximate the correlations between them. For standardized data, these approximate the factor loadings in factor analysis and 4) the order of the orthogonal projections of the row markers (points) onto a column marker (vector) approximates the order of the row elements (values) in that column. The larger the projection of an individual point onto a variable vector is, the more this individual deviates from the average of that variable.

Several measurements are essential for a correct HJ-biplot interpretation (Galindo 1986; Galindo & Cuadras 1986). For instance, the relative contribution of the factor to the element is the relative variability of the variable explained by a factor or dimension. Also, for a point (row or column) on a factorial plan, the quality of representation can be defined by adding the two relative contributions of these factors to the element. Only points with a high quality of representation can be properly interpreted. These indices will be indicated, for the rows and for the columns of the data matrix. HJ-biplot has the advantage that it is a simultaneous representation and achieves an optimum quality of representation for both for rows and columns, both represented on the same reference system.

Calculations and graphs shown in this work were done using Classical Biplot software (Villardón [cited 2009; see Acknowledgements]).

RESULTS

The variation patterns of the environmental and biological parameters, at sites Berlenga and Canal are shown in Figure 2. Water temperature showed a similar unimodal pattern at both stations, with max.in summer (23.5 ºC) and min.in winter (13 ºC) (Figs. 2a and 2d). The concentration of chlorophyll *a* (Chl *a*) shows three peaks at both sites (Figs. 2a and 2d). At Berlenga, two moderate peaks of Chl *a* (May 2006 and January 2007) and broad maxima from July to October 2006 (Fig. 2a). At Canal, one moderate peak appeared in April 2006, and two higher peaks of Chl *a* appeared in October 2006 and December 2006 (Fig. 2d).

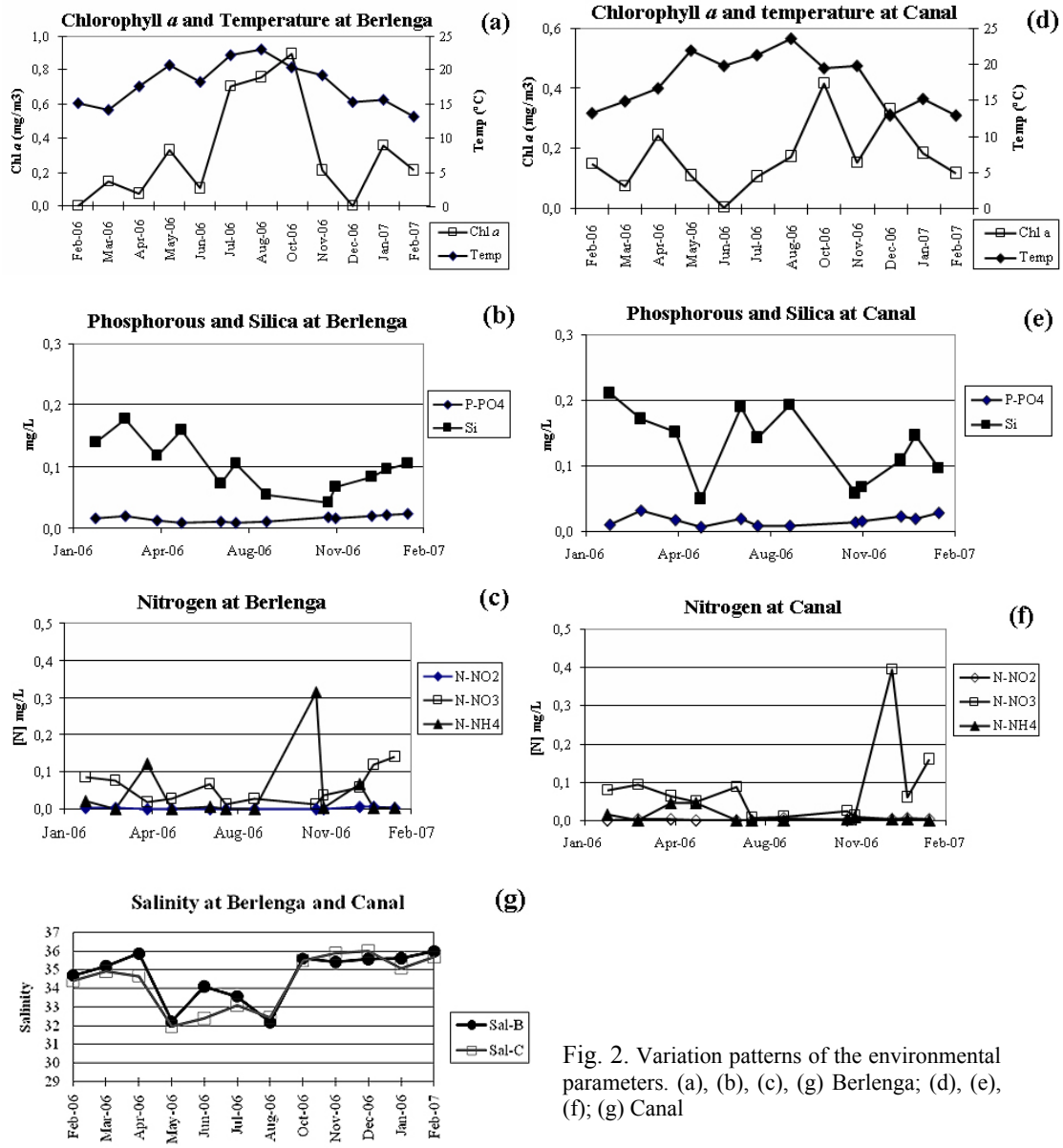


Fig. 2. Variation patterns of the environmental parameters. (a), (b), (c), (g) Berlenga; (d), (e), (f); (g) Canal

The concentration of silica showed different patterns in the two sites (Figs. 2b and 2e). At Berlenga, minimum values coincide with summer and autumn 2006, while at Canal there are two minimum values in May and October-November 2006. Phosphate concentration was fairly constant throughout the year and similar in the two sites, reaching maxima values in March 2006 (Figs. 2b and 2e). Salinity (Fig. 2g) varied between 32 and 36 with maxima in April 2006 and October to December 2006. Nitrogen dynamics are shown in Figs. 2c and 2f. Nitrite concentration was low and fairly similar at Berlenga and Canal. Nitrate and ammonium showed similar variation patterns, at the two sites, from January to September 2006, with a peak of ammonium concentration in April which coincided with a decrease of nitrate concentration. Between September and December 2006 patterns were different, with a peak of ammonium concentration at Berlenga in October, a peak of nitrate at Canal in November.

All metabolic groups were detected throughout the year at both sites (Fig. 3). AHB were, in average, present in greater numbers than SRB or NRB and showed greater number and fluctuations at Berlengas than Canal, with maxima values in February and August 2006 as well as between November and December 2006. SRB were, in average, present in greater numbers than NRB, at Berlengas, but the opposite was observed in Canal, where average numbers of NRB were greater than those of SRB. NRB showed similar patterns (and average numbers as well as standard deviation) and at both sites, although greater numbers at Canal, except for winter 2007 when numbers were greater at Berlengas.

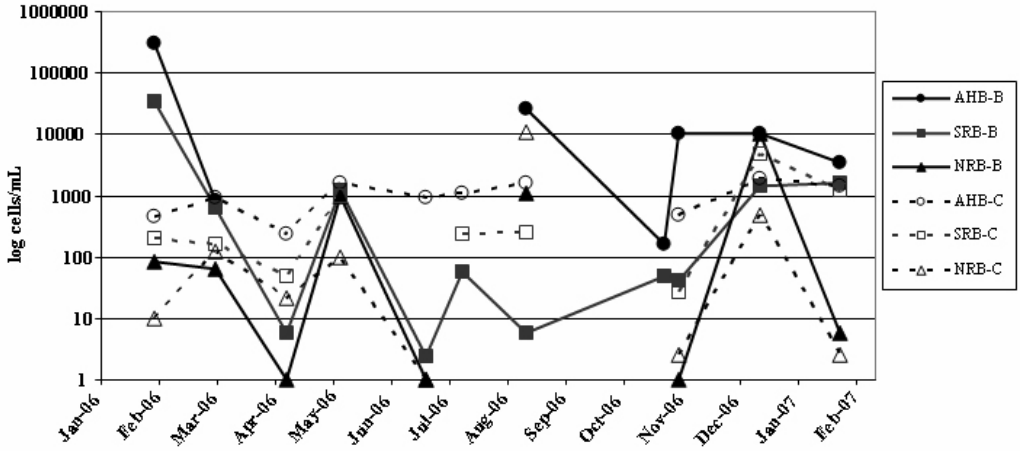


Fig. 3. Variation of physicochemical and biological parameters at Berlenga and Canal sampling sites during the annual cycle. AHB - aerobic heterotrophic bacteria, SRB - sulphate-reducing bacteria, NRB - nitrate-reducing bacteria. Points "B" and "C" correspond to the sites Berlenga and Canal.

The HJ-biplot analysis of environmental and biological parameters indicates that the system dynamics was independent of location (Berlenga, Canal) and mainly controlled by a seasonal gradient – spring/summer versus autumn/winter (Fig. 4a). The first three axes of the HJ-biplot analysis explained 63.52% of data variability (Fig. 4c and Table 1). The first axis (32.54%) revealed a strong opposition between samples collected in spring/summer and those collected in autumn/winter, independently of site of origin (Fig. 4a).

Table 1. Eigenvalues and Explained Variance (EV)

Axis	Eigenvalue	E.V. (%)	Cum. (%)
Axis 1	9.269	32.54	32.54
Axis 2	6.886	17.96	50.50
Axis 3	5.862	13.02	63.52
Axis 4	5.308	10.67	74.20

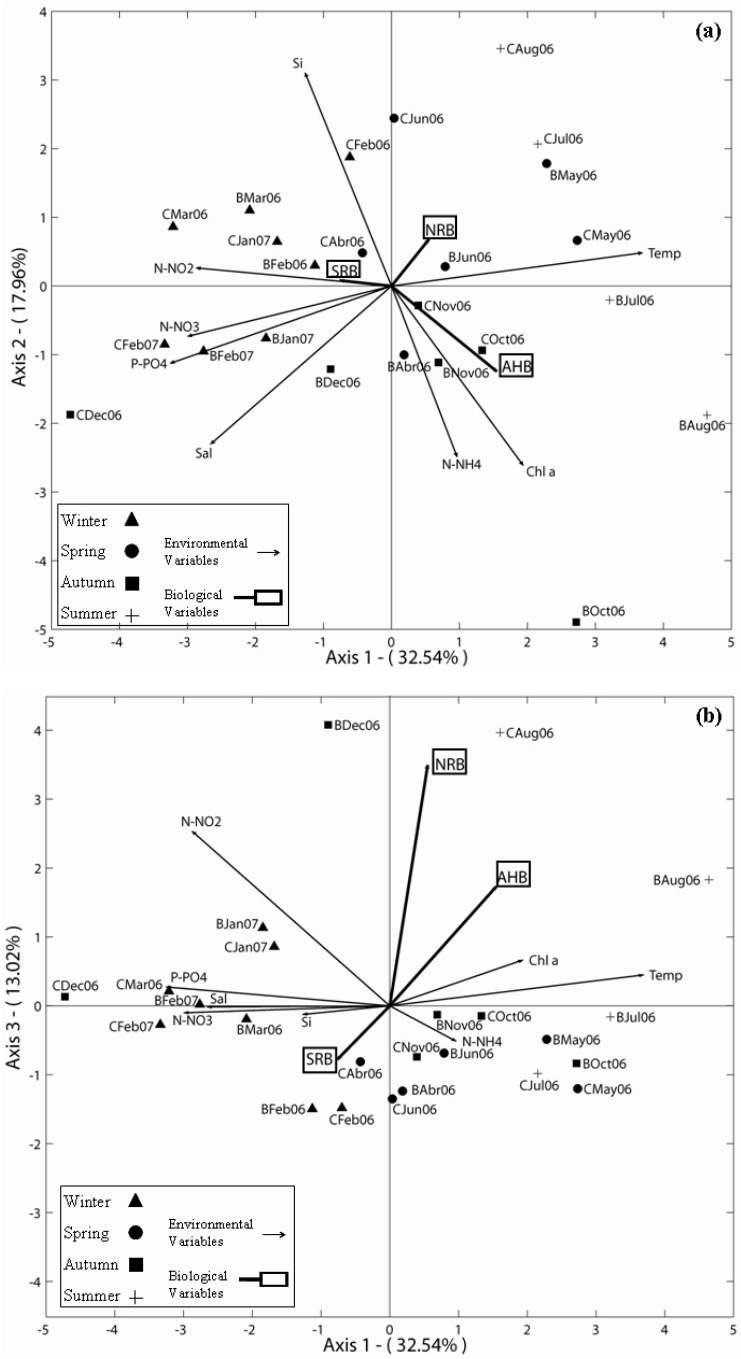


Fig. 4. HJ-Biplot representation. Projection of samples (points) and variables (arrows) onto plan of maximum inertia. Points "B" and "C" correspond to the sites Berlenga and Canal, the first three letters signify the month and the two second numbers indicate the year (2006 or 2007); AHB: aerobic heterotrophic bacteria, SRB: sulphate-reducing bacteria, NRB: nitrate-reducing bacteria; Si: silica, P-PO_4: phosphate, N-NO_2: nitrite, N-NO_3: nitrate, N-NH_4: ammonium, Chl *a*: chlorophyll *a*, S: salinity, T: Temperature. 4a) plan 1-2; 4b) plan 1-3.

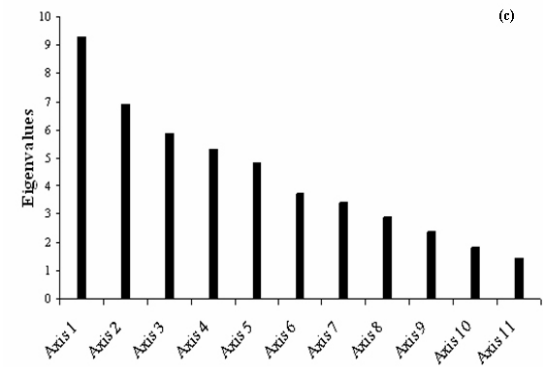


Fig. 4c). Eigenvalues barplot from HJ-Biplot analysis.

Moreover, this axis is characterized by phosphate, nitrate, nitrite, salinity and mostly by temperature. Therefore, the temperature characterizes the positive side of axis 1 (spring/summer samples), while phosphate, nitrate, nitrite and salinity appeared to be characteristic of the opposite side (autumn/winter samples at the left extreme of this axis). Hence, this is the gradient which defines axis 1. The temperature presented a negative correlation with those variables: higher values of temperature imply lower values of phosphate, nitrite, nitrate and salinity. Together, however, phosphate, nitrate, nitrite and salinity were positively correlated. The second axis of variability (17.96%) is mainly characterized by silica, chlorophyll *a* and ammonium (Fig. 4a). Silica defines the top of this axis and presented a negative correlation with the other two variables, chlorophyll *a* and ammonium (which are positively correlated). Furthermore, it was observed that chlorophyll *a* and ammonium are independent of nitrate and phosphate. The same relationship occurred between the pairs of salinity and chlorophyll *a*, silica and nitrate and salinity and phosphate. For the first principal plan AHB, NRB and SRB showed lower quality of representation (Table 2). This fact justifies the weak significance of these variables on the plan 1-2 (Fig. 4a). Particularly, in the plan 1-3 it is possible to observe that axis 3 is characterized by NRB (Fig. 4b and Table 2). Moreover, only in plan 1-4 and 3-4 (graphic representation not showed), AHB had a superior quality of representation (Table 2). These plans revealed that AHB increases with temperature and chlorophyll *a*, but decreases with the increase of phosphate, nitrite and ammonium concentration.

Specific factorial axes that most contributed are summarized in Table 2 (as a function of the relative contribution of the factor to the element, both for the samples as for the environmental and biological variables). Shaded cells indicate the different factorial axes where the greatest relationship with the samples (Table 2a) and environmental and biological variables (Table 2b) were observed.

Table 2. Relative contributions of the factor to the element. The number refers of its maximum quality of representation on each axis (on a scale of 1000):

Table 2a) Samples

Samples	Axis 1	Axis 2	Axis 3
BFeb06	37	3	65
BMar06	**607**	169	5
BAbr06	3	90	137
BMay06	**541**	330	25
BJun06	152	19	115
BJul06	**715**	3	2
BAug06	**468**	77	73
BOct06	154	**499**	15
BNov06	73	188	3
BDec06	35	64	**723**
BJan07	**493**	84	186
BFeb07	**821**	97	0
CFeb06	42	**304**	190
CMar06	**615**	44	3
CAbr06	80	104	**293**
CMay06	**500**	29	97
CJun06	0	**517**	158
CJul06	**439**	**404**	92
CAug06	72	336	**441**
COct06	**362**	178	4
CNov06	31	16	108
CDec06	**604**	95	0
CJan07	**520**	77	136
CFeb07	**831**	54	6

Table 2b) Variables

Variables	Axis 1	Axis 2	Axis 3	Axis 4
Si	99	**593**	1	23
P-PO4	**652**	78	5	1

Table 2b). Variables (continuation)

Variables	Axis 1	Axis 2	Axis 3	Axis 4
N-NO2	**508**	4	395	3
N-NO3	**553**	33	1	74
N-NH4	57	**381**	16	367
Chl *a*	231	**420**	27	10
Sal	**435**	327	0	24
Temp	**840**	14	12	2
AHB	148	95	185	**403**
SRB	36	0	37	190
NRB	19	29	**752**	78

DISCUSSION

The Portuguese west coast is affected by weak westerly winds during winter, alternating with relatively strong northerly winds during spring and summer. Because of the north-south orientation of the coast, the increase and steadiness of these northerly along-shore winds support the occurrence of coastal upwelling that, in this area, has its maximum intensity from April to September (Fiúza 1983; Moita et al. 2003; Resende et al. 2007). The period between August and October (late summer–early autumn) is generally characterized by a relaxation of the upwelling events (Fiúza 1983). During the upwelling period, primary production is enhanced in shelf areas (Tenore et al. 1995). The fate of this new dissolved and particulate matter largely depends upon the recycling efficiency within the pelagic microbial food web (Barbosa et al. 2001). Heterotrophic bacteria play a pivotal role in the biogeochemical cycles of marine ecosystems and are the major consumers of dissolved organic carbon (Barbosa et al. 2001). Besides the heterotrophic AHB, NRB and SRB two other metabolic groups were targetted (ammonium-oxidising bacteria and nitrite-oxidising bacteria) by the MPN method in the present study, although unsuccessfully, possibly due to their low numbers. Also carbon and sulphate were not determined, which would have been interesting, as they are central to understanding the biogeochemical cycles in which AHB, NRB and SRB are involved. It would have also been important to determine total bacterial numbers, in addition to the performed viable counts by MPN, however this was not possible to accomplish. Culturable viable bacterial numbers are lower than total bacteria and each set of data allows different analysis of information. In this work we chose to study dynamics of the three main metabolic bacterial groups involved in biogeochemical cycles which would have not been possible if considering the total number of bacteria.

In this study, AHB showed a general increase during late spring indicating that the main factors regulating their abundance in the studied aquatic systems are temperature, organic and inorganic substrates. In fact, the increase in abundance of AHB and also NRB during the annual period was mainly associated with increasing temperature and chlorophyll *a* and, to a lesser degree, ammonium, as found by other authors (Ducklow & Shiah 1993; Hoch & Kirchman 1993; Cunha et al. 2000; Bacelar-Nicolau et al. 2003). Moreover, globally AHB numbers decrease with the increase of phosphate, nitrite and nitrate concentrations. The relation between AHB and chlorophyll *a* and ammonium suggests that phytoplankton is an important source of organic substrates for bacteria (e.g. Barbosa et al. 2001; Bacelar-Nicolau et al. 2003). The chlorophyll *a* content is widely used as a measure of the organic carbon available to support oxygen consuming processes (Jonas & Tuttle 1990), where degradation of phyto-plankton-derived organic matter has an essential ecological role for carbon and nitrogen bio-geochemistry in coastal ecosystems (Wetz et al. 2008). The negative correlation existing between AHB numbers and silica concentration also sup-ports the importance of phytoplankton to AHB. Phosphorus (as phosphate) appeared as a limiting resource both to the bacterioplankton and phyto-plankton communities, as found by other authors (Currie 1990; Rivkin & Anderson 1997; Bacelar-Nicolau et al. 2003). In fact this may also be due to the fact that N:P ratio was above the Redfield only ratio value during November 2006 at Berlenga and December 2006 at Canal, and slightly below that value in April 2006 and May 2006 (not shown). At both sites, when N:P ratio was slightly below or above the Redfield value there was an peak in ammonium concentration, and a peak of nitrate in Canal.

The quality of representation (Table 2) of NRB and SRB was very weak on the plan 1-2, similar in plan 1-3 for the SRB, and NRB showed the best representation in axis 3. However, the observed opposition between SRB and both NRB, and AHB in the two first plans may be attributed to competition for carbon, as suggested for the Mondego estuary by Bacelar-Nicolau et al. (2003). NRB dynamics, besides being suggestively related to increasing temperature and chlorophyll *a* (upwelling months) also appeared to a great extent correlated with increasing nitrite concentrations (particularly in December and August 2006), a direct product of microbial denitrification (e.g. Herbert 1999) and with increasing silica concentrations (indicating decreasing diatom phytoplankton). This latter relationship is not yet understood.

The present work points to the strong involvement of AHB, NRB and SRB in the biogeochemical cycling of carbon and nitrogen, through the water column. Even though sites were not oxygen depleted (data not shown) anaerobic metabolisms may occur in particular microniches such as marine snow particles (e.g. Azam 1998). Future objectives will be to investigate the different proportions of bacterial mortality attributable to viral lysis or bacterivory at different stages of a phytoplankton bloom and to characterize bacterioplankton functional assemblages that transform specific components of the coastal seawater dissolved organic carbon. Future work also aims to further identify and investigate some of the viable cultivable bacteria particularly in relation to their relevance for biogeochemical cycles at the studied sites.

ACKNOWLEDGEMENTS

This study was funded by IMAR-Institute of Marine Research and by Instituto Politécnico de Leiria. The authors want to thanks to Júlio and João Laranjeira, the crew of the vessel *Julius* and to the students of the Escola Superior de Turismo e Tecnologia do Mar (Peniche) who helped in field work. The statistical analyses were run using the Classical Biplot software, available free of charge (http://biplot.usal.es/ClassicalBiplot/ [Villardón cited 2009]). The authors are grateful to the contributors who have made such a valuable tool available. The authors wish to thank the anonymous referees for their helpful and constructive comments on the manuscript.

REFERENCES

Almeida, A. 1996. Structure and spatial variability of the rocky fish fauna in the protected marine «Reserva Natural da Berlenga» (Portugal). *Arquivo. Museu Bocage*, Nova Série II: 633-642.

Azam, F. 1998. Microbial control of oceanic carbon flux: the plot thickens. *Science* 280: 694-696.

Azeiteiro, U., M. Pardal, J. Neto, I. Metelo & N. Bengala 1997. *Caracterização das comunidades planctónicas da Reserva Natural da Berlenga e seu possível padrão de migração: estudo preliminar*. IMAR – Instituto do Mar (CIC), Instituto da Conservação da Natureza (ICN) – Reserva Natural da Berlenga. 36 pp. [In Portuguese]

Bacelar-Nicolau, P., L.B. Nicolau, J.C. Marques, F. Morgado, R. Pastorinho & U.M. Azeiteiro 2003. Bacterioplankton dynamics in the Mondego estuary (Portugal). *Acta Oecologica* 24: 67-75.

Barbosa, A.B., H.M. Galvão, P.A. Mendes, P.A. Álvarez-Salgado, F.G. Figueiras & I. Joint 2001. Short-term variability of heterotrophic bacterioplankton during upwelling off the NW Iberian margin. *Progress in Oceanography* 51: 339-359.

Bengala, N., I. Metelo, J. Neto & M. Pardal 1997a. *Dinâmica anual de espécies chave da ilha da Berlenga*. IMAR – Instituto do Mar (CIC), Instituto da Conservação da Natureza (ICN) – Reserva Natural da Berlenga. 38 pp. [In Portuguese]

Bengala, N., I. Metelo, J. Neto & M. Pardal 1997b. *A ilha da Berlenga. Breve caracterização dos biótopos marinhos da zona intermareal e submareal*. IMAR – Instituto do Mar (CIC), Instituto da Conservação da Natureza (ICN) – Reserva Natural da Berlenga. 38 pp. [In Portuguese]

Bengala, N., I. Metelo, J. Neto & M. Pardal 1997c. *Variação sazonal das principais espécies de macroalgas da Reserva Natural da Berlenga*. IMAR – Instituto do Mar (CIC), Instituto da Conservação da Natureza (ICN) – Reserva Natural da Berlenga. 18 pp. [In Portuguese]

Cabrera, J.M.G., M.R.F. Martínez, E.J.M. Mateos & S. Vicente-Tavera 2006. Study of the evolution of air pollution in Salamanca (Spain) along a five-year period (1994–1998) using HJ-Biplot simultaneous

representation analysis. *Environmental Modelling & Software* 21: 61-68.

Cunha, M.A., M.A. Almeida & F. Alcântara 2000. Patterns of ectoenzymatic and heterotrophic bacterial activities along a salinity gradient in a shallow tidal estuary. *Marine Ecology Progress Series* 204: 1-12.

Currie, D.J. 1990. Large-scale variability and interactions among phytoplankton, bacterio-plankton, and phosphorous. *Limnology and Oceanography* 35: 1437-1455.

Ducklow, H.W. & F.K. Shiah 1993. Estuarine bacterial production. Pp. 261-264 in: Ford, T. (Ed.). *Aquatic Microbiology: An ecological approach*. Blackwell, London.

Fiúza, A. 1983. Upwelling patterns of Portugal. Pp. 85–98 in: Suess, E. & J. Thiede (Eds). *Coastal Upwelling, its Sedimentary record. Part A: Responses of the Sedimentary Regime to Present Coast Upwelling*, Plenum Press, New York.

Gabriel, K.R. 1971. The biplot-graphic display of matrices with application to principal component analysis. *Biometrika* 58: 453-467.

Gabriel, K.R. 1972. Analysis of meteorological data by means of canonical decomposition and Biplots. *Journal of Applied Meteorology* 11: 1071-1077.

Gabriel, K.R. & C.L. Odoroff 1990. Biplots in biomedical research. *Statistics in Medicine* 9: 469-485.

Galindo, M.P. 1986. Una alternativa de representación simultánea: HJ-Biplot. *Qüestiio* 10: 13-23. [In Spanish] [In Spanish]

Galindo, M.P. & C.M. Cuadras 1986. Una extensión del método Biplot y su relación con otras técnicas. *Publicaciones de Bioestadística y Biomatemática*. Universidad de Barcelona. 17 pp. [In Spanish]

García-Talegón, J., M.A. Vicente, E. Molina-Ballesteros & S. Vicente-Tavera 1999. Determination of the origin and evolution of building stones as a function of their chemical composition using the inertia criterion based on an HJ-biplot. *Chemical Geology* 153: 37-51.

Gasol, J.M. & C.M. Duarte 2000. Comparative analyses in aquatic microbial ecology: How far do they go? *FEMS Microbiology Ecology* 31: 99-106.

Herbert, R.A. 1999. Nitrogen cycling in coastal marine ecosystems. *FEMS Microbiology Ecology* 563-590.

Hoch, M.P. & D.L. Kirchman 1993. Seasonal and interannual variability in bacterial production and biomass in a temperate estuary. *Marine Ecology Progress Series* 98: 283-295.

Jonas, R.B. & J.H. Tuttle 1990. Bacterioplankton and Organic Carbon Dynamics in the Lower Meso-haline Chesapeake Bay. *Applied and Environ-mental Microbiology* 56: 747-757.

Martínez-Ruiz, C., B. Fernández-Santos & J.M. Gómez-Gutiérrez 2001. Effects of substrate coarseness and exposure on plant succession in uranium-mining wastes. *Plant Ecology* 155: 79-89.

Martínez-Ruiz, C. & B. Fernández-Santos 2005. Natural revegetation on topsoiled mining-spoils according to the exposure. *Acta Oecologica* 28: 231-238.

Martínez-Ruiz, C., B. Fernández-Santos, P.D. Putwain & M.J. Fernández-Goméz 2007. Natural and man-induced revegetation on mining wastes: Changes in the floristic composition during early succession. *Plant Ecology* 155: 79-89.

Metelo, I. 1999. Caracterização das comunidades de crustáceos da ilha da Berlenga. MSc Thesis, F.C.T.U., Coimbra. 77 pp. [In Portuguese]

Moita, M.T., P.B. Oliveira, J.C. Mendes & A.S. Palma 2003. Distribution of chlorophyll a and *Gymnodinium catenatum* associated with coastal upwelling plumes off central Portugal. *Acta Oecologica* 24: 125-132.

Neto, J. 1999. Variação espacial e temporal dos moluscos do substrato rochoso da ilha da Berlenga. MSc Thesis, F.C.T.U., Coimbra. 95 pp. [In Portuguese]

Neto, J., N. Bengala, I. Metelo & M. Pardal 1999. *Farilhões: breve caracterização do substrato rochoso da zona intermareal e submareal*. IMAR – Instituto do Mar (CIC), Instituto da Conservação da Natureza (ICN) – Reserva Natural da Berlenga. 21 pp.

Pardal, M. & U.M. Azeiteiro 2001. Zooplankton biomass, abundance and diversity in a shelf area of Portugal (the Berlenga Marine Natural Reserve). *Arquipélago.* Life and Marine Sciences 25: 49-62.

Pedraz, C., M.P. Galindo, C. De Hoyos, R. Escribano & V. Salazar-Villalovos 1985. Estudio de los factores socioculturales que influyen en la selección de lactancia natural. *Archivos de Pediatría* 36: 469-477. [In Spanish]

Resende, P., U. Azeiteiro, F. Gonçalves & M.J. Pereira 2007. Distribution and ecological preferences of diatoms and dinoflagellates in the west Iberian Coastal zone (North Portugal). *Acta Oecologica* 32: 224-235.

Rivkin, R.B. & M.R. Anderson 1997. Inorganic nutrient limitation of oceanic bacterioplankton. *Limnology and Oceanography* 42: 730-740.

Rodrigues, N.V., P. Maranhão, P. Oliveira & J. Alberto 2008. *Guia de Espécies Submarinas – Portugal, Berlengas* (1st edition). Instituto Politécnico de Leiria. 231 pp. [In Portuguese]

Strickland, J.D. & T.R. Parsons 1972. A practical hand-book of seawater analysis. *Bulletin of Fisheries Research Board*. Canada 167: 1-311.

Tenore, K.R., M. Alonso-Noval, M. Alvarez-Ossorio, L.P. Atkinson, J.M. Cabanas, R.M. Cal, H.J. Campos et al. 1995. Fisheries and oceanography off Galicia, NW Spain: Mesoscale spatial and temporal changes in physical processes and resultant patterns of biological productivity. *Journal of Geophysical Research* 100: 10943–10966.

Wetz, M.S., B. Hales & P.A. Wheeler 2008. Degradation of phytoplankton-derived organic matter: Implications for carbon and nitrogen biogeochemistry in coastal ecosystems. *Estuarine, Coastal and Shelf Science* 77: 422-432.

Accepted 19 November 2009.

PAPER II

The efficiency of Partial Triadic Analysis method: an ecological application

Mendes S, Fernández-Gómez MJ, Pereira MJ, Azeiteiro U,

Galindo-Villardón MP

Biometrical Letters 47: 83-106, 2010 (#)

(#) Additional information:

- Publ. Poznań : Polskie Towarzystwo Biometryczne
- ISSN: 1896-3811
- **Indexed in**: Current Index to Statistics and Polish Scientific Journal Contents

Biometrical Letters
Vol. 47 (2010), No. 2 , 83-106

The efficiency of the Partial Triadic Analysis method: an ecological application

Susana Mendes[1,2], Mª José Fernández Gómez[2], Mário Jorge Pereira[3], Ulisses Miranda Azeiteiro[4], Mª Purificación Galindo-Villardón[2]

[1]GIRM - Marine Resources Research Group, School of Tourism and Maritime Technology, Polytechnic Institute of Leiria – Campus 4, 2520-641 Peniche, Portugal, susana.mendes@estm.ipleiria.pt
[2]University of Salamanca, Department of Statistics, 37007 Salamanca, Spain, mjfg@usal.es, pgalindo@gugu.usal.es
[3]University of Aveiro, Department of Biology, 3810-193 Aveiro, Portugal, mverde@ua.pt
[4]Centre for Functional Ecology (CFE), Department of Life Sciences, University of Coimbra and Universidade Aberta, Department of Sciences and Technology, 4200-055 Porto, Portugal, ulisses@univ-ab.pt

SUMMARY

In this paper we present a Partial Triadic Analysis (PTA) method that can be applied to the analysis of series of ecological tables. The aim of this method is to analyse a three-way table, seen as a sequence of two-way tables. PTA belongs to the family of STATIS methods and comprises three steps: the interstructure, the compromise and the trajectories. The advantage of this method is related to the fact that it works with original data instead of operators, which permits all the interpretations to be performed in a directly way. In this study we present an efficient application of the PTA method in the simultaneous analysis of several data tables and show how well-adapted it is to the treatment of spatio-temporal data. Two kinds of matrices were constructed: a species abundance table and an environmental variables table. Both matrices had the sampling sites in rows. All computations and graphical displays were performed with the free software ADE-4. An example with phytoplankton and environmental factors data is analysed, and the results are discussed to show how this method can be used to extract the stable part of species and environment relationships.

Key words: Partial triadic analysis, multi-table analysis, STATIS, species abundance, environmental factors

1. Introduction

Many generalizations of standard linear multivariate analysis, like principal component analysis (PCA) or canonical correlation analysis (CCA), have been

proposed for studying three or more sets of variables. In this paper, we present an application of the Partial Triadic Analysis (PTA) method (Jaffrenou 1978), a multi-table technique, using a simple ecological data set. In particular we analysed the main temporal structure of the species assemblages and their spatial changes (and did the same for the environmental factors).

Introduced in ecological studies by Thioulouse and Chessel (1987), the PTA method aims at investigating three-dimensional data analysis (e.g. a data cube) seen as a sequence of two-way tables. In PTA all the tables must have the same rows and the same columns, but its advantage or potential is related to the fact that it works with original data instead of operators, which permits all the interpretations of the results to be performed in a direct manner. This method belongs to the family of STATIS methods, and in comparison, the PTA used in the present work allowed the optimal projection of trajectories. For example, Gaertner (2000) made an approach to studying the organization patterns of demersal species in the Gulf of Lions on a seasonal scale. However Rossi (2006) and Ernoult et al. (2006) also used the PTA for other problems. In particular, Ernoult et al. (2006) investigated the overall landscape variability through its different dimensions (space vs. time) and demonstrated the relative importance of each dimension.

Starting from an ecological perspective, the application of this method here aims to analyse the stability of the seasonality across sites of dinoflagellates assemblage in the near-shore shallow coastal area (300 m from the coast and before the surf zone) off the north-western Portuguese coast, in terms of bio-ecological categories as well as studying the stability across the sample sites of the temporal covariation structures between some environmental variables. The data for the Vila Praia de Âncora coast used in this work have been previously analysed and have already been published by Resende et al. (2007), but in a different context, with a different statistical approach and with one more phytoplankton community, the diatoms. In that work, the data were especially designed to identify the environmental variables governing the composition and structure of the species assemblages. The data were analysed from a global

point of view by performing a CCA (Canonical Correspondence Analysis) (ter Braak 1986), so the between-site and within-site variability cannot be separated, which may present a problem when working with phytoplankton communities. Therefore the results may fail to be significant and reasonable.

Our purpose, therefore, is to take an approach to these data, taking into account their three-dimensional structure, analysing the dinoflagellates data cube and the environmental one by means of an appropriate multi-table analysis technique, since with this kind of data the examinations of simultaneous data sets are, in a general way, a recent practice – in particular with phytoplankton data, so we could investigate the common temporal structure derived from each site: of the dinoflagellates and of the environmental variables.

The mathematical description of PTA is presented using simple matrix notations.

2. The partial triadic analysis method

To analyse each one of the data cubes, the statistical approach used was the PTA (Blanc et al. 1998; Thioulouse and Chessel 1987). The aim of this method is to identify the structure which is common to the series of tables having the same rows (n) and the same columns (p). More precisely, PTA searches for structures that are stable among the sequence of tables.

Let K be the number of tables with n lines and p columns. The intersection of line i with the column j gives the value of the variable j at the condition i. After the initial transformations (by centralization, normalization, etc) the K tables $\mathbf{X}_k$ are obtained. Each $\mathbf{X}_k$ is a data matrix of J quantitative variables measured on the same I observations (or objects), where each element is $x_{i,j}^{k}$. According to the PTA methodology, $(\mathbf{X}_k,\mathbf{D}_p,\mathbf{D}_n)$ defines a statistical triplet, where $\mathbf{D}_p$ and $\mathbf{D}_n$ are positive definite weighting matrices for variables and observations and whose positive diagonal elements sum to 1. The PTA is a three-step procedure, namely the interstructure, the compromise and the

intrastructure analyses (Lavit *et al.* 1994). Below, these three steps are explained with a description of PTA in matrix form.

First a matrix of scalar products between tables is computed (i.e., the matrix whose elements are: COVV($\mathbf{X}_k$,$\mathbf{X}_l$) = tr($\mathbf{X}_k^T \mathbf{D}_n \mathbf{X}_l \mathbf{D}_p$)). The diagonalization of this matrix provides eigenvectors. The k coefficients α_k of the first eigenvectors are then used to weight the *k* tables in the calculation of the compromise table. Then, a PCA is performed in order to establish the ordination of the different matrices. Alternatively, a matrix of vector correlations (RV) can be used to rescale the importance of the tables. Each element in this table is:

$$\mathbf{RV}(\mathbf{X}_k, \mathbf{X}_l) = \frac{\mathrm{COVV}(\mathbf{X}_k, \mathbf{X}_l)}{\sqrt{\mathrm{VAV}(\mathbf{X}_k)\mathrm{VAV}(\mathbf{X}_l)}}, \quad (1)$$

where VAV($\mathbf{X}_k$) is the variance of the vector obtained by putting all the columns of table $\mathbf{X}_k$ one below the other. It is basically the vector variance of table $\mathbf{X}_k$, i.e. VAV($\mathbf{X}_k$) = tr($\mathbf{X}_k^T \mathbf{D}_n \mathbf{X}_k \mathbf{D}_p$). The vector correlation matrix and the vector covariance matrix are linked by the same relationships as the normal correlation and covariance matrices. Each table is projected onto the factorial plan obtained from the analysis and represented by an arrow, in order to establish the ordination of the different tables, which summarizes the global structure and the relationships between tables. This configuration (based upon the covariance matrix) allows an overall graphical comparison of the tables and shows proximities between the configurations of the same observations.

The second step of this method is analysis of the compromise, a fictitious table which is computed as the weighted mean of all the tables of the series, using the components of the first eigenvector of the interstructure as weights (i.e. issued from the eigenvalues of the vector covariance matrix) (Thioulouse et al. 2004). In other words, it consists in calculating a linear combination of the *k* initial tables with the aim of constructing a mean table of maximum inertia (Gaertner 2000):

$$\mathbf{X}_c = \sum_k \alpha_k \mathbf{X}_k , \quad (2)$$

where $\mathbf{X}_c$ represents the compromise and captures (optimally) the similarities among the individual matrices. Once obtained, $\mathbf{X}_c$ (which has the same dimensions and the same structure and meaning as the tables of the series) is then analyzed by principal component analysis (PCA) and the rows and columns of the individual matrices are projected onto the analysis as supplementary individuals and supplementary variables, respectively. Thus analysis of the compromise gives a factor map that can be used to interpret the structures of this compromise. In other words, it gives a picture of the structures which are common to all the tables (Thioulouse et al. 2004).

The third step summarizes the variability of the succession of tables in comparison to the common structure defined by the compromise. The rows and columns of all the tables of the three dimensional array are projected onto the factor map of the PCA of the compromise as additional elements (Thioulouse et al. 2004) in order to summarize the reproducibility of the structure across the series of tables. Denote by $\mathbf{U}$ the matrix of the eigenvectors of the analysis of the compromise. The coordinates of the rows of the table $\mathbf{X}_k$ are:

$$\mathbf{R}_k = \mathbf{X}_k\mathbf{D}_p\mathbf{U}, \tag{3}$$

and the coordinates of its columns are

$$\mathbf{C}_k = \mathbf{X}_k^T\mathbf{D}_n\mathbf{X}_c\mathbf{D}_p\mathbf{U}\boldsymbol{\Lambda}^{-1/2}, \tag{4}$$

$\boldsymbol{\Lambda}^{-1/2}$ being the diagonal matrix of the inverses of the square root of the eigenvalues of the analysis of the compromise.

Each row of each table is represented by a point in the space of its p columns, and can be projected as a supplementary individual onto the principal axes of compromise. The same procedure is applied (similarly) for the columns (Simier et al. 1999). The points can then be linked, for example by lines, to underline their trajectories; their study constitutes the third step of the method.

For the dinoflagellates and environmental data studied in this work, PTA offered the possibility of studying these three-dimensional data in the way that Figure 1 shows, and studying the dynamic trajectories of the species and

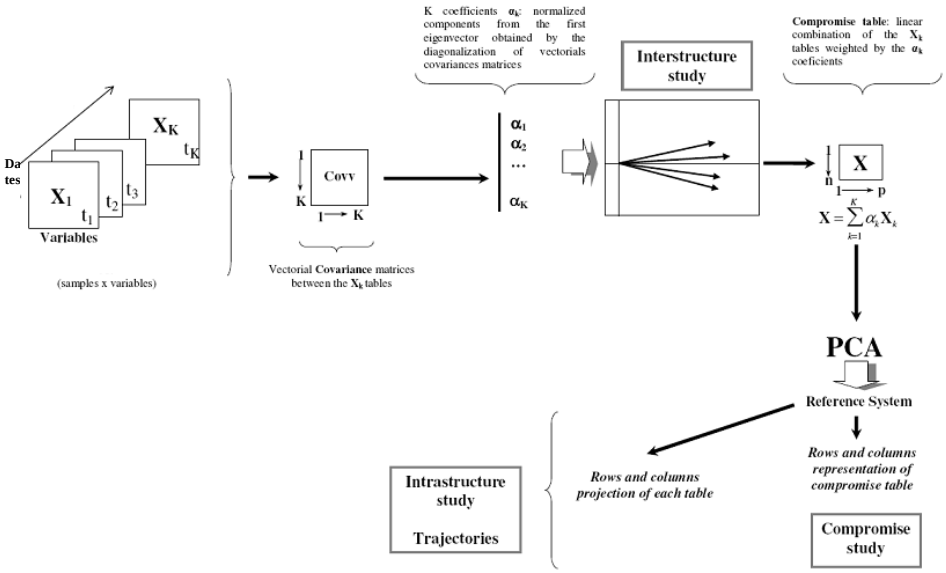


Figure 1. General scheme of the PTA: construction of the interstructure matrix and extraction of the compromise table.

environmental variables per site (each site considers one sampling station in one of the two conditions: high and low tide; data were considered as a series of tables for each site, i.e. table-site: dates in rows vs. variables/species in columns). The main aim in this study is to identify the common temporal structure derived from each station table. The calculations and graphs shown in this work were made using ADE-4 software (Thioulouse et al. 1997). This software is available free of charge from http://pbil.univ-lyon1.fr/ADE-4. Species abundances were transformed to *log(x + 1)* prior to the calculations, in order to minimize the dominant effect of exceptional catches.

3. Application Example

Data

The example data set was extracted from Resende et al. (2007). The data used in this work was collected in Vila Praia de Âncora coast (Fig. 2), which is located

on the north-western tip of Portugal (41º49.26'N; 8º51.50'W). Vila Praia de Âncora's coast is characterized by a vast rocky shore and a small beach, which forms a sandy inlet. The beach receives the estuarine waters of the river Âncora (Resende et al. 2007).

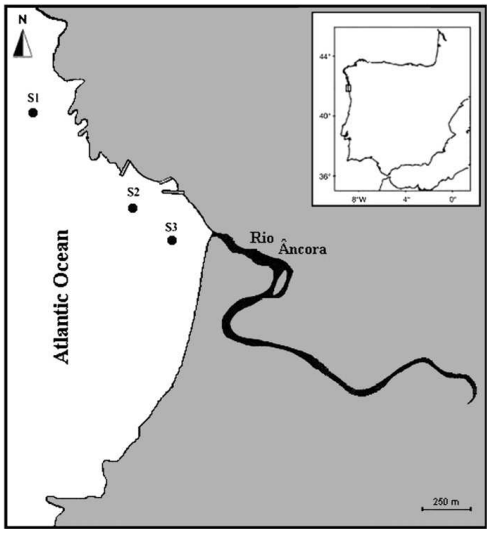


Figure 2. Map of the Northwest Portuguese coast and the study area with location of the three sampling sites (Resende et al. 2007)

Three sites were sampled near the shore (Fig. 2): S1, facing the rocky shore (41º49.26'N; 8º52.64'W); S2, adjacent to the fishing port (41º48.83'N; 8º52.24'W); and S3, opposite the Âncora estuary (41º48.27'N; 8º52.11'W). Sampling took place monthly, in daylight hours, always at full moon, at low tide (L) and high tide (H), from August 2002 to October 2003. A detailed description of the sampling sites is given by Resende et al. (2007). In total, 90 samples were collected between August 2002 and October 2003: 30 at S1, 30 at S2 and 30 at S3.

During the study period, the following environmental data were collected, in situ, for each site: pH, salinity, water temperature, dissolved oxygen and transparency (Secchi disc). Water samples for chemical analysis and chlorophyll *a* and volatile solids quantification were collected at the water subsurface. The concentrations of dissolved nutrients were also measured:

nitrate, nitrite, ammonia, phosphate. A more detailed description of the environmental methodologies and temporal variation of the environmental parameters can be found in Resende et al. (2007). The N:P ratio and the zooplankton biomass were also included. Zooplanktonic oblique tows were made at 1.5–2 km, and a detailed description of the sampling methodology is given by Azeiteiro et al. (2006).

Samples for taxonomic and quantitative dinoflagellates study were collected with a glass bottle (1 litre capacity) and immediately preserved with Lugol 1% (iodine/iodide potassium and distilled water). Only the armoured dinoflagellates were recorded and identified to the lowest possible taxon. The taxa selected for the investigation are listed in Table 1 (Resende et al. 2007).

Organization of matrices

The data were organized in two three-way tables: one for environmental factors (dates x variables x sites) and another one for the species abundances (dates x species x sites). Three sampling sites were considered in two conditions – high tide and low tide. Consequently, each multi-table was made up of six matrices. Fifteen dates were considered, from August 2002 to October 2003, being the same for each type of data matrix. All the species matrices had the same species, and all the environmental matrices had the same variables. Hence the data could be seen as "data cubes": a "species data cube" and an "environmental data cube", each one presented as a sequence of two-way tables.

4. Results

Interstructure analysis – similarity between different stations

The map of the interstructure analysis corresponds to a global representation, presents an ordination of the sampling sites and shows the vectors for individual stations on the plan made by the first and second axes, and consequently the similarities between stations tables. For environmental factors and dinoflagellates abundances the decreasing values of the eigenvalues (Fig. 3a,c)

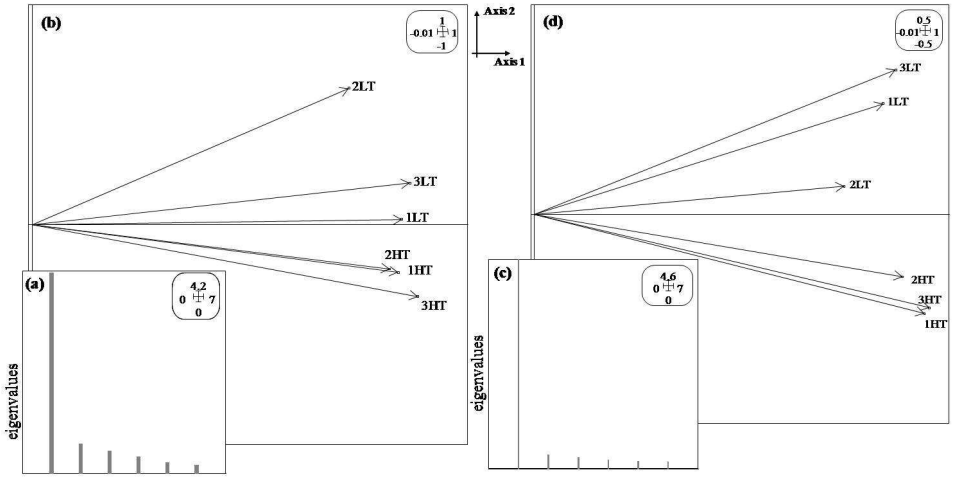


Figure 3. Interstructure analysis and eingenvalue diagrams: (a) Eingenvalue diagram of the series of environmental factors tables; (b) Interstructure factor map of the series of environmental factors tables; (c) Eingenvalue diagram of the series of dinoflagellates abundance tables; (d) Interstructure factor map of the series of dinoflagellates abundance tables. Axis 1 is the first principal component; Axis 2 is the second principal component. The scale of the graph is given by the box in the top right; 1HT – Station 1 at high tide; 1LT – Station 1 at low tide; 2HT – Station 2 at high tide; 2LT – Station 2 at low tide; 3HT – Station 3 at high tide; 3LT – Station 3 at low tide. Note the different scales.

allowed exploration of the first two factorial axes. Indeed, spatial variations of environmental factors and dinoflagellates are mainly projected on the first axis. All the sites display the same sign on the principal axis 1 (69% and 81% of the total inertia for environment and dinoflagellates, respectively), whilst axis 2 (10% and 6% of the total inertia for environment and dinoflagellates, respectively) presented two distinct groups: an opposition between low tide sampling stations and the high tide stations (Fig. 3b,d). All stations with the same sign had a positive correlation between the corresponding set of matrices (the stations' vectors on the first axis presented a uniform distribution) and indicates a relatively strong common sites structure, which indicates a large similarity among stations. The structure expressed through the first axis of the interstructure therefore corresponded to an environmental (Fig. 3b) and

dinoflagellates (Fig. 3d) temporal pattern common at the different sampling stations. The only isolated site appeared to be 2LT (this environment appears with a high positive value on the second axis, Fig. 3b). Besides this, in the interstructure analysis of dinoflagellates three stations come out more closely than the others: 1, 2 and 3, at high tide, with negative values on the second axis. However, detailed description of spatial variations of environmental factors and dinoflagellates were not necessary and were omitted in the further analyses.

Compromise analysis

Figure 4a and Figure 6a shows the factor maps of the compromise, for the environmental variables (Fig. 4a) and for the dinoflagellates species (Fig. 6a). Additionally, Figure 5a and Figure 7a show the factor maps of the compromise, for the sampling dates related to environmental variables (Fig.5a) and to the dinoflagellates species (Fig. 7a).

Stable part of the environment and sampling dates

Analysis of the compromise for environmental variables was carried out to reveal the common temporal pattern and to better explain the differences/similarities among stations. The graphical illustration of the analysis of the compromise (Fig. 4a) shows the average position of each environmental parameter in respect to the first and second axes, and the projection onto the compromise plan of the fifteen sampling dates (Fig. 5a) shows the temporal dynamics of the environmental factors for a mean station.

The first two axes of the analysis of the compromise account for 52% of the cumulative inertia, hence providing a summary of the environment attributes' spatial dynamics. Interpretation of this figure provides a good summary of the spatial environmental dynamics. The length of the arrows on the factor map of the compromise in the PTA analysis indicates that the most relevant environmental variables (variables with long arrows contributed more to the definition of the axes, compared with the variables with short arrows) are temperature, pH and volatile solids followed by, in decreasing order, nitrates,

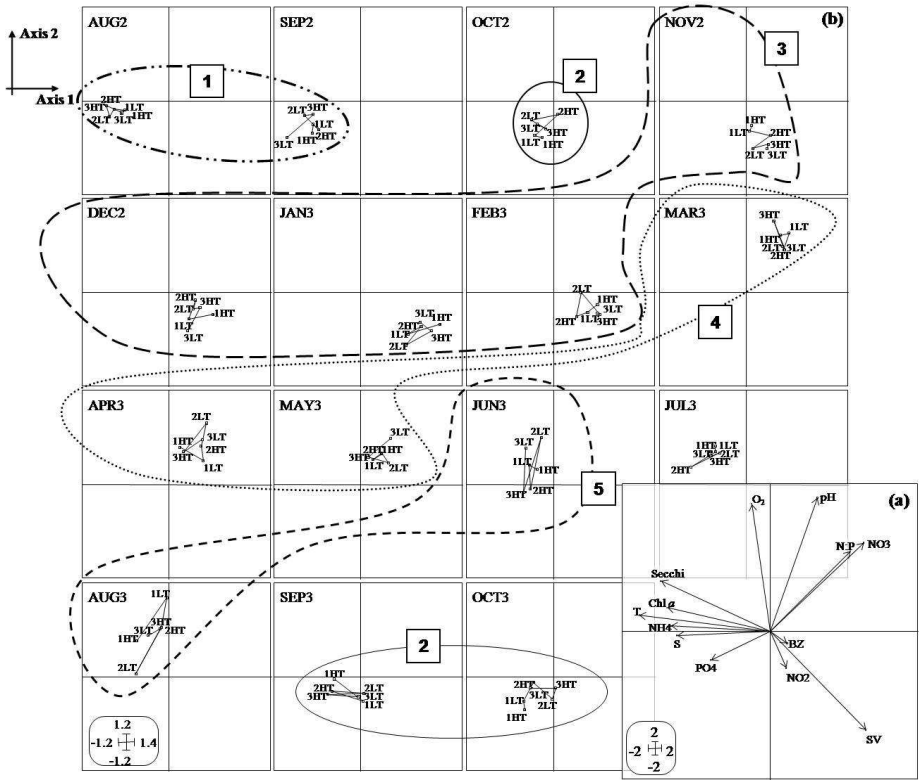


Figure 4. Compromise and trajectories factor maps of the PTA analysis: (a) Compromise factor map of the PTA analysis: environmental variables. This map shows the stable part of the environment relationships on plan 1-2. The scale of the graph is given by the box in bottom left. T – water temperature; O2 – dissolved oxygen; S – salinity; NP – N:P ratio; NH4 – ammonia; Chl a – chlorophyll *a*; PO4 – phosphate; NO2 – nitrite; NO3 – nitrate; Secchi – water transparency; BZ – zooplankton biomass; SV – volatile solids; (b) Trajectories factor maps of the PTA analysis: row-dates projection of each table-site in plan 1-2 of the compromise. The scale of the graph is given by the box in the bottom left. Station codes as Fig. 3. Each date is identified by the three first letters of the month followed by a number: 2 for the year 2002 and 3 for the year 2003 (e.g. AUG2 – August 2002). Axis 1 is the first principal component; Axis 2 is the second principal component. Note the different scales.

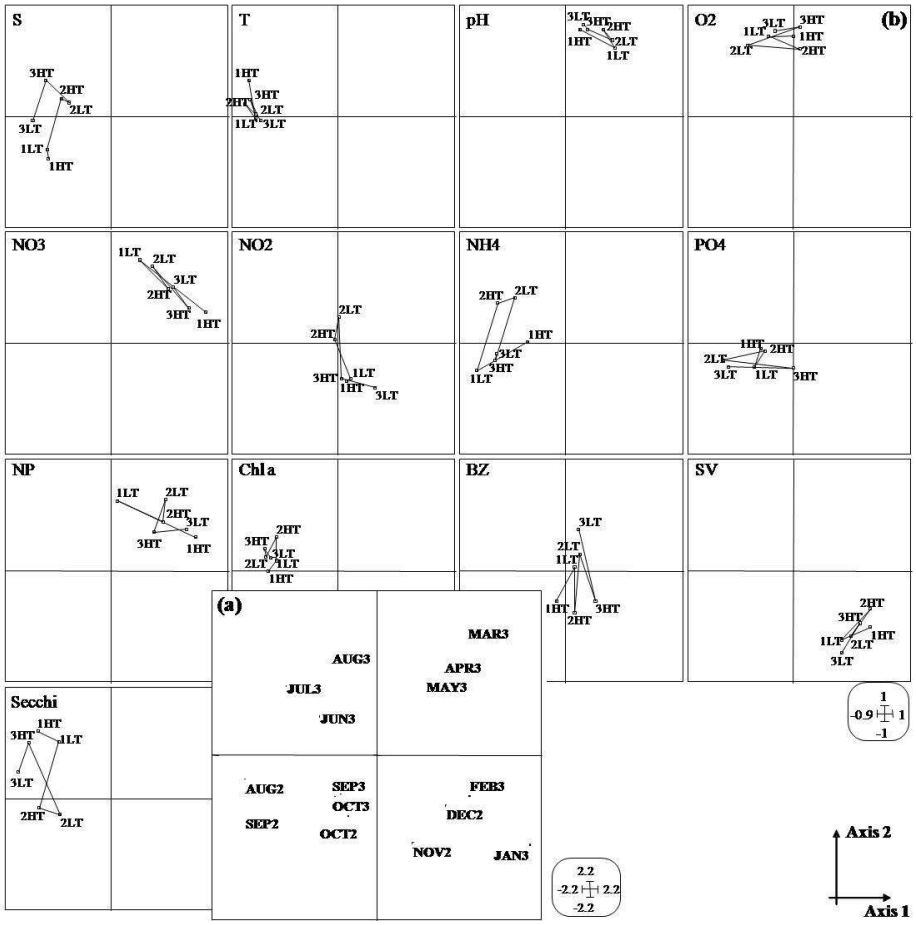


Figure 5. Compromise and trajectories factor maps of the PTA analysis: (a) Compromise factor map of the PTA analysis: sampling dates. This map shows the stable part of the date relationships. The scale of the graph is given by the box in the bottom right. Month codes as Fig. 4; (b) Trajectories factor map of the PTA analysis: variables-environmental factors projection of each table-site in plan 1-2 of the compromise. The scale of the graph is given by the box in the bottom right. Station codes as Fig. 3; Environmental codes as Fig. 4. Axis 1 is the first principal component; Axis 2 is the second principal component. Note the different scales.

dissolved oxygen, transparency, N:P ratio, chlorophyll *a*, ammonia and salinity. In the fourth quadrant, volatile solids appeared isolated. Although nitrite and biomass appear in the same direction, the arrows that represent them have a small length. This means a poor representation on the plan 1-2 and probably does not reflect the proper position. The first axis of this analysis is mainly characterised by temperature, transparency, chlorophyll *a*, ammonia and salinity. Therefore this axis is marked for the left side with higher values of those variables (normally a summer characteristic, see Fig. 5a). Additionally the angles among them were small, which denoted that the variables were strongly correlated. Therefore the opposite side of this axis (on the right side) is where the months that are normally characterised by lower values of temperature, transparency, chlorophyll *a*, ammonia and salinity (usually a winter characteristic, see Fig. 5a) were located. Axis 2 appeared mainly characterised by dissolved oxygen and pH and, in decreasing measure, by nitrates, N:P ratio and volatile solids. In fact, these three variables are variables from the factor plan, although marking different positions. Besides this, the pH and dissolved oxygen have higher values in the months located in the superior part of axis 2 (spring/summer season, see Fig. 5a). The opposite situation occurs at the inferior extreme. Overall the first axis mainly separates spring and winter months from summer and autumn months, while the second axis separates spring and summer months winter and autumn ones (Fig. 5a).

The stable part of dinoflagellates and sampling dates

For dinoflagellates the compromise was performed on the six tables for each site. This gave an average picture of the dinoflagellates abundances (see Table 1 for species codes) which best explained the variations of the species pattern at the fifteen dates for each site (Fig. 6a). The first two axes represented, respectively, 80% and 8% of the total variability. A large abundance of *Prorocentrum micans*, *Ceratium fusus var fusus*, *Dinophysis acuminate* and *Ceratium furca* was observed. These were strongly correlated with the positive part of axis 1 (species from axis 1) and therefore characterized the temporal

Table 1. List of the 17 dinoflagellates species, with codes and category [adapted from Resende et al. (2007)]

Code	*Taxa*
ACIP	*Actiniscus pentasterias*
CEFS	*Ceratium fusus var. fusus*
CEFU	*Ceratium furca*
CEHO	*Ceratium horridum*
CEKO	*Ceratium kofoidii*
DIAC	*Dinophysis acuminata*
DIAT	*Dinophysis acuta*
DICA	*Dinophysis caudata*
GONS	*Gonyaulax spinifera*
PHRO	*Phalacroma rotundatum*
PRDB	*Protoperidinium diabolus*
PRDE	*Protoperidinium depressum*
PRDI	*Protoperidinium divergens*
PRMI	*Prorocentrum micans*
PRPE	*Protoperidinium pellucidum*
PRPT	*Protoperidinium pentagonum*
PYHO	*Pyrophacus horologium*

organizational pattern of the described dinoflagellates. Along axis 2, *Ceratium kofoidii* was opposed to *C. furca. Dinophysis acuta* was strongly correlated with the positive part of axis 2 (species from axis 2). The remainder were considered less abundant (within these, species which are near the origin are slightly abundant).

The analysis of the compromise for sampling dates (Fig. 7a) showed the temporal dynamics of the dinoflagellates assemblage for a mean station and the stable part of the dates' relationships. The two major gradients were determined by the interpretation of the first two axes of the compromise. Axis 2 mainly opposed the dates of September 2002 to August and July 2003. October 2002 and 2003 are strongly correlated with axis 2, while March 2003 has an opposed position and is an axis 1 date. The remaining dates were considered less significant (within these, months which are near the origin, like January and February 2003, are weakly important).

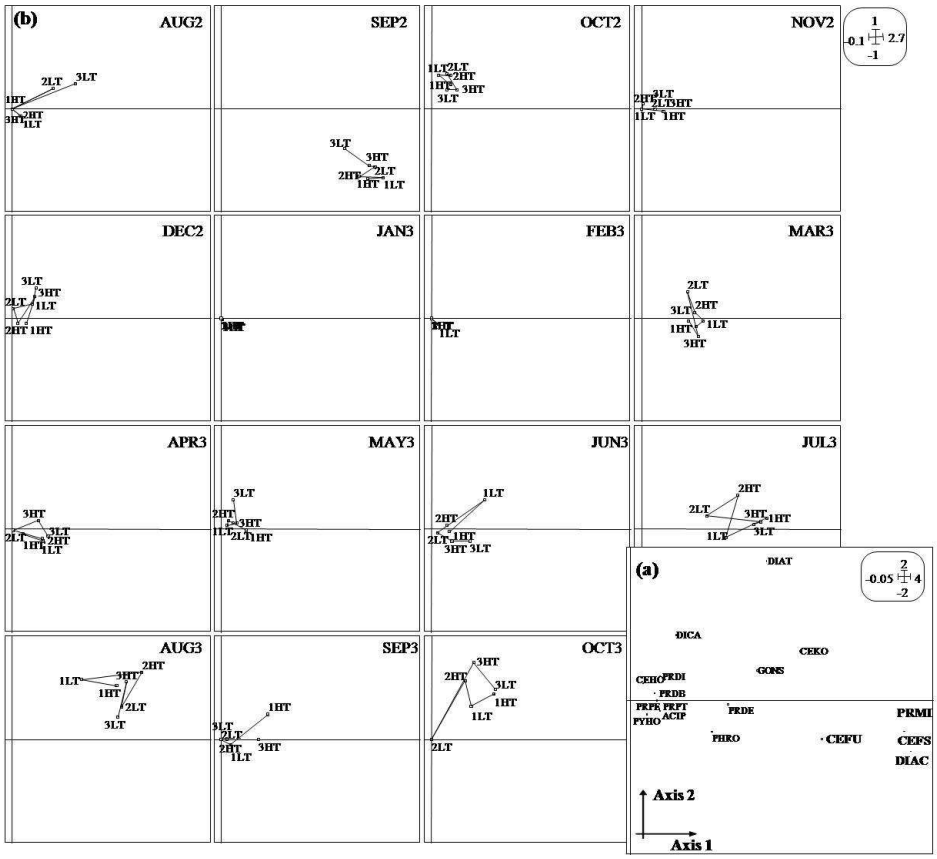


Figure 6. Compromise and trajectories factor maps of the PTA analysis: (a) Compromise factor map of the PTA analysis: the dinoflagellates. The scale of the graph is given by the box in the top right. See Table 1 for species codes. (b) Trajectories factor maps of the PTA analysis: row-dates projection of each table-site in plan 1-2 of the compromise. The scale of the graph is given by the box in the top right. Station codes as Fig. 3. Month codes as Fig. 4. Axis 1 is the first principal component; Axis 2 is the second principal component. Note the different scales.

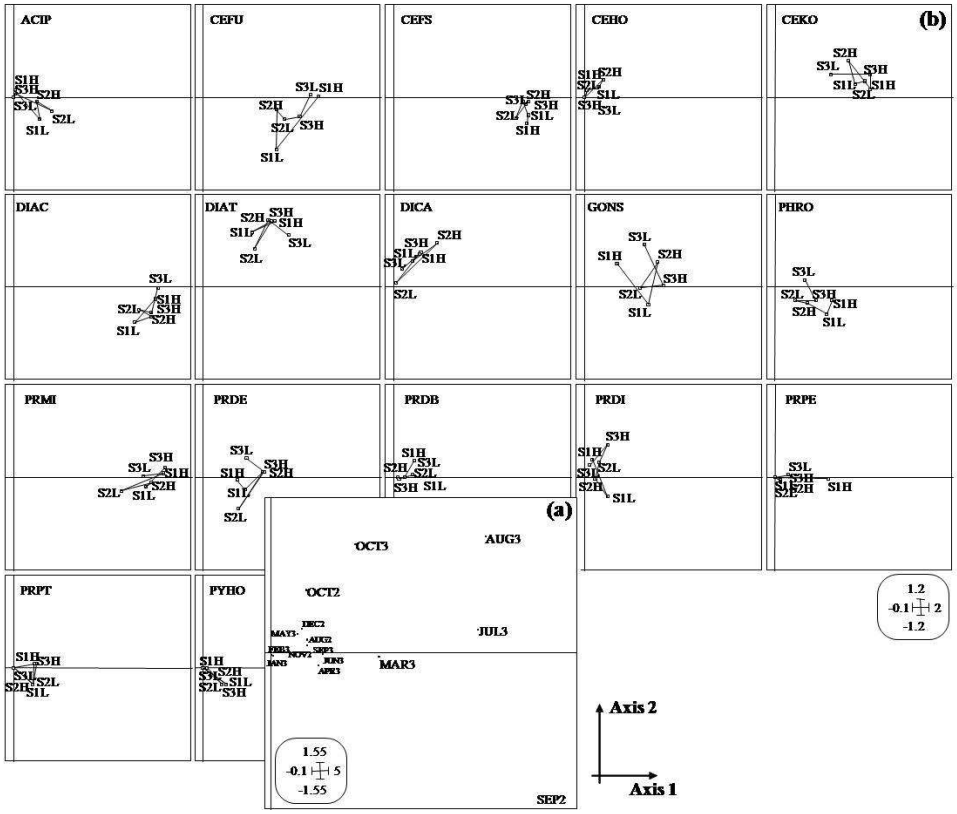


Figure 7. Compromise and trajectories factor maps of the PTA analysis: (a) Compromise factor maps of the PTA analysis: sampling dates. The scale of the graph is given by the box in the bottom left. Month codes as Fig. 4. (b) Trajectories factor map of the PTA analysis: variables-species projection of each table-site in plan 1-2 of the compromise. The scale of the graph is given by the box in the bottom right. Station codes as Fig. 3. See Table 1 for species codes. Axis 1 is the first principal component; Axis 2 is the second principal component. Note the different scales.

5. Trajectories

Time and space effect

For environmental factors, the projection of each date at the six sampling stations on plan 1–2 of the compromise (Fig. 4b) allowed us to visualize the trajectory of each sampling date. Each site (at low and high tides) is represented by a point. The graphical illustration shows the positions of the six stations on the compromise plan connected in the form of trajectories for each date. The trajectories factor maps revealed the possible distortion of the temporal structure of the sampling sites during the study period, that is to say, the diverse shapes of trajectories indicate deviations of single dates from the general pattern. On the whole, although stations followed a chronological scenario, changes of pattern with time can be seen. The main evidence was observed in August and June 2003. Five other distributional patterns of dates can be observed reflecting changes in the system. These patterns fall into distinct date groups: (1) August and September 2002, determined by the information given by the environmental variables transparency, chlorophyll *a*, water temperature, ammonia and salinity (Fig. 4a), which means that these sampling sites within those dates had higher values of these parameters in contrast to the other ones; (2) October 2002, September and October 2003, presented a clearly intermediate pattern between spring/summer and autumn; (3) November and December 2002, January and February 2003 were characterised by volatile solids (Fig. 4a) and evidenced, for the sampling sites, the passage to the autumn/winter season; (4) March, April and May 2003 represents the spring season, with higher values of pH, N:P ratio and nitrates (Fig. 4a), for the majority of the sampling sites; (5) June and August 2003 presented the most different shapes of trajectories when compared to the other ones. Here the trajectories of the stations are dispersed, meaning that their dynamics varied from station to station. At these two dates, variations at stations are determined by the information that was obtained mainly from higher values of pH and dissolved oxygen (Fig. 4a). However, the two dates do

not present the same variations. In June 2003 axis 1 separated stations 2 and 3 at high tide from the others, while in August 2003 all stations were in the upper part of axis 2, and station 2 at low tide had lower values of pH and dissolved oxygen (Fig. 4a) when compared to the other stations. July 2003 appeared as an isolated date pattern, which is characterised by higher values of dissolved oxygen (Fig. 4a) and 2HT was mainly determined by higher observations of transparency, chlorophyll *a*, temperature, ammonia and salinity (Fig. 4a).

It is also possible to project the six sampling sites at the fifteen dates onto the factor map of the compromise and relate it to the dinoflagellates community. Figure 6b allowed us to study this, i.e. the positions of the six sampling sites on the compromise plan connected in the form of "trajectories" for each study date. The dynamics between the sites (at low and high tide) varied according to the sampling dates, and this shows the dynamics between those.

September 2002, October 2002 and August 2003 have a constant occurrence over the selected sites as their trajectories are concentrated on one part of the compromise map, while trajectories of the other dates are dispersed, meaning that their dynamics varied from site to site. In March 2003, axis 1 contrasted station 2 at high and low tide with the other stations. This resulted in a general increase of *P. micans*, *C. fusus var fusus*, *D. acuminate* and *C. furca* (Fig. 6a). March, June and July 2003 presented the most variant shapes, which means, in this scenario, more fluctuations across the information given by axis 2. From October 2002 to May 2003 (with the exception of March 2003), the dynamics were similar, corresponding to winter/spring dynamics, while summer dates (August, September 2002 and June to September 2003) have an opposite pattern. Some of the station points are positioned at the origin of the factor map (0, 0); probably in none of these stations was there any observation for the sampling date or the number of observations was too low (France and Mozetic 2006). This is the case with January, February 2003, which corresponds to a winter period.

Environment and space effect

Figure 5b shows the projections of the thirteen environmental factors at the six stations (represented by points) onto the factor map of the PCA of the compromise. This presents the stability of the dynamics of the environmental parameters, where the positions of each parameter through the study sites are connected in trajectories as well. Overall, obvious changes of variation patterns at any date can be seen through the sites, with the exception of pH, chlorophyll *a* and volatile solids. These are the factors that presented constant values over the selected sites, as their trajectories are mainly concentrated on one part of the compromise map. This stable distribution pattern means that these parameters do not have large fluctuations in the dates across the stations. Salinity, temperature, ammonia, chlorophyll *a*, transparency and phosphate are mainly determined by the information given by the spring and autumn seasons (Fig. 5a). Dissolved oxygen had higher concentrations in the summer/spring season (Fig. 5a), for all study sites. For nitrite concentration, the separation between station 2 (both tides) and the others is clear. Stations 3 and 1 (at the two tides) were mainly characterised by the conditions of winter/autumn season (Fig. 5a), while station 2 (at the two tides) are mainly determined by the spring/summer season (Fig. 5a). Another clear separation was observed between the three stations at low tide and at high tide for the zooplankton biomass: the low tide is determined by spring/summer, while high tide is defined by the winter season (Fig. 5a).

The species and space effect

Figure 7b shows the variables-species projection of each table-site in plan 1-2 of the compromise. Generally, the more irregular shapes were for *C. furca*, *Gonyaulax spinifera*, *Phalacroma rotundatum*, *P. micans*, *Protoperidinium pellucidum* and *Protoperidinium divergens*. *C. fusus var fusus*, *C. kofoidii*, *D. acuminata*, *D. acuta* and *Dinophysis caudata* have a constant occurrence over the selected sites as their trajectories are concentrated on one part of the compromise map, while trajectories of the other species are dispersed, meaning

that their dynamics varied from site to site. For *P. rotundatum*, axis 1 contrasted station 3 at low tide with the other stations (Fig. 7b). This resulted in a general abundance increase in March, July August 2003 and September 2002 (a spring/summer period) (Fig. 7a). In a same way and for *G. spinifera*, axis 1 contrasted stations 1 and 2 (at low tide) with the other stations. An increase of *P. micans* abundances was observed in the summer period (July 2003, August 2003 and September 2002), mostly at stations 1 and 3 (at high tide) (Fig. 7a, b). *C. fusus var fusus* and *D. acuminata* were quite stable at the six sampling stations and for all summer periods (July and August 2003 and September 2002). Some of the station points are positioned at the origin of the factor map (0, 0); probably at none of these stations was any observation from the species detected in water samples or the number of findings was too low (France and Mozetic 2006). This is the case with *Ceratium horridum*, *Protoperidinium depressum*, *P. pellucidum*, *Protoperidinium pentagonum* and *Pyrophacus horologium.*

6. Discussion

This example shows that the PTA method can be used as a useful tool to analyse three-way table species and/or environmental data, interpreting the results in a direct way. This benefit results from the fact that it works with original data instead of operators. Besides this, along with the results interpretation it is possible to summarize the global structure and the relationships between the tables (by means of the interstructure analysis), provide a picture of the structures common to all the tables (by the compromise analysis), and summarize the variability of the series of tables around the common structure defined by the compromise (with the trajectories). The only constraints from PTA are that all the cross-tables must have the same rows and the same columns (i.e. that the species and environmental variables must be the same in all the pairs of tables).

In this work, analysis of the results from two PTA analyses allowed us to group the sites together in terms of their environment attributes or species and their trajectories–histories. Our results showed that: (1) sites may have similar temporal features and exhibit different trajectories; (2) sites may have similar trajectories but different temporal features; and (3) sites may be congruent for both temporal features and trajectories. Nevertheless, a general pattern of species abundance and environmental factors that persists through space (derived from different sampling stations) was recognized by applying PTA analysis and was demonstrated by the results given from interstructure and compromise analysis, which are not far from what really happens for all stations.

The interstructure analysis (similarity between stations) of the environment factors and dinoflagellates community reveals a relatively strong common sites structure (which indicates a large similarity among stations) and two distinct groups: low tide and high tide. Indeed, the contribution of the six sampling sites was well-balanced, indicating that no site was either favoured or ignored in the constitution of the average matrix. Only the second station at low tide was positioned separately from the main group of sites, indicating a different spatial dynamic, which means that their structure was not as well reflected by the compromise as the other five sampling stations.

The projection of the thirteen environmental variables onto the compromise axes when compared to the projection on the same axes of the fifteen sampling dates provided a good summary of the temporal pattern of the environment variables shared by the six sampling sites. In fact, the compromise matrix provided a good approximation of the time organizational pattern of the environment over the six sampling sites. In addition, the projection of the fifteen sampling dates onto the compromise axes was compared to the projection on the same axes of the thirteen environmental variables, in order to summarize the environmental pattern in the sampling months shared by the six sampling sites. Furthermore, the compromise analysis of dinoflagellates (the stable part of species) revealed an independent pattern between the most

abundant dinoflagellates. Those, when well represented (species *P. micans, C. fusus var fusus, D. acuminate, C. furca, C. kofoidii, D. acuta, G. spinifera, P. depressum, P. rotundatum* and *D. caudata*), are mainly associated with warm months (summer and early autumn) and the distribution is forced by volatile solids, nitrates and N:P ratio (which appears with a plan position and were negatively related with temperature, transparency, ammonia, salinity, dissolved oxygen and pH). The projection of the species onto the compromise axes and its association with the projection on the same axes of the fifteen sampling dates summarized the temporal pattern of the dinoflagellates abundances shared by the six sampling sites. In reality, this work has shown that the spatial organization patterns of dinoflagellates assemblage in the Âncora coastal zone were persistent during the course of the considered seasons. The variations among sites of each species around the reference structure were generally low. Only a very limited number of species exhibited a strong variation of their abundance in space at this scale. The reported temporal distributions are similar to what is known from published papers (Resende et al. 2007). There, the relationship between dinoflagellates assemblage and the environmental parameters governing their composition and structure was performed with the Canonical Correspondence Analysis (CCA) (ter Braak, Verdonschot 1995). With this technique CCA extracts synthetic gradients from the biotic and environmental matrices, which are quantitatively represented by arrows in graphical biplots. The results obtained from the PTA method have been a successful approach in evaluating the prevailing inter-space and intra-space structure of the species, which with the CCA technique cannot be observed (the same applies to the environment). This becomes a significant problem when we are working with phytoplankton ecological studies. In this case, results may be insignificant and do not represent the reality of this assemblage. On the other hand, spatial stability has been weakly explained to date. The inclusion of PTA in ecology may be important for revealing that behind an overall stability of environmental gradients over the study period, the spatial changes are important and thus may have an impact on local communities. Besides this, it allows

researchers to take into account more biological information, to use methods more adapted to the data, and to produce more accurate statistical results.

Our study analysed both dinoflagellates composition and structure (species abundances and environmental factors), which was likely to capture changes that may not be detected by several traditional methods alone. At this stage we will refer to the importance of dinoflagellates changes at both levels of biodiversity.

Acknowledgements

The statistical analyses were run using the ADE-4 package (Thioulouse et al. 1997). The authors are very grateful to the contributors who have made such a valuable tool available. Sincere thanks to the Department of Statistics (University of Salamanca) for their input and constructive discussions. We also thank Maria José Sá for editing the English text.

REFERENCES

Azeiteiro U., Bacelar-Nicolau L., Resende P., Gonçalves F., Pereira M.J. (2006): Larval fish distribution in shallow coastal waters off North Western Iberia (NE Atlantic). Estuarine, Coastal and Shelf Science 69: 554–566.

Blanc L., Chessel D., Dolédec S. (1998): Etude de la stabilité temporelle des structures spatiales par analyses d'une série de tableaux faunistiques totalement appariés. Bulletin Français de la Pêche et de la Pisciculture 348: 1–21.

Ernoult A., Freiré-Diaz S., Langlois E., Alard D. (2006): Are similar landscapes the result of similar histories? Landscape Ecology 21: 631–639.

France J., Mozetic P. (2006): Ecological characterization of toxic phytoplankton species (Dinophysis spp., Dinophyceae) in Slovenian mariculture areas (Gulf of Trieste, Adriatic Sea) and the implications for monitoring. Marine Pollution Bulletin 52: 1504–1516.

Gaertner J.C. (2000): Seasonal organization patterns of demersal assemblages in the Gulf of Lions (north-western Mediterranean Sea). Journal of the Marine Biological Association of United Kingdom 80: 777–783.

Jaffrenou P.A. (1978): Sur l'analyse des familles finies de variables vectorielles. Bases algébriques et application à la description statistique. Thèse de 3° cycle, Sciences et Technique du Languedoc, Montpellier-II, Montpellier.

Lavit C., Escoufier Y., Sabatier R., Traissac P. (1994): The ACT (STATIS) method. Computational Statistics & Data Analysis 18: 97–119.

Resende P., Azeiteiro U., Gonçalves F., Pereira M.J. (2007): Distribution and ecological preferences of diatoms and dinoflagellates in the west Iberian Coastal zone (North Portugal). Acta Oecologica 32: 224–235.

Rossi J.P. (2006): The spatiotemporal pattern of a tropical earthworm species assemblage and its relation with soil structure. Pedobiologia 47: 497–503.

Simier M., Blanc L., Pellegrin F., Nandris D. (1999): Approche simultanée de K couples de tableaux: application à l'étude des relations pathologie végétale – environnement. Revue de statisque appliquée 47: 31–46.

ter Braak C.J.F. (1986): Canonical Correspondence Analysis: A New Eigenvector Technique for Multivariate Direct Gradient Analysis. Ecology 67: 1167–1179.

ter Braak C.J.F., Verdonschot P.F.M. (1995): Canonical correspondence analysis and related multivariate methods in aquatic ecology. Aquatic Sciences 57: 255–289.

Thioulouse J., Chessel D. (1987): Les analyses multitableaux en écologie factorielle. I. De la typologie d'état à la typologie de fonctionnement par l'analyse triadique. Acta Oecologica, Oecologia Generalis 8: 463–480.

Thioulouse J., Chessel D., Dolédec S., Olivier J.M. (1997): ADE-4: a multivariate analysis and graphical display software. Statistics and Computing **7**: 75–83.

Thioulouse J. Simier M., Chessel D. (2004): Simultaneous analysis of a sequence of paired ecological tables. Ecology 85: 272–283.

PAPER III

Zooplankton distribution in a Marine Protected Area: the Berlengas Natural Reserve (Western coast of Portugal)

Mendes S, Marques SC, Azeiteiro UM, Fernández-Goméz MJ, Galindo-Villardón MP, Maranhão P, Morgado F, Leandro, SM

Fresenius Environmental Bulletin 20: 496-505, 2011 (#)

(#) Additional information:

- Publ. Freising: Parlar Scientific Publication
- ISSN: 1018-4619
- **Indexed in:** Journal Citation Report - Science Citation Index
- **Impact Factors:** IF 2010 (2 years): 0.716; IF 2010 (5 years): 0.704
- **Category:** ENVIRONMENTAL SCIENCES. Ranking: 164/193 – Q. 4 (Quartile 4)
- **Also indexed in**: Biology & Environmental Sciences, BIOSIS, C.A.B. International, Cambridge Scientific Abstracts, Chemical Abstracts, Current Awareness, Current Contents/ Agriculture, CSA Civil Engineering Abstracts, CSA Mechanical & Transportation Engineering, IBIDS database, Information Ventures, NISC, Research Alert, Science Citation Index (SCI), SciSearch, Selected Water Resources Abstracts.

ZOOPLANKTON DISTRIBUTION IN A MARINE PROTECTED AREA: THE BERLENGAS NATURAL RESERVE (WESTERN COAST OF PORTUGAL)

Susana Mendes[1,4,*], Sónia Cotrim Marques[2], Ulisses Miranda Azeiteiro[2,3], Mª José Fernández-Gómez[4], Mª Purificación Galindo-Villardón[4], Paulo Maranhão[1,2], Fernando Morgado[5] and Sérgio Miguel Leandro[1]

[1]School of Tourism and Maritime Technology, Marine Resources Research Group, Polytechnic Institute of Leiria, 2520-641 Peniche, Portugal
[2]University of Coimbra, Centre for Functional Ecology, Department of Life Sciences, 3001-455 Coimbra, Portugal
[3]University Aberta, Department of Science and Technology, 4200-055 Porto, Portugal
[4]University of Salamanca, Department of Statistics, 37007 Salamanca, Spain
[5]University of Aveiro, Centre for Environmental and Marine Studies, Department of Biology, 3810-193 Aveiro, Portugal

ABSTRACT

Zooplankton distribution in the Berlengas Natural Reserve (Portugal) was studied over a period of one year (February 2006 to February 2007). Monthly sampling was performed at 6 stations, differentiated according to depth and distance to the coastline. The aim of this study was to investigate the overall zooplankton variability through its different dimensions (space vs. time). The Partial Triadic Analysis (PTA) was used to study the spatial variability of the zooplankton community and its dynamics in time and the dynamic trajectories of the zooplankton species for each site. It was possible to distinguish a neritic-ocean gradient of the zooplankton composition and a temporal variability. Four distinct periods can be highlighted considering the distribution of the dates and the arrangement of the species: (i) the first one comprised August to November, (ii) the second one was related to June and July, (iii) the third one associated with spring (April and May) and, (iv) the latest one was related to winter (February, March and December 2006 and January and February 2007). The PTA method showed the similarities between the successive data tables and proved to be useful for investigating biotic structures and detecting spatial-temporal patterns in zooplankton distribution.

KEYWORDS: Berlengas Natural Reserve, partial triadic analysis, spatio-temporal distribution, zooplankton.

1. INTRODUCTION

The Berlengas Natural Reserve (BNR) is an archipelago formed by 3 groups of islands (Berlenga, Estelas and Farilhões) located on the western coast of Portugal. The Reserve was created in 1981, aiming to preserve a rich natural heritage and to ensure sustainable development of human activities in the area. More recently, BNR was proposed to be a Biosphere Reserve. This denomination is attributed by UNESCO to sites where is recognized the existence of innovate approaches to conservation and sustainable development. It has a total area of 9560 hectares, 9456 of which are marine. It is located on the Portuguese continental shelf at an average distance of 5.7 miles from the mainland (Cape Carvoeiro - Peniche). The geographical location gives singular characteristics to the archipelago, which enhanced the interest of ecological studies, because it is located in a zone with a temperate maritime climate and is influenced by seasonal coastal upwelling controlled by the atmospheric circulation associated with the Azores anticyclone. Persistent northerlies (upwelling favourable) are observed in summer (June to September) [1,2]; it is, however, during the non-upwelling season (late winter-spring) that a large amount of meroplankton species are observed over the shelf [3]. Concerning coastal circulation, other important aspects are the Portugal Current flowing off the continental slope westward of 10ºW [4], the Iberian Poleward Current that flows over the slope [5] and the Western Iberia Buoyant Plume (WIBP) [1]. Moreover, it is located at the top of the escarpment of the Nazaré Canyon, one of the most worldwide important submarine canyons in the transition zone between the Mediterranean and European subregions. This location contributes to the remarkable productivity and diversity of marine species and habitats and to a landscape unique in the region. Previous studies have investigated the distribution and composition of zooplankton along the Berlenga shelf area [6]; however information on the zooplankton community remains limited. This work intended to be a preliminary study in this area and pretended to analyze the variations in zooplankton species abundance and the different compositions at shelf and oceanic sites through its different dimensions (time

* Corresponding author

vs space). To achieve such objective the spatio-temporal zooplankton community structure was assessed by means of a Partial Triadic Analysis (PTA) [7-9]. The PTA is a method of analysis for three-way data sets presented as a sequence of two-way tables. It is useful when analyzing the same variables (in this study, the species density) measured on the same items (dates or sites) and done for the same occurrences (sites or dates). Its general aim is to determine the proportion of variability in the variables that depends on space or on time.

2. MATERIALS AND METHODS

2.1. Survey design

From February 2006 to February 2007, zooplankton samples were collected monthly from 6 stations: E1 (39°25´N 9°30´W), E2 (39°25´N 9°31'W), E3 (39°25'N 9°31´W), E4 (39°25´N 9°30´W), E5 (39°25´N 9°27´W) and E6 (39°21´N 9°23´W), located along a transect perpendicular to the coastline (between the Peniche coast and Berlenga islands) (Fig. 1). September 2006 was not sampled due to poor ocean conditions. The samples were obtained from inshore (E6), shelf (E5) and offshore regions (E1, E2, E3 and E4). Zooplankton samples were collected through horizontal hauls and were performed during day time from 1 m below the surface using a 500 µm mesh net and samples were preserved in 4% formalin in seawater for further analysis. A Hydro-Bios digital model 438 110 flow-meter was fitted to the net in order to measure the volume of water filtered. For zooplankton taxonomic and quantification analyses sub-sampled using a Folsom-splitter were used until a minimum of 500 individuals were counted. Abundance data were standardized to number of individuals per cubic meter.

2.2. Data Analysis

Only the most abundant 50 taxa out of the 90 identified, having a minimal mean occurrence of 0.1% of the total density observed in the study area were considered. This cut-off eliminated the species that occurred rarely, some being observed on few or rare occasions. Moreover, well-represented species can be viewed as potential indicators of zooplankton dynamics and ecosystem functioning.

In order to investigate the temporal and spatial variability in the zooplankton community structure, the species densities were arranged in a tridimensional matrix (species, dates and sites) comprising 50 columns and 72 rows. These data offered the possibility to study the three-dimensional array in two ways (Fig. 2): (A) the spatial variability of the zooplankton community and its dynamics in time (data were organized as a series of tables for each date, where each column corresponded to the species density and each row corresponded to a sample) and (B) the dynamic trajectories of the zooplankton community per site (data were considered as a series of tables for each site, where each column corresponded to the species density and each row corresponded to a sample). In order to down weight the influence of highly dominant species, species density was log(x+1) transformed prior to calculations [10]. Data were subjected to the PTA [7-9], which is based on the logic of the Principal Component Analysis (PCA) [11]. It is designed to study simultaneously several sub-matrices of quantitative data and to detect within the structure any pattern common to these different sub-matrices; in other words, it allows extraction of the multivariate structure that is expressed through the different dates or sites, and describes dominant patterns in its first axes. It is a multivariate technique well-adapted to the statistical study of surveys when the same variables (in this study, species density) are measured on the same individuals [7]. The general

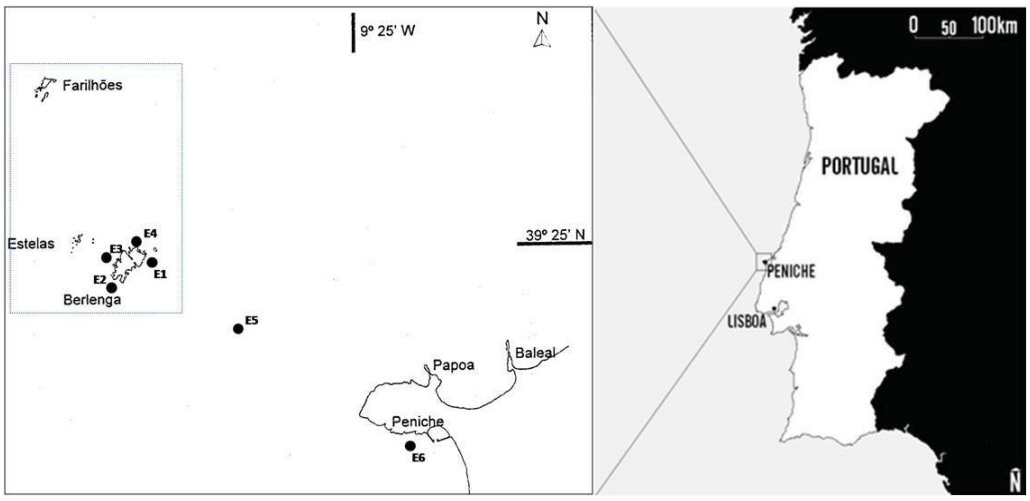


FIGURE 1 - Map of the Berlengas Archipelago and the location of the 6 sampling sites. The rectangle represents the marine protected area.

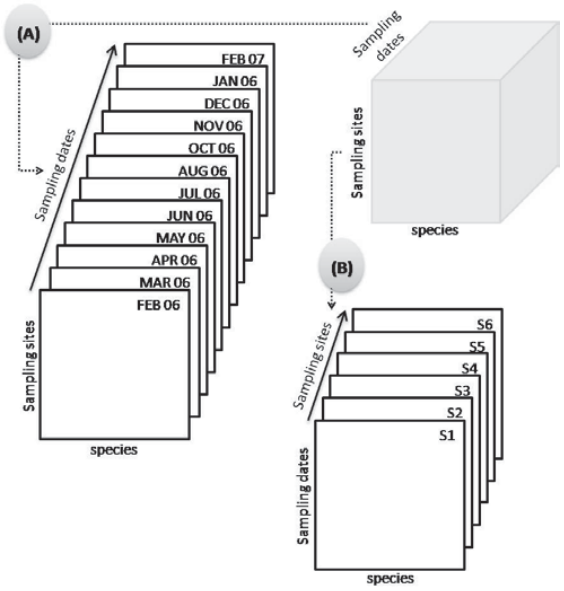


FIGURE 2 - Tridimensional table of data (dates, sites, zooplankton density) used in the PTA analysis. Way (A) – Identification of a spatial structure common to the 12 dates and study of the temporal permanence. Way (B) – Identification of a temporal structure common to the 6 sites and study of the spatial permanence.

functioning principle of PTA consists to find the common part of all separated analysis (for dates or sites). For that purpose, PTA analysis consists of successive steps: (1) the interstructure analysis provided a global description of the sampling points as a function of the typology of the sampling individuals (dates or sites). It consists of the comparison of the structure of the different sub-matrices (dates or sites) and the identification of the individuals sharing a similar structure (by the vectorial correlations presented in the matrix and calculated between dates or sites). The function of this step is to assign a weight to each sub-array and measure the contribution of each to the overall structure. Additionally, and in proportion to the weights is the $\cos^2(x)$. It constitutes an indicator of how much the overall structure expresses the information contained in each table; (2) the compromise analysis provided a description of sampling points as a function of the species typology. It was used to identify the species assemblages that characterized similar patches at different dates or sites. This leads to the establishment of a common spatial or temporal typology shared by those dates or sites, respectively; (3) finally, the analysis of the trajectories (or intrastructure) onto the compromise. The subsets are projected separately onto the compromise to highlight which date or site fits best to the compromise; in other words, allows us to draw the trajectories that represent the temporal or spatial variations of each species around the common structure. All the analyses (calculations and graphs) were run using the package ADE-4 [12]. This software is available free of charge at the following Internet address: http://pbil.univ-lyon1.fr/ADE-4.

3. RESULTS

A total of 90 taxa represented by 30 groups were encountered. Data showed that cladocerans were the most abundant group (30% of the total zooplankton), despite being restricted to warmer months (Fig. 3). Other important zooplankton groups were the copepods, a perennial group (21%), the gelatinous (13%, constituted primarily by 7% doliolids, 4% appendicularians and 2% medusae), meroplankton (18%) and siphonophores (9%). Minor groups also found in the region were euphausiaceans (3%), chaetognaths (1%) and mysidaceans (0.1%). The average abundance for the main zooplankton taxa for each sampling station is presented in Table 1. Among the cladocerans the *Penilia avirostris* were by far the most abundant species in the study area (77% Cladocera), with a limited seasonal occurrence, mostly present in summer and autumn. *Podon leuckarti* (16%) and *Evadne nordmanni* (7%) followed in abundance. More than 40 copepod species were identified during the entire study, but only 6 species accounted for

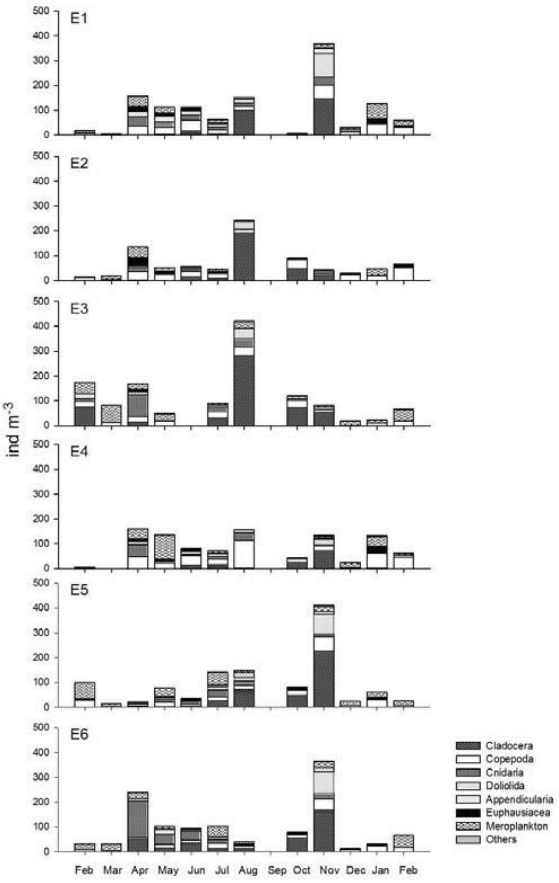


FIGURE 3 - Monthly abundance (ind m^{-3}) of the main zooplankton groups, over different sampling stations, from February 2006 until February 2007.

TABLE 1 - Annual mean abundance (ind m^{-3}) of the main zooplanktonic taxa and their standard deviation (SD) during the studied period. Species abbreviation (abbrev) used in PTA analysis.

		E1	E2	E3	E4	E5	E6
Taxa	abbrev						
Total abundance		1218 ±101	845 ± 64	1298 ± 113	1004 ± 58	1155 ± 109	1206 ± 103
Cladocera		292 ± 48	299 ± 55	541 ± 82	141 ± 21	394 ± 66	363 ± 49
Evadne nordmanni	ENOR	24 ± 3	9 ± 1	19 ± 2	22 ± 2	18 ± 2	53 ± 9
Penilia avirostris	PAVI	241 ± 47	258 ± 55	443 ± 81	93 ± 20	317 ± 65	216 ± 48
Podon leuckarti	PLEU	27 ± 4	31 ± 5	78 ± 12	25 ± 4	60 ± 9	94 ± 15
Copepoda		276 ± 17	245 ± 13	189 ± 10	374 ± 31	195 ± 15	168 ± 11
Acartia clausi	ACLA	15 ± 2	13 ± 1	10 ± 1	15 ± 3	12 ± 1	53 ± 5
Calanus helgolandicus	CHEL	20 ± 1	17 ± 2	9 ± 2	19 ± 2	45 ± 7	12 ± 1
Calanus helgolandicus copepodite	CHELC	40 ± 6	44 ± 6	17 ± 3	36 ± 7	11 ± 2	4 ± 1
Centropages chierchae	CCHI	94 ± 10	59 ± 5	87 ± 8	83 ± 10	64 ± 7	48 ± 5
Oithona plumifera	OITP	17 ± 3	9 ± 2	7 ± 1	87 ± 22	7 ± 1	4 ± 1
Temora stylifera	TSTY	45 ± 7	48 ± 5	27 ± 3	57 ± 9	33 ± 5	36 ± 7
Cnidaria		146 ± 14	41 ± 5	166 ± 26	107 ± 15	94 ± 9	269 ± 42
Lizzia blondina	LBLO	23 ± 4	3 ± 0	14 ± 2	22 ± 4	19 ± 3	20 ± 3
Siphonophora							
Muggiaea atlantica	MATL	110 ± 11	36 ± 5	145 ± 24	71 ± 10	72 ± 7	238 ± 38
Doliolida		139 ± 27	47 ± 9	52 ± 11	54 ± 8	108 ± 23	89 ± 25
Doliolum sp.	DOLI	139 ± 27	47 ± 9	52 ± 11	54 ± 8	108 ± 23	89 ± 25
Appendicularia		72 ± 9	18 ± 1	44 ± 6	30 ± 3	57 ± 6	68 ± 7
Fritillaria borealis	FBOR	35 ± 6	3 ± 0	5 ± 0	5 ± 1	5 ± 1	8 ± 2
Oikopleura sp.	OIKO	37 ± 5	15 ± 1	39 ± 6	25 ± 3	52 ± 5	60 ± 6
Euphausiacea		69 ± 7	60 ± 10	18 ± 3	49 ± 7	23 ± 2	10 ± 1
calyptopsis	ECAL	36 ± 4	30 ± 5	10 ± 1	31 ± 5	17 ± 1	7 ± 1
furcilia	EFUR	33 ± 4	30 ± 5	10 ± 1	31 ± 5	17 ± 1	7 ± 1
Meroplankton		202 ± 18	127 ± 12	274 ± 21	225 ± 29	261 ± 19	229 ± 15
Cirripedia nauplius	NCIRR	5 ± 1	3 ± 0	12 ± 2	7 ± 1	38 ± 10	31 ± 4
Decapoda larvae n id	DLAR	12 ± 1	9 ± 1	15 ± 2	7 ± 1	12 ± 1	28 ± 3
Zoea *Carcinus maenas*	ZCAR	87 ± 12	46 ± 6	183 ± 19	53 ± 8	141 ± 18	108 ± 11
Zoea *Pisidia longicornis*	ZPLO	2 ± 0	3 ± 0	9 ± 1	3 ± 1	7 ± 1	14 ± 1
Echinodermata larvae	ELARV	11 ± 1	3 ± 0	20 ± 4	77 ± 21	17 ± 3	29 ± 7
Ichthyoplankton		7 ± 10	5 ± 11	2 ± 3	6 ± 11	3 ± 6	11 ± 1
Fish egg	FEGG	76 ± 10	51 ± 11	25 ± 3	70 ± 11	38 ± 6	11 ± 1
Others		21 ± 2	9 ± 1	14 ± 1	24 ± 3	22 ± 2	11 ± 1

83% of total copepods. The calanoid *Centropages chierchae* was the most abundant, followed by *Temora stylifera*, *Calanus helgolandicus* (adult and copepodite development stage), *Oithona plumifera* and *Acartia clausi* (Table 1). Meroplankton such as nauplii Cirripedia, Decapoda larvae (particularly Zoea *Carcinus maenas*), Echinodermata pluteus and fish egg were also important species. The siphonophores were mainly represented by *Muggiaea atlantica* and appendicularians by *Oikopleura* spp. and *Fritillaria borealis*.

3.1. Dynamics of zooplankton community spatial variability at dates scale

The first procedure of the PTA was conducted on the 12 tables for each date, denoting the spatial variations (i.e rows = sites and columns = species density). The use of PTA showed that the vectorial correlations (RV in Table 2A) between each of the 12 months had different contributions. Moreover, these results pointed out that the strongest correlation was observed between April and May (RV = 0.73), while June and March reflected the weaker (RV = 0.16). This result highlighted the fact that the associations between species were not stable from one month to another. From the analysis of the correlations and the weights (Table 2A), the common temporal structure appeared to be stronger in November, August, April, July and May. They contributed a larger part in the definition of the compromise (meaning that the compromise will be more influenced by these dates) suggesting that the remaining sampling dates had more particular structures leading to a weaker weight. Finally, by the $\cos^2(x)$ (Table 2A) was possible to evaluate how much the compromise expresses the information contained in each table. July ($\cos^2(x) = 0.68$) was the month that fits best with the compromise, followed by November ($\cos^2(x) = 0.63$). By other way, the compromise represented with less accuracy the zooplankton dynamics in December and March ($\cos^2(x) = 0.19$ and $\cos^2(x) = 0.17$, respectively).

The projection of the sampling sites on the first principal plan 1-2 provides a graphical representation of the compromise, whose interpretation requires consideration of the correspondences with the species (Fig. 4A). The eigenvalues diagram (Fig. 4A1) shows that the first axis was clearly dominant (account for 94% of the explained variance) in contrast with the second axis (4% of the explained variance) which was less significant. Therefore, they provided a good summary and typology of the spatial species organization, on the basis of the common structure, over the 12 sampling dates. The factor plots of the first

TABLE 2 - Typological value indices. Matrix of correlation between surveys (RV) and description of the structure defined for each survey. (A) Temporal scale and (B) spatial scale. Weights (contribution of each table in the construction of the compromise); $Cos^2(x)$ (fit of each table to the compromise).

A

Sampling date	RV												Weight	$Cos^2(x)$
Feb 06	1												0.19	0.31
Mar 06	0.54	1											0.09	0.17
Apr 06	0.37	0.29	1										0.37	0.51
May 06	0.34	0.38	0.73	1									0.32	0.53
Jun 06	0.19	0.16	0.51	0.61	1								0.27	0.53
Jul 06	0.42	0.29	0.62	0.64	0.69	1							0.36	0.68
Aug 06	0.42	0.22	0.39	0.39	0.44	0.59	1						0.38	0.53
Oct 06	0.44	0.23	0.28	0.28	0.34	0.43	0.63	1					0.23	0.45
Nov 06	0.35	0.17	0.33	0.39	0.54	0.55	0.67	0.68	1				0.44	0.62
Dec 06	0.44	0.41	0.23	0.30	0.22	0.26	0.22	0.24	0.34	1			0.10	0.19
Jan 07	0.46	0.42	0.47	0.56	0.50	0.49	0.36	0.42	0.47	0.53	1		0.27	0.48
Feb 07	0.46	0.55	0.38	0.40	0.34	0.40	0.30	0.31	0.36	0.50	0.55	1	0.20	0.32

B

Sampling site	RV						Weight	$Cos^2(x)$
E1	1						0.47	0.68
E2	0.75	1					0.36	0.59
E3	0.64	0.70	1				0.39	0.55
E4	0.76	0.66	0.53	1			0.39	0.54
E5	0.78	0.67	0.67	0.64	1		0.43	0.65
E6	0.69	0.59	0.59	0.59	0.75	1	0.40	0.55

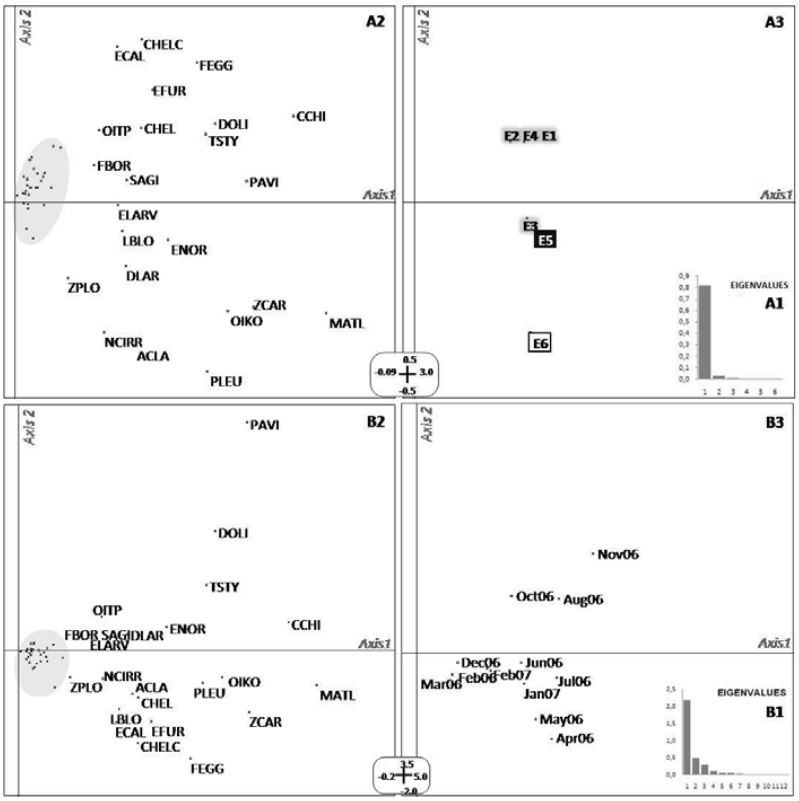


FIGURE 4 - Compromise factor maps of the PTA analysis. (A) Dynamics of zooplankton spatial variability at the temporal scale: (A1) Eigenvalues bar plot of the compromise; (A2) Zooplankton density projected on the first factorial plan; (A3) Sampling sites projected on the same factorial plan. (B) Variability of the zooplankton dynamics at the spatial scale: (B1) Eigenvalues bar plot of the compromise; (B2) Zooplankton density projected on the first factorial plan; (B3) Sampling dates projected on the same factorial plan. Only the labels of most abundant species and thus of importance in these graphics, were identified for clarity. Species codes according Table 1. Each date is identified by the three first letters of the month (e.g. Feb07 – February 2007). Axis 1 - the first principal component; Axis 2 - the second principal component. The scales of the graphs are given in the rounded box. Note for different scales.

two axes of the compromise analysis are shown for the zooplankton community (Fig. 4A2) and sites (Fig. 4A3). In order to facilitate the interpretation, only the most abundant species were identified in the graph (see Table 1 for code correspondence). The analysis of the first principal plan (Fig. 4A) distinguished three zones characterized by its particular species composition: one related to the inshore station E6, the second zone comprises stations with different characteristics, a shelf station (E5) and an offshore station (E3), and finally one zone that group the further offshore stations (E1, E2, E4). Axis 2 mainly opposed homogeneous stations (E1, E2, and E4; Fig. 4A3) dominated mainly by *C. chierchae*, *P. avirostri, T. stylifera* and *Doliolum* spp. to stations less homogeneous (E3, E5 and E6). The second zone defines a transition area characterized by the occurrence of echinodermata and decapoda larvae, the cladoceran *E. nordmanni* and medusae *L. blondina*. Finally, the siphonophore *M. atlantica*, the appendicularians *Oikopleura* sp., and meroplanktonic larvae such as Zoea of *C. maenas* and nauplii of cirripedia, and the cladoceran *P. leuckarti* were clearly under the influence of the characteristics of the neritic station E6. The trajectories of the compromise for the sampling dates were represented for species (Fig. 5A1) and for sites (Fig. 5A2). It allows us to identify temporal patterns for the species and sites around the common structure. The trajectories maps focused on November, August, April, July and May, when the observed co-structure of species densities was the most significant (see Table 2A). The projection of species variables for each month on the compromise plan showed that each one is close to the compromise structure (see Fig. 4A), which confirms that the analysis is in accordance with the pattern pointed out by the compromise.

The projection of each table-date in plan 1-2 of the compromise (Fig. 5A1) showed that *P. avirostris* exhibited the strongest temporal variation. The density found for this species was particularly lower in April, May, June and July, in contrast with August and November. It was the clear seasonal dynamics of *P. avirostris*, in conjunction with some Copepoda species (e.g. *O. plumifera, C. chierchae, T. stylifera*), that strongly affected the structure of zooplankton community and caused the observed pattern between E1, E3 and E5 (Fig. 5A2) in these two months. Moreover, *Doliolum* sp. and *P. avirostris* appeared as the dominant species during this period, following a similar pattern, however the former species showed an earlier increase in abundance (June and July). *M. atlantica* was the most stable during the study period, with its position always in the negative half of axis 2. Also, *P. leuckarti* follow this pattern with an exception for June, where it appeared closer to *Doliolum* sp. and *C. chierchae*. Regarding the period in analysis, the inshore station E6 was always located in the negative half of axis 2 (Fig. 5A2). Furthermore, while in May and November, offshore sites were in opposition to E6, in July those sites had the particularity of being nearest the inshore station, indicating a similar pattern in species composition. In addition, the dynamics between E4 and E2, from July to August, resulted in a general increase of *P. avirostris* and a decrease of Echinodermata larvae and *A. clausi* (Fig. 5A1 and 5A2). To summarize, based on the analysis of the two representations of the trajectories, the conditions of the sites showed a more pronounced stability over time compared with the annual pattern of the species distribution.

3.2. Variability of zooplankton community dynamics at the site scale

The second procedure of the PTA was performed on the 6 tables for each site, denoting the temporal variations (rows = dates and columns = species density).

The matrix presenting the RV between the sampling sites sub-matrices (Table 2B) showed that the strongest correlation (RV = 0.78) was observed between the sites E1 and E5 whereas the sites E3 and E4 pointed out the weakest one (RV = 0.53). Also, it was observed that the contribution of the different sampling sites for the construction of the compromise, were well-balanced (weighting 0.36-0.47). However, the sub-matrices E1, E5 and E6 contributed a major part in the definition of the compromise (0.47, 0.43, 0.40, respectively) suggesting that other sites had more particular structures (leading to a lower weight). The observation of the $\cos^2(x)$ (Table 2B) indicates that E1 was the one that fits the best with the compromise ($\cos^2(x) = 0.68$), followed by stations E5 ($\cos^2(x) = 0.65$) and E2 ($\cos^2(x) = 0.59$). On the other hand, for the stations E3, E6 and E4, the compromise represented with less precision the annual dynamics of zooplankton ($\cos^2(x) = 0.55$, 0.55 and 0.54, respectively).

The projection of the species on the plan 1–2 (Fig. 4B2) provides a graphical representation of the compromise, whose interpretation requires consideration of the correspondences with the months (Fig. 4B3). The first two axes of the compromise accounted for 82% of the total inertia with 67% for axis 1 and 15% for axis 2 (Fig. 4B1). They provided a good summary of the temporal species organization over the 6 sampling stations for the 12 sampling months. As previously, only the most abundant species were identified in the graph (see Table 1 for code correspondence).

The first axis distinguished a group of species that were typical of warmer months such as *C. chierchae* and *M. atlantica* and to a lesser extent Zoea *C. maenas*, *Oikopleura* sp., *T. stylifera* and *P. leuckarti*. These were all observed to be relatively more common between June to November (Fig. 4B3).

The second axis was defined by two distinct species assemblages. One assemblage was mainly dominated by *P. avirostris* and *Doliolum* sp. whereas the second one was essentially composed by fish eggs and *C. helgolandicus* (copepodites and adults). In addition, the distribution of the sampling dates on the compromise reflected a transition from April-June to August-November (Fig. 4B3).

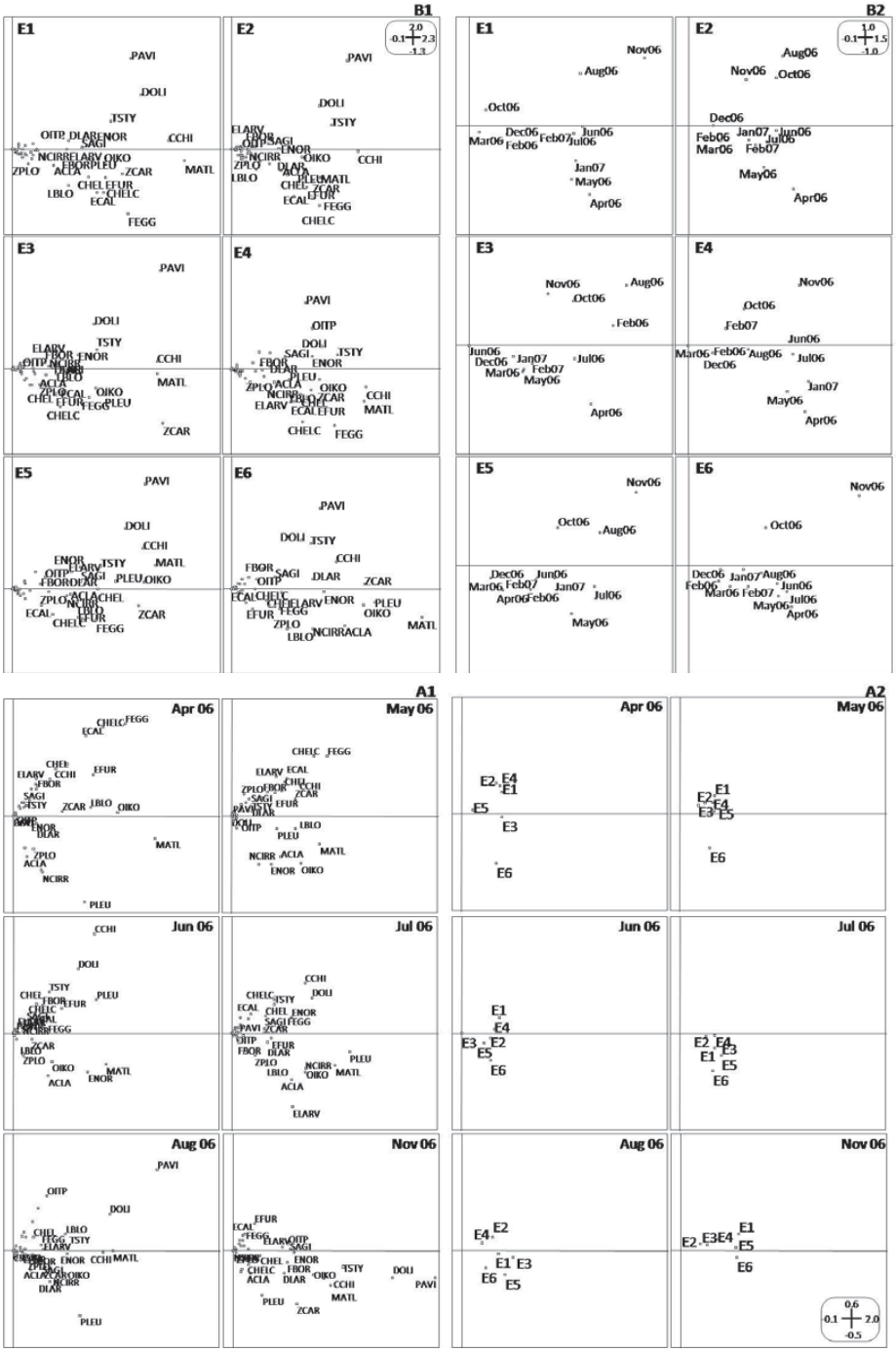


FIGURE 5 - Trajectories factor plots of the PTA analysis. (A) Dynamics of zooplankton spatial variability at the temporal scale: (A1) Projection of the zooplankton species on the first factorial plan (only for species that stood out on the compromise diagram, see Fig. 4A2). (A2) Projection of the sampling stations on the first factorial plan. Graphs are given only for the 6 dates that showed the highest contribution to the compromise (see Table 2A). (B) Variability of the zooplankton dynamics at the spatial scale: (B1) Projection of the zooplankton species on the first factorial plan (only for species that stood out on the compromise diagram, see Fig. 4B2). (B2) Projection of the sampling dates on the first factorial plan. Species codes according Table 1. Legend of the months according Fig. 4. Axis 1- the first principal component; Axis 2 - the second principal component. The scales for axes are given in the rounded box.

From the separated analyses of each table, the compromise of species (Fig. 5B1) and dates (Fig. 5B2) has been projected onto the first principal plan. This allowed the discussion of the spatial variations of species and dates, i.e. the internal structure of each table per site. Generally, these projections were in agreement with the patterns noted by the compromise. However, the analysis can be mainly focused on sampling stations E1 and E5, since its general pattern of species and dates distribution, is what is more aligned with the compromise (Fig. 5B). This observation implies that there were a few differences between seasonal changes in the composition of the zooplankton community for these two sites and the compromise. Moreover, this confirms the information given by the values of the weights and the $\cos^2(x)$ (see Table 2B). Although slight, the general pattern observed for the stations E2, E3, E4 and E6, for the distribution of dates and species (see Fig. 5B1 and Fig. 5B2) differed from the distribution achieved by the compromise. Even though not very strong, they had their own internal typology (each sites presented a particular species composition and abundance dynamic) which indicates that the zooplankton community was characterized by its own seasonal dynamics for those sampling stations.

4. DISCUSSION AND CONCLUSIONS

The analysis of zooplankton community in the upper waters of the shelf area of Berlengas Natural Reserve (NW Portugal) suggested high species diversity, with 90 species/genera recorded. Data from zooplankton samples showed that cladocerans were the most abundant group (30% of the total zooplankton), but restricted to warm periods. An important perennial group was represented by copepods (21%). It is worth stressing the importance of these taxa in offshore zooplankton studies, as has been already observed off the NW Iberian Peninsula [13-15] and in other coastal areas [16, 17]. In the region, these groups have been also highlighted as the most abundant by a previous study performed by Pardal and Azeiteiro [6].

The PTA method showed the similarities between the successive data tables (arranged according to different scales, time vs. space) and proved to be useful for investigating biotic structures and detecting different patterns in the temporal development and spatial distribution of zooplankton communities. Therefore, regarding the zooplankton abundance pattern and the species association it was possible to distinguish a pattern related to a neritic-ocean gradient of the zooplankton composition and a temporal variability. In addition, following the analysis of the principal factor plan, four distinct periods can be highlighted considering the distribution of the dates and the arrangement of the species on the first two axes: (i) the first one comprised August to November, (ii) the second one was related to June and July, (iii) the third one associated with spring (April and May) and, (iv) the latest one was related to winter (February, March and December 2006 and January and February 2007).

The geographic location of the Berlengas Archipelago and the specific hydrographic conditions in the area were clearly important in structuring the zooplankton community. As observed in other coastal areas, associated with upwelling events, shelf and oceanic sites have different zooplankton composition in response to different hydrographic conditions [17,18]. The hydrodynamics off the Portugal coast has been described by several authors (e.g. [1, 19]). As described, the general surface circulation follows a seasonal pattern. Wind regime seasonality can be very important in the shelf currents in the region [20]. Dominant south-westerly winds in autumn-winter induce downwelling and favour poleward flow (IPC), while dominant northerly winds in spring-summer are associated with coastal upwelling and equatorward flow [5]. The transition from winter downwelling to summer upwelling conditions occurs around April [21]. This transition is reflected in the large variability in circulation and in the distribution of physical properties, observed both on the shelf and offshore. These seasonal changes in the currents flowing along the Portuguese coast and the different upwelling patterns originated by differences in the shelf topography were reflected in the changes in the horizontal distribution of zoo-plankton. According to PTA analysis was possible to define 3 spatial areas based on species composition. What emerges from our results is that the inshore region was the more unstable, owing to the influence of upwelling events. Indeed different species-assemblages were observed during the study period at that station. Taxa linked to high phytoplankton concentrations such as *C. helgolandicus* and *Calanoides carinatus* [22] dominated during warmer months, which coincide with upwelled waters, rich nutrient waters that favored phytoplankton development/spring bloom [23]. Furthermore, it is noteworthy that during the warmer months (between August to October) a considerable increase in the abundance of cladocerans was observed, mainly due to *P. avirostris*. This species is one of the more abundant and widespread members of the crustacean zooplankton in nearshore tropical and subtropical waters [24], although recently it has spread to higher latitudes (e.g. the North Sea, [25]). It seems that its occurrence is mainly related to the availability of adequate food. *P. avirostris* ingests a wide spectrum of microbial organisms, from flagellates <2 mm to chain-forming diatoms, showing a clear advantage over copepods [26]. In addition, it is known that among the different phytoplankton groups observed in coastal upwelling ecosystems, diatoms and dinoflagellates can take advantage of different oceanographic conditions [23,27,28], since those organisms are the main food source of *P. avirostris*, that situation could be positive for their positive development. Besides, during the study period, regional and local climate showed a general pattern toward dried and warmer conditions which extend through-outthroughout the autumn period, as indicated by a reduction in precipitation and a warming of air temperatures (http://web.meteo.pt/pt/clima/clima.jsp). These conditions could favour a dominance of flagellates resulting in a change in food availability for *P. avirostris*. Another

feature that can distinguish species from continental shelf and more offshore sites during stratified waters was species linked to lower salinity, such as *A. clausi* and meroplankton, which are quite common inshore.

A good knowledge of the geographical distribution of the species-assemblages and their temporal variability is crucial to facilitate and detect ecological interpretation. However, we recognize that monitoring the dynamics of pelagic ecosystems at long-term time series is needed in order to be aware of important aspects such as global warming trend on marine communities, invasion of exotic species ("biologic pollution"), and to identify the important species or groups that can act as indicator of global changes or environmental contaminants [29]. As denoted by [30] the decrease in subarctic species, associated with an increase in temperate pseudoceanic species, have a possible link with change along the European shelf-edge. For all the mentioned aspects, the Berlengas Natural Reserve (NW Portugal), due to their special hydrodynamics features, can be considered a suitable area to biological long-term studies to detect changes, at local but also at regional scale. More studies should be conducted in near future in order to better understanding the zooplankton dynamics and their relationship with hydrodynamic processes that occur on the west coast of Portugal, and namely on the vicinity of Berlengas Natural Reserve as a consequence of the Nazaré submarine canyon.

ACKNOWLEDGEMENTS

A special thanks to all colleagues that helped during this fieldwork. The statistical analyses were run using the ADE-4 package [12]. The authors are very grateful to the contributors who have made such a valuable tool available.

REFERENCES

[1] Peliz, A., Rosa, T., Santos, A.M.P. and Pissarra, J. (2002) Fronts, jets, eddies and counterflows in the western Iberia upwelling system. Journal of Marine Systems 35, 61-77.

[2] Álvarez-Salgado, X. A., Figueiras, F. G., Pérez, F. F., Groom, S., Nogueira, E., Borges, A. V., Chou, L., Castro, C. G., Moncoiffé, G., Ríos, A. F., Miller, A. E. J., Frankignoulle, M., Savidge, G. and Wollast, R. (2003) The Portugal coastal counter current off NW Spain: new insights on its biogeochemical variability. Progress in Oceanography 56, 281-321.

[3] Santos, A., Peliz, A., Ré, P., Dubert, J., Oliveira, P. and Angélico, M. (2004) Impact of a winter upwelling event on the distribution and transport of sardine eggs and larvae off western Iberia: A retention mechanism. Continental Shelf Reseach 24, 149-165.

[4] Saunders, P. (1982) Circulation in the eastern North Atlantic. Journal of Marine Research 40, 641-657.

[5] Haynes, R. and Barton, E.D. (1991) Lagrangian observations in the Iberian Coastal Transition Zone. Journal of Geophysical Research-Oceans 96, 14731-14741.

[6] Pardal, M. and Azeiteiro, U.M. (2001) Zooplankton biomass, abundance and diversity in a shelf area of Portugal (the Berlenga Marine Natural Reserve). Arquipélago Life and Marine Sciences 18A, 25-33.

[7] Thioulouse, J. and Chessel, D. (1987) Les analyses multitableaux en écologie factorielle. I. De la typologie d'état à la typologie de fonctionnement par l'analyse triadique. Acta Oecologica, Oecologia Generalis 8, 463-480.

[8] Kroonenberg, P.M. (1989) The analysis of multiple tables in factorial ecology. III. Three-mode principal component analysis: ''analyse triadique complète''. Acta Oecologica, Oecologia Generalis 10, 245–256.

[9] Blanc, L., Chessel, D. and Dolédec, S. (1998) Etude de la stabilité temporelle des structures spatiales par analyses d'une série de tableaux faunistiques totalement appariés. Bulletin Français de la Pêche et de la Pisciculture 348, 1-21.

[10] Legendre, L. and Legendre, P. (1979) Ecologie numerique. 1. Le traitement multiple des donnees ecologiques; 13 Cdée, editor. Paris: Masson et Le presses de l'université du Québec. 197 p.

[11] Hotelling, H. (1933) Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology 24, 498-520.

[12] Thioulouse, J., Chessel, D., Dolédec, S. and Olivier, J.M. (1997) ADE-4: a multivariate analysis and graphical display software. Statistics and Computing 7, 75-83.

[13] Villate, F. and Valencia, V. (1997) Mesozooplancton community indicates climate changes in a shelf area of the Bay of Biscay throughout 1988 to 1990. Journal of Plankton Research 19, 1617-1636.

[14] Beaugrand, G., Ibanez, F. and Reid, P.C. (2000) Spatial, seasonal and long-term fluctuations of plankton in relation to hydroclimatic features in the English Channel, Celtic Sea and Bay of Biscay. Marine Ecology Progress Series 200, 93-102.

[15] Lindley, J.A. and Daykin, S. (2005) Variations in the distributions of the *Centropages chierchiae* and *Temora stylifera* (Copepoda: Calanoida) in the north-eastern Atlantic Ocean and western european shelf waters. ICES Journal of Marine Science 62, 860-877.

[16] Fernández de Puelles, M.L, Valencia, J., Jansá, J. and Morillas, A. (2004) Hydrographical characteristics and zooplankton distribution in the Mallorca channel (Western Mediterranean): spring 2001. ICES Journal of Marine Science 61, 654-666.

[17] Ayón, P., Criales-Hernandez, M.I., Schwamborn, R. and Hirche, H.-J. (2008) Zooplankton research off Peru: a review. Progress in Oceanography 79, 238-255.

[18] Blanco-Bercial, L., Alvarez-Marques, F. and Cabal, J.A. (2006) Changes in the mesozooplankton community associated with the hydrography off the northwestern Iberian Peninsula. ICES Journal of Marine Science 63, 799-810.

[19] Relvas, P., Peliz, A., Oliveira, P.B., da Silva, J., Dubert, J., Barton, E.D. and Santos, A.M.P. (2007) Western Iberia upwelling ecosystem: an oceanographic overview. Progress in Oceanography 74, 149-173.

[20] Vitorino, J., Oliveira, A., Jouanneau, J.M. and Drago, T. (2002) Winter dynamics on the northern Portuguese shelf. Part 2: bottom boundary layers and sediment dispersal. Progress in Oceanography 52, 155-170.

[21] Blanton, J.O., Atkinson, L.P., Fernandez-Castillejo, F. and Lavin Montero, A. (1984) Coastal upwelling off the Rias Bajas, Galicia, Northwest Spain I; Hydrographic studies. Rapports et Procés-Verbaux des Réunions Conseil International pour l'Exploration de la Mer 183, 78-90.

[22] Laabir, M., Poulet, S.A., Harris, R.P., Pond, D.W., Cueff, A., Head, R.N. and Ianora, A. (1998) Comparative study of the reproduction of *Calanus helgolandicus* in well-mixed and seasonally stratified coastal waters of the western English Channel. Journal of Plankton Research 20, 407-421.

[23] Pereira, C., Azeiteiro, U.M and Ferreira, M.J. (2010) Diatoms and dinoflagellates of the outer Aveiro Estuary, Portugal: annual variation and ecology. Fresenius Environmental Bulletin 19 (4a), 704-716.

[24] Rose, K., Roff, J. and Hopcroft, R.R. (2004) Production of *Penilia avirostris* in Kingston Harbour, Jamaica. Journal of Plankton Research 26, 605-615.

[25] Johns, D.G., Edwards, M., Greve, W. and SJohn, A.W.G. (2005) Increasing prevalence of the marine cladoceran *Penilia avirostris* (Dana, 1852) in the North Sea. Helgoland Marine Research 59, 214-218.

[26] Katechakis, A., Stibor, H., Sommer, U. and Hansen, T. (2004) Feeding selectivities and food niche separation of *A. clausi, Penilia avirostris* (Crustacea) and *Doliolum denticulatum* (Thaliacea) in Blanes Bay (Catalan Sea, NW Mediterranean). Journal of Plankton Research 26, 589-603.

[27] Resende, P., Azeiteiro, U., Gonçalves, F. and Pereira, M.J. (2007) Distribution and ecological preferences of diatoms and dinoflagellates in the west Iberian Coastal zone (North Portugal). Acta Oecologica 32, 224-235.

[28] Oliveira, P.B., Moita, T., Silva, A., Monteiro, I.T. and Palma, A.S. (2009) Summer diatom and dinoflagellate blooms in Lisbon Bay from 2002 to 2005: pre-conditions inferred from wind and satellite data. Progress in Oceanography 83, 270-277.

[29] Pereira, M.E., Abreu, S., Pato, P., Coelho, J.P, Azeiteiro, U.M., Pardal, M.A., and Duarte, A.C. (2007) Seasonal Hg concentrations in plankton-net material from a contaminated coastal lagoon (Ria de Aveiro, Portugal). Fresenius Environmental Bulletin 16 (11b), 1442-1450.

[30] Beaugrand, G., Reid, P.C., Ibanez, F., Lindley, J.A. and Edwards, M. (2002) Reorganization of North Atlantic marine copepod biodiversity and climate. Science 296, 1692-1694.

Received: September 10, 2010
Accepted: November 04, 2010

CORRESPONDING AUTHOR

Susana Mendes
School of Tourism and Maritime Technology
Marine Resources Research Group
Polytechnic Institute of Leiria
2520-641 Peniche
PORTUGAL

Phone: +351 262 783 607
Fax: +351 262 783 088
E-mail: susana.mendes@ipleiria.pt

FEB/ Vol 20/ No 2a/ 2011 – pages 496 - 505

PAPER IV

Spatio-temporal structure of diatom assemblages in a temperate estuary. A STATICO analysis

Mendes S, Fernández-Gómez MJ, Resende P, Pereira MJ,

Galindo-Villardón MP, Azeiteiro UM

Estuarine, Coastal and Shelf Science 84: 637-664, 2009 (#)

(#) Additional information:

- Publ. London: Academic Press / Elsevier Science
- ISSN: 0272-7714
- **Indexed in:** Journal Citation Report - Science Citation Index
- **Impact Factors:** IF 2009 (2 years): 1.970; IF 2009 (5 years): 2.366
- **Category:**
  MARINE & FRESHWATER BIOLOGY. Ranking: 20/88 – Q. 1 (Quartile 1)
  OCEANOGRAPHY. Ranking: 16/56 – Q. 2 (Quartile 2)
- **Also indexed in**: BIOBASE, BIOSIS databases/Zoological Records, CAB Internacional, Chemical Abstracts Service, Chemical Abstracts Service, Current Awareness in Biological Sciences, Current Awareness in Biological Sciences, Current Contents ASCA/Engineering Technology & Applied Science/Science Citation Index/SCISEARCH Data, Current Contents/Agriculture, Biology & Environmental Sciences, Current Contents/Physics, Chemical, & Earth Sciences, Engineering Index, Environmental Periodicals Bibliography, Geo Bib & Index, INSPEC Data/Cam Sci Abstr, Marine Literature Review, Meteorological and Geoastrophysical Abstracts, Oceanbase, Oceanographic Literature Review, Research Alert. Scisearch, Scopus.

Estuarine, Coastal and Shelf Science 84 (2009) 637–644

Contents lists available at ScienceDirect

Estuarine, Coastal and Shelf Science

journal homepage: www.elsevier.com/locate/ecss

Spatio-temporal structure of diatom assemblages in a temperate estuary. A STATICO analysis

Susana Mendes [a,*], Mª José Fernández-Gómez [b], Paula Resende [c], Mário Jorge Pereira [d], Mª Purificación Galindo-Villardón [b], Ulisses Miranda Azeiteiro [e,c]

[a] GIRM-Research Group on Marine Resources, Polytechnic Institute of Leiria, School of Tourism and Maritime Technology – Campus 4, 2520-641 Peniche, Portugal
[b] University of Salamanca, Department of Statistics, 37007 Salamanca, Spain
[c] IMAR-Institute of Marine Research, University of Coimbra, Department of Zoology, 3004-517 Coimbra, Portugal
[d] University of Aveiro, Department of Biology, 3810-193 Aveiro, Portugal
[e] University Aberta, Department of Sciences and Technology, 4200-055 Porto, Portugal

ARTICLE INFO

Article history:
Received 2 May 2009
Accepted 7 August 2009
Available online 12 August 2009

Keywords:
environmental forcing
Diatoms
Estuaries
multitable analysis
STATICO

ABSTRACT

This study examines the spatio-temporal structure of diatom assemblages in a temperate estuary (Ria de Aveiro, Western Portugal). Eighteen monthly surveys were conducted, from January 2002 to June 2003, at three sampling sites (at both high and low tide) along the estuarine salinity gradient. The relationship of diatom assemblages and environmental variables was analysed using the STATICO method, which has been designed for the simultaneous analysis of paired ecological tables. This method allowed examination of the stable part of the environment-diatom relationship, and also the variations of this relationship through time. The interstructure factor map showed that the relationship between the 11 environmental variables and the abundance of the 231 diatom species considered was strongest in the months May and September 2002 and January, February and May 2003. The stable part of the species–environment relationships mainly consisted of a combined phosphate, chlorophyll *a* and salinity gradient linked to a freshwater-marine species gradient. A more pronounced gradient was observed in January, February and May 2003. Diatom assemblages showed clear longitudinal patterns due to the presence of both marine and freshwater components. May and September 2002 had the least structured gradients with marine-estuarine species appearing in the freshwater side of the gradient. The most complete gradient in February 2003 could be considered, in terms of bio-ecological categories, as the most structured period of the year, with a combination of strong marine influence in the lower zone and freshwater influence in the upper. The best-structured gradients were during periods of a diatom bloom. Stable diatom assemblages (with a strong structure and a good fit between the diatoms and environment) are described and characterized. This study shows the efficiency of the STATICO analysis. The inclusion of space-time data analysis tools in ecological studies may therefore improve the knowledge of the dynamics of species–environmental assemblages.

1. Introduction

Various studies on phytoplankton communities in estuaries have concluded that diatoms are the most important taxonomic groups, either in terms of abundance or in terms of diversity or both (Trigueros and Orive, 2001; Lemaire et al., 2002; Adolf et al., 2006; Gameiro et al., 2007). Diatoms can survive in systems with a high turbidity and a short water retention time (Lionard et al., 2008). These communities are composed of dynamic multi-species assemblages characterized by high diversity and rapid successional shifts in species composition in response to environmental changes. Identifying the ecological variables that regulate the seasonal succession of diatom communities is essential to understand the consequences of eutrophication and climate change. Beyond that, phytoplankton composition and abundance are intimately linked to higher trophic levels through grazing by herbivores and cascading effects on ecosystem trophodynamics (Mallin and Paerl, 1994; Pinckney et al., 1998; Urrutxurtu et al., 2003).

* Corresponding author. GIRM – Grupo de Investigação em Recursos Marinhos, Instituto Politécnico de Leiria, Escola Superior de Turismo e Tecnologia do Mar, *Campus* 4, Santuário Nossa Senhora dos Remédios, 2520-641 Peniche, Portugal.
E-mail addresses: smendes@estm.ipleiria.pt (S. Mendes), mjfg@usal.es (M.J. Fernández-Gómez), presende@zoo.uc.pt (P. Resende), mverde@ua.pt (M. Jorge Pereira), pgalindo@usal.es (M.P. Galindo-Villardón), ulisses@univ-ab.pt (U.M. Azeiteiro).

0272-7714/$ – see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ecss.2009.08.003

Fig. 1. Map of Canal de Mira – Ria de Aveiro and the study area with location of the 3 sampling sites (Resende et al., 2005).

In many temperate estuaries and coastal areas, seasonal patterns of phytoplankton community-composition are characterized by a spring diatom bloom (Pinckney et al., 1998; Domingues et al., 2005; Domingues and Galvão, 2007) or a late autumn/winter-spring diatom bloom (Adolf et al., 2006; Lopes et al., 2007). Diatom abundances usually decreases in the summer, due to Nitrogen (N) and/or Silicates (Si) limitation (Kocum et al., 2002; Domingues et al., 2005; Domingues and Galvão, 2007), and pelagic and benthic grazing (Vaquer et al., 1996; Domingues et al., 2005; Domingues and Galvão, 2007).

Resende et al. (2005) described for Ria de Aveiro (Western Portuguese Atlantic Coast) a diatom composition that resembled other European temperate estuaries but found no seasonal pattern of diatom density. Salinity and temperature were described as the environmental drivers of diatoms' distribution and composition patterns. Canonical Correspondence Analysis (CCA, ter Braak, 1986) was used by Resende et al. (2005) to identify the environmental variables governing the composition and structure of diatom assemblages (together with the study of ecological preferences). The CCA is a global analysis that does not consider the *a priori* information on time or space, information being presented *a posteriori*, when the results are displayed. Therefore a mixture of space-time effects is produced, showing evidence of the strongest effects compared to the weaker ones. Multitable analyses focus on the identification of spatial or temporal structures and the permanence of these structures in time or space. They thus represent a good alternative to CCA and to other methods that do not take into account the temporal variations of environmental factors and biological communities and the effect of the former on the latter (Carassou and Ponton, 2007).

In this study the main spatial structure of diatom assemblages and its temporal changes, in terms of bio-ecological categories and of their relation to some environmental variables, were studied using the STATICO method (Simier et al., 1999; Thioulouse et al., 2004) since it performs a simultaneous analysis of a sequence of paired ecological tables. The method has been used to obtain a clear representation of diatom-environment relationship, its evolution in time and to characterize the typology of the studied stations according to this temporal evolution. The efficiency of this methodology was demonstrated by Carassou and Ponton (2007) in a study with the spatio-temporal structure of pelagic larval and juvenile fish assemblages in coastal areas of New Caledonia (Southwest Pacific) and by Simier et al. (2006) on fish assemblages from the Gambia River. The present work aims at: (1) describing diatom dynamics in a temperate estuary; (2) determining the applicability of the STATICO method to diatom estuarine dynamics and (3) comparing the advantages of STATICO over the classical CCA technique.

2. Material and methods

2.1. Study area and sampling sites

Ria de Aveiro is located in the Northwest coast of Portugal (40°38' N and 8°44' W). Canal de Mira is an elongated shallow arm that can be considered a small estuary in itself (Resende et al., 2005). Three sites were sampled in Canal de Mira (Fig. 1): S1 – Barra (40°38' N; 08°44' W), S2 – Costa Nova (40°36' N; 08°44' W) and S3 – Vagueira (40°33' N; 08°45' W). Sampling took place monthly, in sub-surface water, always at new moon periods and at both low and high tide, from January 2002 to June 2003. A detailed description of the sampling sites is given by Resende et al. (2005).

Table 1
Ranges of environmental parameters during the study period, in the three sampling places.

	Barra (S1)			Costa Nova (S2)			Vagueira (S3)		
	Min	Max	Average ± SD	Min	Max	Average ± SD	Min	Max	Average ± SD
Sal (g l^{-1})	17.6	36.9	31.9 ± 4.5	1.0	36.7	24.8 ± 11.0	0.0	33.7	15.1 ± 10.8
T (°C)	11.2	19.7	15.7 ± 2.2	9.5	20.7	15.7 ± 2.6	10.2	21.8	16.6 ± 3.3
pH	6.39	8.31	8.0 ± 0.4	6.83	8.30	8.0 ± 0.3	7.50	8.45	8.0 ± 0.2
DO_2 (% sat)	66.0	171.2	97.0 ± 21.2	66.0	115.2	85.7 ± 13.7	50.3	152.3	86.0 ± 22.0
NO_3^- (mg N l^{-1})	0.002	0.388	0.113 ± 0.095	0.002	1.079	0.233 ± 0.282	0.002	1.498	0.303 ± 0.374
NO_2^- (mg N l^{-1})	ND	0.011	0.004 ± 0.003	ND	0.024	0.006 ± 0.005	ND	0.026	0.009 ± 0.008
NH_4^+(mg N l^{-1})	0.003	0.444	0.060 ± 0.088	ND	0.146	0.042 ± 0.031	ND	0.132	0.043 ± 0.035
PO_4^{3-}(mg P l^{-1})	0.001	0.227	0.036 ± 0.051	0.003	0.150	0.038 ± 0.040	0.011	0.168	0.055 ± 0.042
N:P	0.6	212.0	16.5 ± 35.4	0.9	148.9	16.0 ± 26.2	0.5	23.5	6.5 ± 5.1
Chl *a* (μg l^{-1})	0.53	33.38	5.02 ± 7.13	ND	14.27	4.09 ± 3.46	ND	22.96	6.41 ± 5.17

ND – undetermined value. Sal – salinity; T – water temperature; DO_2 – dissolved oxygen; NO_3^- – nitrate; NO_2^- – nitrite; NH_4^+ – ammonium; PO_4^{3-} – phosphate; N:P – N:P ratio; Chl *a* – chlorophyll *a*.

Table 2
List of the codes and habitat affinity of diatoms species that stood out in the compromise factor map (see Fig. 3b): (a) For Brackish species (B) and Freshwater species (F); (b) For Marine species (M). [adapted from Resende et al. (2005)].

Code	*Taxa*	Habitat
(a)		
BAPA	*Bacillaria paxillifer*	B
FAPY	*Fallacia pygmaea*	B
GYFA	*Gyrosigma fasciola*	B
MENU	*Melosira numuloides*	B
NZBR	*Nitzschia brevissima*	B
NZSG	*Nitzschia sigma*	B
PASU	*Paralia sulcata*	B
STSP	*Stauroneis specula*	B
TYAC	*Tryblionella acuminata*	B
TYAP	*Tryblionella apiculata*	B
AUGR	*Aulacoseira granulata*	F
CYME	*Cyclotella meneghiniana*	F
NARA	*Navicula radiosa*	F
NACA	*Navicula capitata*	F
NZCL	*Nitzschia clausii*	F
RHAB	*Rhoicosphenia abbreviata*	F
SUBR	*Surirella brebissonii*	F
SYPU	*Synedra pulchella*	F
STPH	*Stauroneis phoenicenteron*	F
TAFE	*Tabellaria fenestrata*	F
(b)		
ACLO	*Achnanthes longipes*	M
AMCO	*Amphora commutata*	M
ANEX	*Anorthoneis excentrica*	M
ATSE	*Actinoptychus senarius*	M
BIAL	*Biddulphia alternans*	M
CODI	*Cocconeis disculus*	M
COPS	*Cocconeis pseudomarginata*	M
COSC	*Cocconeis scutellum*	M
DIMI	*Dimeregramma minor*	M
DPDY	*Diploneis didyma*	M
GRMA	*Grammatophora marina*	M
GROC	*Grammatophora oceanica*	M
LIGD	*Licmophora grandis*	M
LYAB	*Lyrella abrupta*	M
ODMO	*Odontella mobiliensis*	M
OPPA	*Opephora pacifica*	M
PEMO	*Petroneis monilifera*	M
PLEL	*Pleurosigma elongatum*	M
PLNO	*Pleurosigma normanii*	M
PLST	*Plagiogramma staurophorum*	M
RHAD	*Rhabdonema adriaticum*	M
TEAM	*Terpsinoë ammericana*	M
THWS	*Thalassiosira weissflogii*	M

2.2. Environmental data

In total, 108 samples were collected between January 2002 and June 2003: 36 in Barra (S1), 36 in Costa Nova (S2) and 36 in Vagueira (S3). At each site, pH, salinity, water temperature (°C) and dissolved oxygen (% sat) were measured, *in situ*, with a WTW MultiLine P4 portable meter. Water samples for chemical analyses and chlorophyll *a* quantification were collected and immediately stored in the dark and at low temperature (4 °C), until further processing was possible. At the laboratory, these water samples were filtered through GF/C filters (1.2 μm pore diameter) for quantification of photosynthetic pigments. Filtrates were used for the determination of nutrient contents (Resende et al., 2005). Chlorophyll *a* concentration was determined spectrophotometrically at 665 and 750 nm, before and after acidification (Strickland and Parsons, 1972). Nitrate and nitrite concentrations were determined using sodium salicilate and sulfanilic acid and α–naphtylamine method respectively (Rodier, 1984). Ammonium concentration was determined by the indophenol blue technique following the recommendations and procedures of Hall and Lucas (1981). Phosphate in the form of orthophosphate was determined using the stannous chloride method (APHA, 1992). The N: P ratio and distance to the mouth of the estuary were also considered in following analyses. The ranges of environmental parameters during the study period, in the three sampling places are presented in Table 1. A more detailed description can be found in Resende et al. (2005).

2.3. Biological data

During the eighteen months' study period (from January 2002 to June 2003) samples for taxonomic and quantitative study were collected with a glass bottle (1 l capacity) at the water subsurface and immediately preserved with Lugol 1% (iodine/iodide potassium) (Resende et al., 2005). A total of 231 species were identified (Resende et al., 2005) in this study using the standard floras of Peragallo and Peragallo (1897–1908); Germain (1981); Hustedt (1985); Krammer and Lange-Bertalot (1986, 1988, 1991a,b); Round et al., (1990); Sims (1996); Tomas (1996) and Witkowski et al. (2000).

2.4. Data analysis

Data were organized in two series of tables: one for the 11 environmental variables and the other one for 231 diatoms species abundances. Each pair of tables corresponded to the three sites at two tidal conditions, which means six rows (sampling sites) per table. Species abundance was changed to $\log(x+1)$ prior to calculations (Legendre and Legendre, 1979), to minimize the dominant effect of exceptional catches and environmental data were normalized to homogenize the table.

The common structure between environmental and species abundances tables and the stability of this structure over the sampling period were assessed by STATICO method (Simier et al., 1999; Thioulouse et al., 2004). The STATICO method proceeds in three stages: (1) the first stage consists in analysing each table by a one-table method (normed PCA of the environmental variables and centered PCA of the species data); (2) each pair of tables is linked by the Co-inertia analysis (Dolédec and Chessel, 1994) which provides an average image of the co-structure (species-variables); (3) Partial Triadic Analysis (Thioulouse and Chessel, 1987) is finally used to analyze this sequence. It is a three-step procedure, namely the interstructure, the compromise and the intrastructure analyses. STATICO also enables to plot the projection of the sampling sites of each original table on the compromise axes (of the PCA factor map), in terms of species abundances and environmental factors structures. Hence, it is possible to discuss the correlation between species distribution and environmental factors. Calculations and graphs were done using ADE-4 software (Thioulouse et al., 1997). This software is available free of charge at the following Internet address: http://pbil.univ-lyon1.fr/ADE-4.

3. Results

3.1. Interstructure

The interstructure factor map of the STATICO analysis, based on the 11 environmental variables and on the abundances of the 231 diatom species, showed that the relationship between environmental variables and diatoms appeared to be stronger in September 2002 (with the longest arrow) followed, in decrease order of importance in the compromise, by May 2003, May 2002, January 2003 and February 2003 (meaning that the compromise will be more influenced by these dates) (Fig. 2). The remaining

sampling dates presented short arrows, which means that the corresponding tables are less structured and that their importance in the compromise will be lower. The first two axes represented, respectively, 21% and 10% of the total variability (Fig. 2b).

3.2. Compromise

The first axis was clearly dominant, and accounted for 78% of the explained variance in contrast with the second axis which accounted for 13% of the explained variance and was much less significant (Fig. 3c). Temperature and N: P ratio presented a weak representation on this factorial plan. For the factor map of the environmental variables (Fig. 3a), the first axis describes a salinity gradient, with high values of dissolved oxygen, pH and salinity (positively correlated) on the left side and high values of chlorophyll *a*, phosphate, nitrite, nitrate and low salinity waters (positively correlated) on the right side of the factorial plan. This opposition linked the "marine side" to the "freshwater side" (Fig. 3b), with the majority of marine species on the left quadrants and the majority of freshwater species on the right quadrants of the ordination. Diatom species were therefore ranked according to their salinity affinities. As a result, the stable part of the species–environment relationships mainly consisted of a combined phosphate, chlorophyll *a* and salinity gradient linked to a freshwater-marine species gradient (Table 1 and Fig. 3).

3.3. Trajectories

The trajectories maps focused on May and September 2002 and January, February and May 2003, when the observed co-structure between environment and diatom abundances was the most significant (see Fig. 2). For each sampling month, the projection on the compromise axes of the environmental variables (Fig. 4a) and of the species (Fig. 4b) allowed to visualize the relationships between environmental factors and species abundances and distribution. In general, the trajectories factor maps indicated that the relationships between the most abundant species in the assemblages and environmental factors differed between sampling months (Fig. 4). May 2003 presented a performance best approximated to January and February 2003, contrasting with May and September 2002. In January and February 2003, salinity, pH, dissolved oxygen and temperature appeared positively correlated on the left side of axis 1, along which diatom species appeared to distribute. Salinity, pH, dissolved oxygen and temperature were less correlated during May and September 2002 (at the same time as the distribution of diatom species abundances were most correlated with axis 2). The trajectories factor maps showed a different gradient, the line bisecting the origin on the plan 1–2, with a contrast between brackish, freshwater and marine species (Figs 4a, b).

The co-structure graphics (divided according to sampling dates) clearly showed the dynamics of diatom species–environment relationships and highlighted differences between sites (Fig. 5). Whatever the date, the species points (circles) were more stable than the environmental points. This expresses the steady establishment of the diatom assemblages, in spite of the high environmental variability (salinity in particular). The ends of the arrows (environmental variables) had comparatively different values (generally presented separated) and simultaneously close diatom abundances (the circles were always in the same area). Indeed, from the species point of view, the sites were regularly projected on the right-hand side of the first axis, characterized by the highest diatom abundances (see Fig. 3b).

In general, site one was the one that often presented the shortest arrow, which means that the environmental factors explained well the distribution of species for that site. At this site, located at the mouth of estuary, at low tide, and in particular for May 2002 and September 2002 the arrows were mostly short. This means a higher correlation between the distribution of diatom abundances and the environment, under the direct influence of high values of phosphate, ammonium and chlorophyll *a*. On the other hand, and notwithstanding the strong dispersion of the environmental points and a poor fit between the diatoms and environment (long arrows), sites 1LT (station 1 low tide), 1HT (station 1 high tide) and 2HT (station 2 high tide) were regularly grouped together, as well as sites 2LT (station 2 low tide), 3LT (station 3 low tide) and 3HT

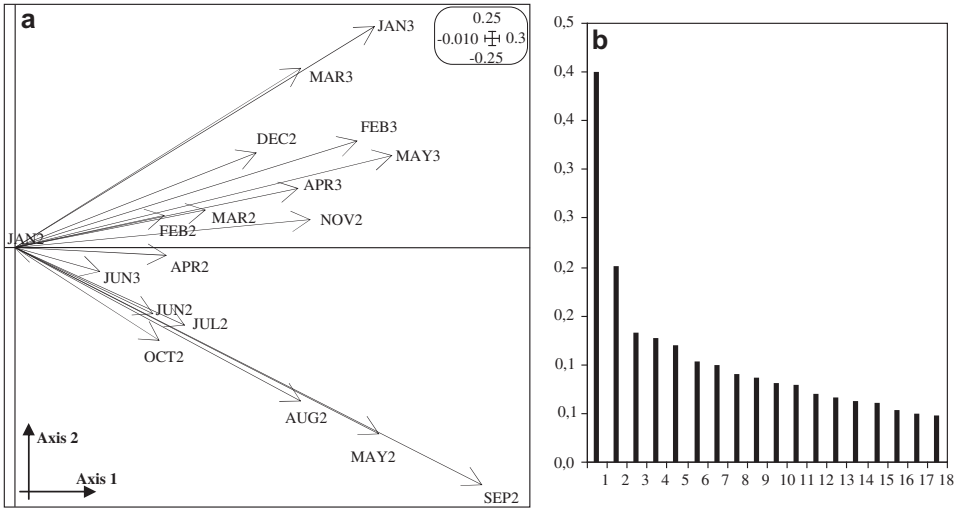


Fig. 2. Interstructure factor map of the STATICO analysis on the Ria de Aveiro data. (a) This map shows the importance of each sampling date in the compromise. Each date is identified by the three first letters of the month followed by a number: 2 for the year 2002 and 3 for the year 2003 (e.g. JAN2 – January 2002). Axis 1 the first principal component; Axis 2 the second principal component. The scale of the graph is given in the rounded box (upper right side). (b) Eigenvalues bar plot of the interstructure.

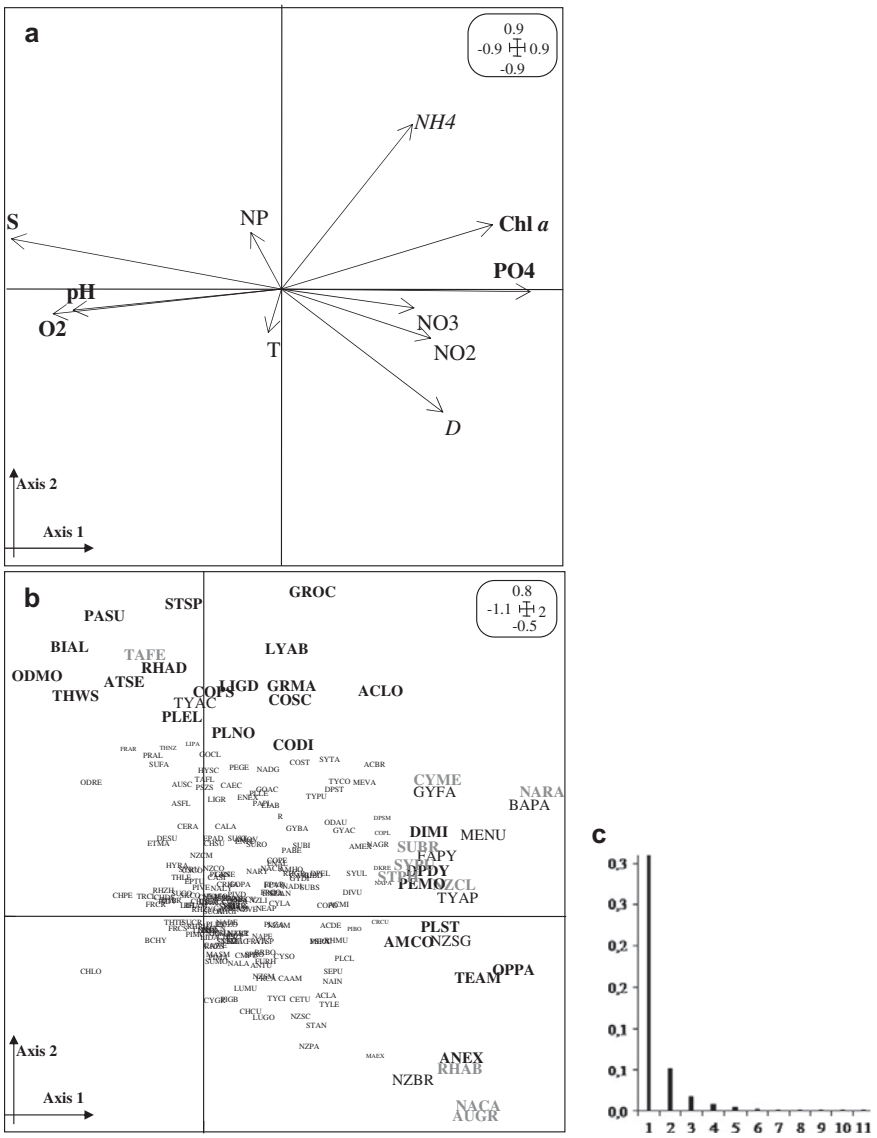


Fig. 3. Compromise factor maps of the STATICO analysis. These maps show the stable part of the diatoms–environment relationships: (a) Environmental variables projected on the first factorial plan. See Table 1 for environmental codes; D – distance to the mouth of the estuary. (b) Diatom abundances projected on the same factorial plan. The labels of most abundance diatom species and thus of importance in these graphics, were with big size letter for clarity. Different gray levels are used to distinguish marine species from freshwater species. See Table 2 for species affinities and codes. Axis 1 the first principal component; Axis 2 the second principal component. The scale of the graph is given in the rounded box (upper right side). Note for different scales. (c) Eigenvalues bar plot of the compromise.

(station 3 high tide) although the latter expressed a better consensus in the relationship species–environment. In fact, this was most clear for January 2003 and February 2003. Specifically the environmental points (end of arrows) for the first group (sites 1LT, 1HT and 2HT) were located on the left-hand side of the first axis corresponding to saline waters with the warmest temperatures, higher values of pH and dissolved oxygen. On the contrary, environmental points (end of arrows) corresponding to the aggregation of sites 2LT, 3LT and 3HT (freshwater stations) were located on the right-hand side of the first axis, which means higher nutrient concentrations (ammonium, phosphate, nitrite, nitrate) and chlorophyll *a* (opposed to salinity). Besides this, for these ones the arrows were mostly short when compared to the other group and in particular for February 2003. For that case, this revealed a strong structure, which means a best fit between the diatoms (corresponding to a majority to freshwater and/or brackish affinities) and environment (short arrows).

4. Discussion

This study has focused on the months with the highest contribution to the co-structure between environmental factors

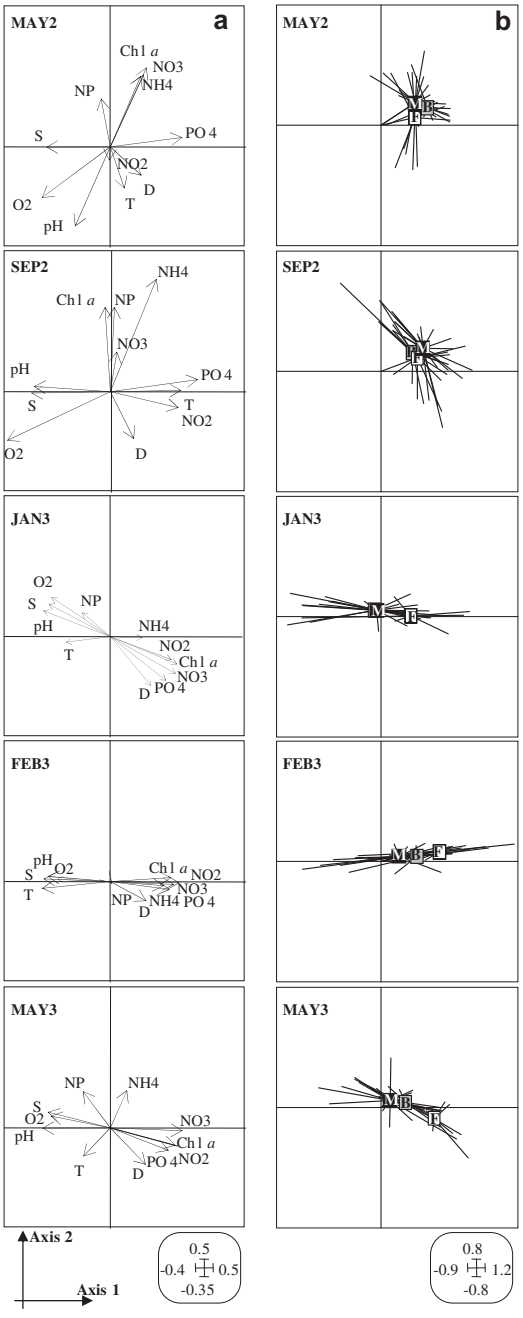


Fig. 4. Trajectories factor plots of the STATICO analysis: (a) Projection of the environmental variables on the first factorial plan. See Table 1 for environmental codes; D – distance to the mouth of the estuary. (b) Projection of the average positions of diatom habitat affinities (only for species that stood out on the compromise diagram). See Table 2 for species habitat affinities. Graphs are given only for the five dates that showed the highest contribution to the co-structure between environmental factors and diatom abundance. See Fig. 3 for legend of months. Axis 1 the first principal component; Axis 2 the second principal component. The scales for axes are given in the rounded box. Note for different scales.

and diatom abundances (May and September 2002 and January, February and May 2003). The relationship between environmental variables and diatoms was strongest in May 2002, September 2002, January 2003, February 2003 and May 2003 while the remaining sampling dates are less structured. This result revealed by the interstructure was failed to be achieved by CCA analysis in Resende et al. (2005). However, these sampling months corresponded to an increase of diversity from April 2002 to June 2002 (afterwards species diversity dropped in the summer months, when community was dominated by *Pseudo-nitzschia seriata* accounting for 73% of the diatom community), January 2003 (an increase of diversity was verified and the community was dominated by *Aulacoseira granulata* and *Tryblionella apiculata*), and the registered higher diatoms densities in September 2002 (5.6×10^5 cells l^{-1}) (Resende et al., 2005). The winter dominance by *A. granulata*, reaching maximal abundances (in the most downstream areas) is a common feature (Muylaert et al., 2000) and the interstructure was able to reveal ecological structure where CCA was inefficient.

In the Ria de Aveiro estuary, the simultaneous analysis of diatom abundances and environmental variables has emphasized the preponderant role of salinity in the spatio-temporal structuring of diatom assemblages. The main spatial structure was a longitudinal gradient from marine to freshwater assemblages. The most abundant diatoms (marine species *Paralia sulcata*, *Biddulphia alternans*, *Odontella mobiliensis*, *Thalassiosira weissflogii*, *Actinoptychus senarius*, *Rhabdonema adriaticum*, *Stauroneis spicula*, *Pleurosigma elongatum*, *Cocconeis pseudomarginata* and freshwater species *Cyclotella meneghiniana*, *Navicula radiosa*, *Surirella brebissonii*, *Synedra pulchella*, *Stauroneis phoenicenteron*, *Nitzschia clausii*, *Rhoicosphenia abbreviata*, *Navicula capitata* and *Aulacoseira granulata*) were associated with different environmental variables. As a result, the stable part of the species–environment relationships mainly consisted of a combined phosphate, chlorophyll *a* and salinity gradient linked to a freshwater-marine species gradient. The marine component of the community was associated with saline waters, high values of pH, dissolved oxygen and low phosphates, while the freshwater component was characteristic of low saline waters and high concentrations of phosphates. Although this salinity gradient is relatively stable in time and space, changes in river discharge and marine waters intrusion induce variations in its position through the course of the year (Soetaert and Herman, 1995; Muylaert et al., 2000). McLusky (1993) claimed that the use of a fixed reference frame to determine spatial variation in a spatio-temporal data set in an estuarine environment will fail to capture all spatially structured variation because longitudinal estuarine gradients vary overtime, what Martin (2003) also designates as the contamination of the spatial data with temporal effects.

The water characteristics of Ria de Aveiro indicates that during low freshwater flow, in late spring and summer months, there is enhanced salinity intrusion upstream estuary. On the other hand, increased precipitation promoted higher freshwater flows moving the salinity intrusion seawards (Lopes et al., 2007). External forcing features (meteorological events, river discharge, and nutrient loading) are major determinants of ecosystem response (Pinckney et al., 1998) and seasonal patterns are strongly influenced by freshwater flow (Adolf et al., 2006). The river discharge (Si loading, and other nutrients) of unusual rainy periods may contribute to a prolonged supply of Si into the system. This river discharge is important, determining diatom variability (Adolf et al., 2006; Gameiro et al., 2007; Lopes et al., 2007) over these periods. On the other hand increases in discharge during late summer were shown to result in washout of phytoplankton from the freshwater tidal reaches of an estuary (Muylaert et al., 2000;

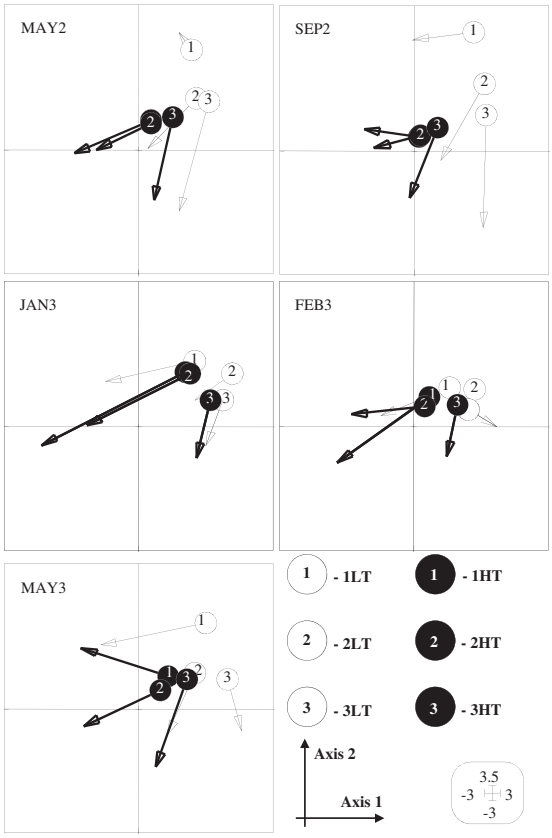


Fig. 5. Projection of the sampling sites on the compromise axes, in terms of both environmental and diatom structure. Graphs are given for each sampling date that showed the highest contribution to the co-structure between environmental factors and diatom abundance. Each site is represented by two points: one is the projection of the row of the diatoms table (circle: origin of arrows), and the other is the projection of the row of the environmental table (end of arrows). The length of the connecting line reveals the disagreement or the consensus between the two profiles (species–environment), i.e., the length of the line is proportional to the divergence between the datasets. When the datasets agree very strongly, the arrows will be short. Likewise, a long arrow demonstrates a locally weak relationship between the environment and diatoms features for that case. See Fig. 3 for legend of months. Axis 1 the first principal component; Axis 2 the second principal component. The scale of the graph is given in the rounded box.

Gameiro et al., 2004; Lionard et al., 2008). In Chesapeake Bay Adolf et al. (2006) reported phytoplankton composition forcing driven mainly by River flow (and its effect on spatial and temporal variability of light and nutrients). These authors found a high diversity of taxa in summer, and environmental forcing by extremes of freshwater flow that elicited a shift to diatoms. This complexity contrasts with other systems with relatively predictable spring diatom blooms. These stable ecosystems with good species evenness and diversity, with abundance patterns related to seasonal cycles contrasts with the unstable situations, in terms of spatial homogeneity and agreement with the environmental conditions.

The more pronounced gradients were observed in January, February and May 2003. Diatom assemblages showed clear longitudinal patterns due to the presence of both marine and freshwater components. In May and September 2002, least structured gradients, marine-estuarine species were in the "freshwater side" of the gradient. This strong longitudinal organization varied on a temporal scale. The most complete gradient in February 2003 could be considered, in terms of bio-ecological categories, as the more structured period of the year, with a combination of strong marine influence in the lower zone and freshwater influence in the upper zone. The best-structured gradients are periods of diatom blooms (abundance and/or diversity).

The dynamics observed in stations 2LT, 3LT and 3HT (at high tide but the furthest from the sea) in January and February 2003 and May 2003 can be explained by the river inputs of water and sediment containing dissolved and particulate nutrients which causes higher nutrient loadings to the estuary (that can be caused by highest freshwater flows) (Domingues and Galvão, 2007; Lopes et al., 2007). The January and February 2003 spatial homogeneity of the intermediate brackish (in low tide) and freshwater zones of the estuary and strong structure corresponded to a good fit between diatoms and environment. Stations 1LT, 1HT and 2HT (with higher marine affinities) presented the most heterogeneous pattern from both an environmental and diatoms point of view (independently from the sampling month). This high heterogeneity was associated with a poor fit between the species abundances and environment. These sites of the estuary were directly under the marine influence and consequently subject to both short-term ebb/flood changes and monthly variations, in particular spring differences. In the marine influenced zone of the estuary was associated a good fit between diatoms and the environment at low tide in May and September 2002. In a system with suggested phosphate limitation (Lopes et al., 2007), May and September 2002 were associated with high values of phosphate.

CCA is useful and effective evaluating the relationships between environmental variables and species distribution. Variability can however be masked or not shown in the first principal axes of CCA. STATICO graphical complementary plots revealed the co-structure between environmental variables and diatom species distribution. The application of this method proved to be well adapted in taking into account the spatio-temporal dynamics of both environmental factors and abundances of diatoms, and the relationships between these two datasets. This study corroborates the conclusions of Carassou and Ponton (2007) demonstrating the method's ability to distinguish the environmental factors which have a general effect on species distribution from those which act only for a given period, location or condition.

5. Conclusions

Like many other temperate estuaries, nutrient enrichment of the catchment area of Ria de Aveiro has been responsible for cultural eutrophication which may induce alteration in phytoplankton assemblages (Lopes et al., 2007). Identifying the ecological variables that regulate phytoplankton dynamics is essential for understanding the consequences of eutrophication problems and biological response to climate change in estuaries. The earlier study of Resende et al. (2005), based on a CCA analysis, while correct in broadly characterizing the ecosystem, failed to capture variability associated with different strengths of environmental forcing. The results presented in the present study add new ecological information on diatom estuarine dynamics (relationships between diatoms and environmental parameters and their space-time structures) and are of importance for the understanding of estuarine ecosystems. In further studies the inclusion of physical parameters, e.g. water column depth, spring-neap cycle as a major driver of temporal variations and diel cycles, could be interesting to explore fully the role of the estuarine and coastal zones.

Acknowledgments

The statistical analyses were run using the ADE-4 package (Thioulouse et al., 1997). The authors are very grateful to the contributors who have made such a valuable tool available. Sincere thanks to Ana Lillebø from CESAM (University of Aveiro) and to Department of Statistics (University of Salamanca) for their inputs and constructive discussions. We also thank Maria José Sá and Annemarie Belenkov in editing the English text. We thank the anonymous referees for their helpful and constructive comments on the manuscript.

References

APHA (American Public Health Association), 1992. Standard Methods for the Examination of Water and Wastewater, 18th ed. American Public Health Association, Washington, DC, p. 991.

Adolf, J.E., Yeager, C.L., Miller, W.D., Mallonee, M.E., Harding Jr., L.W., 2006. Environmental forcing of phytoplankton floral composition, biomass, and primary productivity in Chesapeake Bay, USA. Estuarine. Coastal and Shelf Science 67, 108–122.

Carassou, L., Ponton, D., 2007. Spatio-temporal structure of pelagic larval and juvenile fish assemblages in coastal areas of New Caledonia, southwest Pacific. Marine Biology 150, 697–711.

Dolédec, S., Chessel, D., 1994. Co-inertia analysis: an alternative method for studying species-environment relationships. Freshwater Biology 31, 277–294.

Domingues, R.B., Barbosa, A., Galvão, H., 2005. Nutrients, light and phytoplankton succession in a temperate estuary (the Guadiana, south-western Iberia). Estuarine, Coastal and Shelf Science 64, 249–260.

Domingues, R.B., Galvão, H., 2007. Phytoplankton and environmental variability in a dam regulated temperate estuary. Hydrobiologia 586, 117–134.

Gameiro, C., Cartaxana, P., Brotas, V., 2007. Environmental drivers of phytoplankton distribution and composition in Tagus Estuary, Portugal. Estuarine, Coastal and Shelf Science 75, 21–34.

Gameiro, C., Cartaxana, P., Cabrita, M.T., Brotas, V., 2004. Variability in chlorophyll and phytoplankton composition in an estuarine system. Hydrobiologia 525, 113–124.

Germain, H., 1981. Flore des diatomées. Diatomophycées eaux douces et saumâtres du Massif Armoricain et dês contrées voisines d'Europe Occidentale. Société Nouvelle dês Éditions. Boubée, Paris, pp. 444.

Hall, A., Lucas, M., 1981. Analysis of ammonia in brackish waters by the indophenol's blue technique: comparison of two alternative methods. Revista Portuguesa de Química 23, 205–211.

Hustedt, F., 1985. The Pennate Diatoms, a translation of Husted's "Die Kieselalgen, 2. Teil". With supplement by Norman G. Jensen. Koeltz Scientific Books, Koenigstein, p. 918.

Kocum, E., Underwood, G.J., Nedwell, D.B., 2002. Simultaneous measurement of phytoplankton primary production, nutrient and light availability along a turbid, eutrophic UK east coast estuary/the Colne Estuary. Marine Ecology Progress Series 231, 1–12.

Krammer, K., Lange-Bertalot, H., 1986. Bacillariophyceae, 1. Teil: Naviculaceae. In: Ettl, H., Gerloff, J., Heying, H., Mollenhauer, D. (Eds.), Süßwasserflora von Mitteleuropa, 2/1. Gustav Fischer Verlag, Stuttgart, p. 876.

Krammer, K., Lange-Bertalot, H., 1988. Bacillariophyceae, 2. Teil: Bacillareaceae, Epithemiaceae, Surirellaceae. In: Ettl, H., Gerloff, J., Heying, H., Mollenhauer, D. (Eds.), Süβwasserflora von Mitteleuropa, 2/2. Gustav Fischer Verlag, Stuttgart, p. 596.

Krammer, K., Lange-Bertalot, H., 1991a. Bacillariophyceae, 3. Teil: Centrales, Fragilariaceae, Eunotiaceae. In: Ettl, H., Gerloff, J., Heying, H., Mollenhauer, D. (Eds.), Süβwasserflora von Mitteleuropa, 2/3. Gustav Fischer Verlag, Stuttgart, p. 577.

Krammer, K., Lange-Bertalot, H., 1991b. Bacillariophyceae, 4. Teil: Achnanthaceae. In: Ettl, H., Gärtner, G., Gerloff, J., Heying, H., Mollenhauer, D. (Eds.), Süβ wasserflora von Mitteleuropa, 2/4. Gustav Fischer Verlag, Stuttgart, p. 437.

Legendre, L., Legendre, P., 1979. Ecologie numerique. 1. Le traitement multiple des donnees ecologiques. et les Presses de 1'Universite du Quebec, Masson, Paris, p. 197.

Lemaire, E., Abril, G., De Wit, R., Etcheber, H., 2002. Distribution of phytoplankton pigments in nine European estuaries and implications for in estuarine typology. Biogeochemistry 59, 5–23.

Lionard, M., Muylaert, K., Hanoutti, A., Maris, T., Tackx, M., Vyverman, W., 2008. Inter-annual variability in phytoplankton summer blooms in the freshwater tidal reaches of the Schelde estuary (Belgium). Estuarine, Coastal and Shelf Science 79, 694–700.

Lopes, C.B., Lillebø, A.I., Dias, J.M., Pereira, E., Vale, C., Duarte, A.C., 2007. Nutrient dynamics and seasonal succession of phytoplankton assemblages in a Southern European Estuary: Ria de Aveiro, Portugal. Estuarine, Coastal and Shelf Science 71, 480–490.

Mallin, M.A., Paerl, H.W., 1994. Planktonic trophic transfer in an estuary: seasonal, diel, and community structure effects. Ecology 75, 2168–2184.

Martin, A.P., 2003. Phytoplankton patchiness: the role of lateral stirring and mixing. Progress in Oceanography 57, 125–174.

McLusky, D.S., 1993. Marine and estuarine gradients – an overview. Netherlands Journal of Aquatic Ecology 27, 489–493.

Muylaert, K., Sabbe, K., Vyverman, W., 2000. Spatial and temporal dynamics of phytoplankton communities in a freshwater Tidal Estuary (Schelde, Belgium). Estuarine, Coastal and Shelf Science 50, 673–687.

Peragallo, H., Peragallo, M., 1897–1908. Diatomées Marines de France et des Districts maritimes voisins. Koeltz Scientific Books, Koenigstein, p. 509, pl. I-CXXXVII.

Pinckney, J.L., Paerl, H.W., Harrington, M.B., Howe, K.E., 1998. Annual cycles of phytoplankton community-structure and bloom dynamics in the Neuse River Estuary, North Carolina. Marine Biology 131, 371–381.

Resende, P., Azeiteiro, U., Pereira, M.J., 2005. Diatom ecological preferences in a shallow temperate estuary (Ria de Aveiro, Western Portugal). Hydrobiologia 544, 77–88.

Rodier, J., 1984. L'analyse de l'eau, eaux naturelles, eaux résiduaires, eau de mer (chimie, physico-chimie, bactériologie, biologie), seventh ed., Dunoud, Bordas, Paris. p. 1365.

Round, F., Crawford, R., Mann, D., 1990. The Diatoms. Biology & Morphology of the Genera. Cambridge University Press, Cambridge, p. 747.

Simier, M., Blanc, L., Pellegrin, F., Nandris, D., 1999. Approche simultanée de K couples de tableaux: application à l'étude des relations pathologie végétale – environnement. Revue de statisque appliquée 47, 31–46.

Simier, M., Laurent, C., Ecoutin, J.-M., Albaret, J.-J., 2006. The Gambia River estuary: a reference point for estuarine fish assemblages studies in West Africa. Estuarine, Coastal and Shelf Science 69, 615–628.

Sims, P. (Ed.), 1996. An Atlas of British Diatoms. Biopress Limited, Bristol, p. 601.

Soetaert, K., Herman, P.M.J., 1995. Estimating estuarine residence times in the Westerschelde (The Netherlands) using a box model with fixed dispersion coefficients. Hydrobiologia 311, 215–224.

Strickland, J., Parsons, T., 1972. A Practical Handbook of Seawater Analysis. Canada, Ottawa.

ter Braak, C.J.F., 1986. Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. Ecology 67, 1167–1179.

Thioulouse, J., Chessel, D., 1987. Les analyses multitableaux en écologie factorielle. I. De la typologie d'état à la typologie de fonctionnement par l'analyse triadique. Acta Oecologica, Oecologia Generalis 8, 463–480.

Thioulouse, J., Chessel, D., Dolédec, S., Olivier, J.M., 1997. ADE-4: a multivariate analysis and graphical display software. Statistics and Computing 7, 75–83.

Thioulouse, J., Simier, M., Chessel, D., 2004. Simultaneous analysis of a sequence of paired ecological tables. Ecology 85, 272–283.

Tomas, C., 1996. Identifying Marine Diatoms and Dinoflagellates. Academic Press, Inc., San Diego, p. 598.

Trigueros, J.M., Orive, E., 2001. Seasonal variations of diatoms and dinoflagellates in a shallow, temperate estuary, with emphasis on neritic assemblages. Hydrobiologia 444, 119–133.

Urrutxurtu, I., Orive, E., Sota, A., 2003. Seasonal dynamics of ciliated protozoa and their potential food in an eutrophic estuary (Bay of Byscay). Estuarine, Coastal and Shelf Science 57, 1169–1182.

Vaquer, A., Trousselier, M., Courties, C., Bibent, B., 1996. Standing stock and dynamics of picophytoplankton in the Thau Lagoon (northwest Mediterranean coast). Limnology and Oceanography 41, 1821–1828.

Witkowski, A., Lange-Bertalot, H., Metzeltin, D., 2000. Diatom flora of marine coasts I. In: Lange-Bertalot, H. (Ed.), Iconographia Diatomologica, p. 925. In: K.G., G.V. (Ed.), Ruggell.

PAPER V

An empirical comparison of Canonical Correspondence Analysis and STATICO in identification of spatio-temporal ecological relationships

Mendes S, Fernández-Gómez MJ, Pereira MJ, Azeiteiro UM,

Galindo-Villardón MP.

Journal of Applied Statistics, 2011 (in press) (#)

(#) Additional information:

- First online: available online: 18 November 2011
  http://dx.doi.org/10.1080/02664763.2011.634393
- Publ. Oxfordshire: Taylor & Francis
- ISSN: 0266-4763
- **Indexed in:** Journal Citation Report - Science Citation Index
- **Impact Factors:** IF 2010 (2 years): 0.306; IF 2010 (5 years): 0.449
- **Category:** STATISTICS & PROBABILITY. Ranking: 106/110 – Q. 4 (Quartile 4)
- **Also indexed in**: ABI/Inform, CompuMath Citation Index, Current Index to Statistics, Econlit, MathSciNet, OCLC Databases, Science Citation Index, Scopus and Zentralblatt Math.

Journal of Applied Statistics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/cjas20

An empirical comparison of Canonical Correspondence Analysis and STATICO in the identification of spatio-temporal ecological relationships

Susana Mendes [a] , M. José Fernández-Gómez [b] , Mário Jorge Pereira [c] , Ulisses Miranda Azeiteiro [d e] & M. Purificación Galindo-Villardón [b]

[a] School of Tourism and Maritime Technology, Marine Resources Research Group, Polytechnic Institute of Leiria, Peniche, Portugal

[b] Department of Statistics, University of Salamanca, Salamanca, Spain

[c] Department of Biology, University of Aveiro, Aveiro, Portugal

[d] Department of Sciences and Technology, Universidade Aberta, Porto, Portugal

[e] CFE-Centre for Functional Ecology, Department of Life Sciences, University of Coimbra, Portugal

Available online: 18 Nov 2011

To cite this article: Susana Mendes, M. José Fernández-Gómez, Mário Jorge Pereira, Ulisses Miranda Azeiteiro & M. Purificación Galindo-Villardón (2011): An empirical comparison of Canonical Correspondence Analysis and STATICO in the identification of spatio-temporal ecological relationships, Journal of Applied Statistics, DOI:10.1080/02664763.2011.634393

To link to this article: http://dx.doi.org/10.1080/02664763.2011.634393

Journal of Applied Statistics
2011, iFirst article

An empirical comparison of Canonical Correspondence Analysis and STATICO in the identification of spatio-temporal ecological relationships

Susana Mendes[a]*, M. José Fernández-Gómez[b], Mário Jorge Pereira[c], Ulisses Miranda Azeiteiro[d,e] and M. Purificación Galindo-Villardón[b]

[a]*School of Tourism and Maritime Technology, Marine Resources Research Group, Polytechnic Institute of Leiria, Peniche, Portugal;* [b]*Department of Statistics, University of Salamanca, Salamanca, Spain;* [c]*Department of Biology, University of Aveiro, Aveiro, Portugal;* [d]*Department of Sciences and Technology, Universidade Aberta, Porto, Portugal;* [e]*CFE-Centre for Functional Ecology, Department of Life Sciences, University of Coimbra, Portugal*

(*Received 12 January 2011; final version received 19 October 2011*)

The wide-ranging and rapidly evolving nature of ecological studies mean that it is not possible to cover all existing and emerging techniques for analyzing multivariate data. However, two important methods enticed many followers: the Canonical Correspondence Analysis (CCA) and the STATICO analysis. Despite the particular characteristics of each, they have similarities and differences, which when analyzed properly, can, together, provide important complementary results to those that are usually exploited by researchers. If on one hand, the use of CCA is completely generalized and implemented, solving many problems formulated by ecologists, on the other hand, this method has some weaknesses mainly caused by the imposition of the number of variables that is required to be applied (much higher in comparison with samples). Also, the STATICO method has no such restrictions, but requires that the number of variables (species or environment) is the same in each time or space. Yet, the STATICO method presents information that can be more detailed since it allows visualizing the variability within groups (either in time or space). In this study, the data needed for implementing these methods are sketched, as well as the comparison is made showing the advantages and disadvantages of each method. The treated ecological data are a sequence of pairs of ecological tables, where species abundances and environmental variables are measured at different, specified locations, over the course of time.

Keywords: multitable analysis; multivariate data; ecological data analysis; simultaneous analysis; paired tables

*Corresponding author. Email: susana.mendes@ipleiria.pt

ISSN 0266-4763 print/ISSN 1360-0532 online

http://dx.doi.org/10.1080/02664763.2011.634393
http://www.tandfonline.com

1. Introduction

Various statistical methods available for ecological studies do not provide the best results for all the ecological situations and controversy over the choice of methods of analysis is still present [3]. The data sets obtained repeatedly over time fall broadly into two categories. Usually investigators have counts of many organisms at various locations (biological data) and measurements of abiotic factors at the same locations (chemical or environmental data). Researchers are not able to control any of these variables, but merely observe the values they take; data are observational, and of multivariate nature. In the treatment of these data, tables containing the composition of species–environmental information selected in the same sampling sites can be analyzed, using canonical correlation (CC) [5], Canonical Correspondence Analysis (CCA) [16], canonical non-symmetrical correspondence analysis (CNCA) [25], redundancy analysis (RDA) [26] or co-inertia analysis [2]. CCA is probably the most popular method for using environmental information in a direct manner. It can be obtained by working linear restrictions into the basic correspondence analysis (CA) [6]. CCA includes a step in the algorithm, a regression of the coordinates of the sites on environmental variables (usually linearly independent), which implies that the number of samples should be high compared with the number of environmental variables [21]. This arises as a restriction on the application of the model. However, when a pair of tables is repeated in time or space, it is possible to investigate the temporal or spatial stability of relations between the two sets of tables by the STATICO method (STATIS and co-inertia) [14,24]. STATICO is an application of the STATIS method [9,10] to operators of co-inertia. The aim of this method is to find the stable part in the dynamics of relationships between species and their environment. Each pair consists of a table of abundance of species (species in columns) and another table of environmental variables (variables in columns). The sampling sites (in rows) must be the same in each pair of tables, but may be different between pairs. The environmental variables must be the same in all tables and the list of species as well (although some species may be absent in some tables). The STATICO is really a partial triadic analysis (PTA) [22] of a series of tables that result from the cross co-inertia analysis of each pair of tables. This method combines the objectives of the STATIS, i.e. finding the stable part of the structure of a series of tables, and objectives of co-inertia analysis, i.e. finding the common structure of two data tables. The STATICO analysis, which has been implemented only in the ADE-4 (free) software [23], is much less common than the CCA. The success of CCA is probably due to the availability of this method in statistical packages and to the previous success of CA. But this success should not hide the fact that CCA is only suited to gradient analysis [12] (i.e. to the study of spatial patterns of species in terms of environmental factors and characteristics of communities) and is not always appropriate for the coupling of two tables. Indeed, CCA is very stringent and requires that the species table is analyzed by CA and that the sites are weighted by their richness [3]. These considerations are not suitable for all situations.

The main goal of this study is to compare CCA and STATICO techniques. The advantages and disadvantages of the two methods in addressing the problem of coupling two tables are presented, showing how they can be used simultaneously to extract the stable part of the relationships between species–environmental variables. STATICO and CCA have already been presented and explored in several journals, so their methodological bases are briefly presented. The comparison of the two methods is done on the same data set as the one used by Mendes *et al.* [11] and Resende *et al.* [13]. The results will be explored from a rather practical point of view, on their respective graphical outputs and on their global properties.

2. Principles of CCA

The procedure is summarized in Figure 1(a). CCA is a multivariate method to elucidate the relationships between biological assemblages of species and their environment. The method is designed to extract synthetic environmental gradients from ecological data sets. The gradients are the basis for succinctly describing and visualizing the differential habitat preferences of taxa through an ordination diagram. CCA operates on (field) data on occurrences or abundances (e.g. counts of individuals) of species and data on environmental variables (the explanatory variables) at sites [20]. CCA drift directly from the classical ordination method of CA; in this sense, CCA

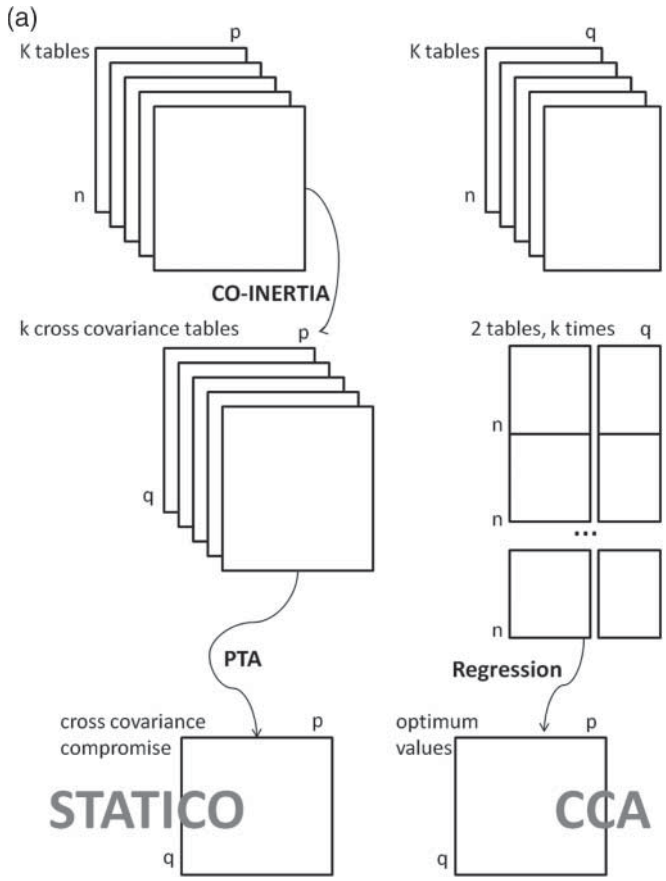


Figure 1. Methods flow charts. (a) STATICO and CCA techniques and (b) PTA technique: construction of the inter-structure matrix and extraction of the compromise table W. The inter-structure between matrices $W_1, \ldots, W_t$ corresponding to the t tables (here sampling months) is analyzed by a simple principal component analysis of the inter-structure matrix. The inter-structure analysis provides an ordination of the t tables and gives a description of the q rows in relation to the typology of the t tables for each of the p variables. The compromise tables are derived from the coordinates of the variables at the different q rows on the principal components of the principal component analysis of the inter-structure matrix. Compromise tables are then analyzed by a simple principal component analysis. This operation indicates which of the variables are of importance in the structure extracted by the principal component analysis of the inter-structure table.

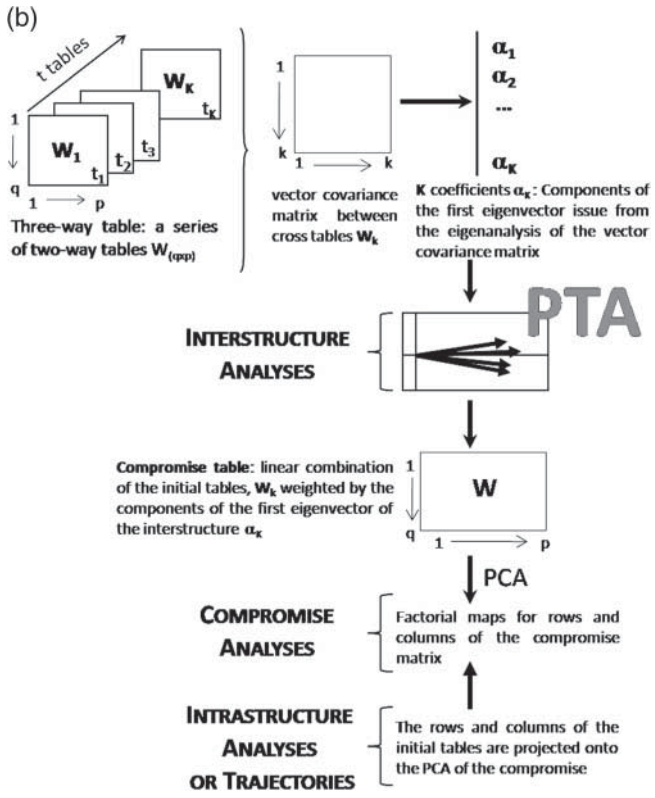


Figure 1. Continued.

can be seen as a constrained CA. In fact, CCA selects ordination axes as in CA, but imposing that they are a linear combination of environmental variables. CCA chooses then the best weights for the environmental variables so as to maximize the dispersion of species scores. Restrictions become less strict as more environmental variables are included in the analysis. When the number of environmental variables included in the analyses is approximately equal to the number of samples collected, CCA is equivalent to CA. Being a restriction of CA, CCA shares properties and drawbacks with CA but some drawbacks can be more easily solved in CCA than in CA [4,8]. The statistical model underlying the CCA is that a species' abundance or frequency is a unimodal function of position along environmental gradients. CCA is an approximation to Gaussian regression under a certain set of simplifying assumptions and is robust to violations of those assumptions [18]. CCA is inappropriate for extremely short gradients, in which species abundance or frequency is a linear or a monotonic function of gradients [17,18]. From a more theoretical point of view, it is possible to arrive at the basic equations of CCA from different perspectives, most of them described in the literature. For instance, the CCA is a maximization of the dispersion of the species scores, using a linear restriction on the site scores [7,17]. Alternatively, CCA has been stated to be weighted least-squares approximation to the weighted averages of the species with respect to the environmental variables [16]. It is possible to put out the CCA in the framework of reciprocal averaging, where the reciprocal averaging algorithm is combined with the regression of site scores onto environmental variables. CCA has also been formulated as a weighted principal component analysis of a matrix of weighted averages [17]. The purpose here is to give a description of CCA following the viewpoint of ter Braak and coworkers [17,20], who described the method as being a combination of CA and multiple regression. It is also important

to stress that CCA is characterized by having a symmetrical objective, i.e. CCA aims to find the agreement between the typology of places from the standpoint of the matrix structure of community composition (i.e. from the point of view of content in terms of species), and the typology of places from the point of view of the structure of the environmental matrix (i.e. from the standpoint of environmental value that occurs in those places). Let

$$u_k = \sum_{i=1}^{n} \frac{y_{ik}}{y_{+k}} x_i \tag{1}$$

the weighted average of k species with respect to any gradient x (environmental variable, synthetic gradient or ordination axis). So, u_k is the weighted average of the gradient values of the sites at which the species occurs. Note that y_{ik} is the abundance of k species in site $i(i = 1, \ldots, n)$ and $k = 1, \ldots, p)$, x_i the value of gradient x at site i and

$$y_{+k} = \sum_{i=1}^{n} y_{ik} \tag{2}$$

the total abundances of k species. For a standardized gradient x, i.e. a gradient for which

$$\sum_{i=1}^{n} \frac{y_{i+}}{y_{++}} x_i = 0 \tag{3}$$

and

$$\sum_{i=1}^{n} \frac{y_{i+}}{y_{++}} x_i^2 = 0 \tag{4}$$

the weighted variance (dispersion) of species scores $u_k(k = 1, \ldots, p)$ is defined by

$$\delta = \sum_{k=1}^{p} \frac{y_{+k}}{y_{++}} u_k^2. \tag{5}$$

Let x be a synthetic gradient, i.e. a linear combination of environmental variables

$$x_i = c_0 + c_1 z_{i1} + c_2 z_{i2} + \cdots + + c_q z_{iq}, \tag{6}$$

where z_{ij} is the value of the environmental variable j in site i and $c_j, j = 1, \ldots, q$, the coefficient or weight. Then, CCA is the method that chooses the optimal weights c_j, i.e. the weights that result in a gradient x for which the weighted variance of the species scores, δ, is maximum. Mathematically, the synthetic gradient x can be obtained by solving an eigenvalue problem; x is the first eigenvector x_1 with the maximum eigenvalue δ [17]. The optimized weights (c_j) are termed canonical coefficients. Each subsequent eigenvector $x_s = (x_{1s}, \ldots, x_{ns})^{\mathrm{T}} (S > 1)$ maximizes the δ, subject to constraint

$$x_i = \sum_{j=1}^{q} c_q z_{ij} \tag{7}$$

and the extra constraint that it is uncorrelated with previous eigenvectors, i.e.

$$\sum_{i=1}^{n} y_{i+} x_{is} x'_{is} = 0 (s' < s). \tag{8}$$

In practice, CCA is used to detect, interpret and predict the underlying structure of the data set based on the explanatory variables (e.g. environmental variables). Furthermore, analyzing the CCA

ordination diagram obtained from the most popular package (CANOCO software [19]), it shows more accurately the dissimilarities between the patterns of occurrence of different species (i.e. the focus scaling is on inter-species distances). Moreover, the scaling type is by biplot scaling meaning that the interpretation is made by the biplot rule. These two options (inter-species distances and biplot scaling) lead to the χ^2 distance as a measure of dissimilarity.

3. Principles of STATICO

The procedure is summarized in Figure 1(a). The STATICO method is particularly well adapted to the study of the changes of species–environment relationships during several T sampling occurrences (in time or space). The STATICO analysis is a symmetric coupling method that joins the STATIS method to operators of co-inertia. In fact, the STATICO begins with the application of the principles of co-inertia analysis. The two ecological data tables Y (containing the values of p species-columns measured at n sites-rows) and Z (containing the values of q environmental variables-columns measured at n sites-rows) produce two representations of the sites in two hyperspaces by separate analysis through a one-table method (like PCA or CA). This separate analysis finds axes maximizing inertia in each hyperspace. Then, each pair of tables is linked by the analysis of the statistical triplet $(Z_t^T D_n Y_t, D_p, D_q)$ and provides an average image of the co-structure (i.e. provides the agreement between the typologies, resulting from species data from environmental data). Note that D_n is the diagonal matrix of row weights (and is what allows the co-inertia analyses of the two tables Y and Z) and D_p and D_q are diagonal matrices of column weights for Y_t and Z_p, respectively. Co-inertia aims to find a couple of co-inertia axes on which the sites are projected. It maximizes the square covariance between the projections of the sites on the co-inertia axes (i.e. maximizes

$$\begin{aligned}\text{COVARIANCE}(ZD_q v_q, YD_p v_p) = &\sqrt{\text{VARIANCE}(ZD_q v_q)} \times \sqrt{\text{VARIANCE}(YD_p v_p)} \\ &\times \text{CORRELATION}(ZD_q v_q, YD_p v_p) \end{aligned} \tag{9}$$

v_p and v_q are D_p and D_q normed vectors in species–environmental space with maximal co-inertia, respectively). The co-inertia leads to a series of cross-tables that will be analyzed by a PTA (Figure 1(b)). PTA is the third step of the STATICO method and begins with the analysis of the statistical triplet $(Z_t^T D_n Y_t, D_p, D_q)$. Let $(Z_t^T D_n Y_t) = W_t$, which has the same p species (in columns) and the same q environmental variables (in rows) at each t occurrence. The compromise co-inertia (according to the definition of PTA, Figure 1(b)) is to find a_t such that $(\sum_{t=1}^{T} a_t W_t, D_p, D_q)$ has a maximum inertia under the constraint

$$\sum_{t=1}^{t} a_t^2 = 1. \tag{10}$$

The t coefficients a_t are normalized components from the first eigenvector obtained by the diagonalization of vectorials covariance matrices between the W_t tables (i.e.

$$\text{COVARIANCE}(W_t, W_t') = Tr(W_t^T D_q W_t', D_p) \tag{11}$$

and $Tr(W)$, the trace of the matrix W, represents the sum of the diagonal of the matrix W). The compromise (W_c) of PTA is here a co-inertia analysis of the fictitious crosstab, form $(Z_c^T D_n Y_c)$, which is a weighted average of T cross-tabulations (in practice the compromise table is a "weighted mean" of $W_t = Z_t^T D_n Y_t$). The compromise table has the columns of Y_t and the rows of Z_t (Figure 1(b)). His analysis provides the eigenvalues, an ordination of its columns (and therefore p variables of Y) and an ordination of its rows (and therefore q variables of Z). Once obtained, W_c is then analyzed by principal components analysis (PCA) and the rows and columns of the individual matrices are projected onto the analysis as supplementary individuals

and supplementary variables, respectively (Figure 1(b)). Thus, the analysis of the compromise gives a factor map that can be used to interpret the structures of this compromise.

4. STATICO versus CCA

The STATICO method is an efficient tool to analyze sequences of paired ecological tables. Its flexibility comes in part from the flexibility of co-inertia analysis [24] which maximizes the square covariance between the species score and the environmental score. CCA also maximizes the square covariance, but with additional constraints, influencing the robustness of the analysis relative to the number of variables. In fact, CCA maximizes the proportion of variance in the species matrix that is explained by the environmental matrix. This difference is due to the metric of CCA (Mahalanobis metric $(X^T DX)^{-1}$), which is linked to the underlying regression step. This metric takes into account the correlation in the data (because it is calculated using the inverse of the variance–covariance matrix), adds constraints in the analysis (the total variance must be equal to 1) and its calculation implies precautions concerning the dimensions of the tables in CCA (tables with numerous environmental variables must be avoided). As a matter of fact, analyses with respect to instrumental variables (as CCA) require a small number of environmental variables. Another problem regarding CCA is related with the abundance effect. CCA removes it and the information given by species absences is not considered. So, when the aim of the study is to analyze limiting factors, species absence is an important part of information and CCA become incongruous and should be avoided [3]. On the other hand, STATICO presents constraints coming from co-inertia analysis (sites are weighted in the same way for the two separate analyses) and from PTA (all the cross-tables must have the same rows and the same columns, i.e. species–environmental variables must be the same in all the pairs of tables). Another problem that emerges with the STATICO analysis is when environmental variables or species vary among tables. In such cases, the analysis of the compromise will not produce a good description of such entities, insofar as the compromise is a square matrix of scalar products among species–environmental variables, respectively. However, despite the restrictions, the STATICO method allows the simultaneous description of species–environmental variables by computing a compromise that represents a species versus environmental variables cross-table. The results between STATICO and CCA are more similar, the more structured data are. This means that when the structure of t tables is strong and all of them have almost the same contribution to this structure, the two methods present comparable and congruous description of species–environment relationship. However, when the stable part of the data is not prominent, indicating that the corresponding tables are less structured (with lower importance in the compromise and presenting of some individual features with minor or different contributions to the structure), the STATICO method (through compromise) will be a better descriptor of the relations between species–environmental variables. This arises because STATICO uses a non-uniform weighting (components of the first eigenvector of the inter-structure) instead of a simple average. In such a case, STATICO will downweight the tables with minor contribution or less structured, and the resulting compromise will be nearer to the other tables [24].

5. Case study

Resende *et al.* [13] described for Ria de Aveiro (Western Portuguese Atlantic Coast) a diatom composition that resembled other European temperate estuaries. However, no seasonal pattern of diatom density was found. CCA was used to identify the environmental variables governing the composition and structure of diatom assemblages (together with the study of ecological preferences). Although CCA is appropriated when the responses of the dependent variables are expected to be unimodal along environmental gradients [19], such as in this case, this technique does not consider the *a priori* information on time or space. This information is presented *a posteriori*,

when the results are displayed. Therefore, a mixture of space–time effects is produced, showing evidence of the strongest effects compared with the weaker ones [11]. This study was revised by Mendes *et al.* [11], using STATICO analysis. This statistical procedure allows the identification of spatial or temporal structures and the permanence of these structures in time or space. In order to widen the discussion in the context of multitable analyses, the aim here is to make a comparison between STATICO and CCA based on the global mathematical properties of the methods, and the biological aims of the study.

Environmental variables and species considered here are the same for both methods and take into account the selection made by Resende *et al.* [13]. These data sets are arranged in two tables: one table with 108 rows and 231 columns, containing the species data, and one table with 108 rows and 7 columns, containing the environmental variables. The rows in both the tables correspond to three sampling stations located on Canal de Mira, a branch of Ria de Aveiro (an elongated shallow arm considered as a small estuary on the northwest coast of Portugal) at two conditions: low and high tide. These three sites were sampled 18 times, from January 2002 to June 2003 (monthly, in sub-surface waters, at new moon periods). The 231 columns of the species data table correspond to 231 diatoms species. The seven environmental variables are physico-chemical measures. A detailed description of the sampling sites, the process of collection of variables and their quantification is given by Mendes *et al.* [11] and Resende *et al.* [13].

Whether from the standpoint of CCA, or STATICO's view, the aim was to investigate the diatom ecological preferences along the Canal de Mira (Ria de Aveiro).

5.1 *CCA results*

The interpretations of detail and graphical outputs were presented by Resende *et al.* [13], so the results are presented here briefly. In summary, it can be concluded that the spatial distribution of diatoms was clearly determined by salinity, whereas the temporal distribution was mainly determined by temperature (Figure 2(a)). The most saline stations (sites S1 and S2 at high tide and S1 at low tide, Figure 3(a)) were characterized by higher abundance of marine species. The seven environmental variables considered in the CCA explained 18.2% of the total variation of the diatom assemblages. The first two axes of the species–environmental variables biplot alone accounted for 11.3% of the total variability and the goodness-of-fit was 62%. In the plan 1-2, salinity, temperature and distance to the mouth of the estuary played a major role, whereas ammonium and phosphate presented a reduced responsibility to explain the total variability of the diatom community (Figure 3(b)). This is an important analysis, and will be developed later (when comparing the two methods), since it is necessary to observe the third and fourth axes of the CCA diagram (Figure 2(a)), in order to observe the influence of these variables in the community. This is a configuration that usually is neglected by researchers. Most of them do not seek information beyond the first principal plan.

5.2 *STATICO results*

The results of the STATICO method were presented previously by Mendes *et al.* [11]. The first result from STATICO analyses, the inter-structure diagram (Figure 2(b)), took into account the 7 environmental variables and the 231 diatoms (identified in the study period) and showed the changes in their relationship over time. The months with a stronger co-structure were September, May and August 2002. Thus, the species–environment relationship was much more structured and stronger in those dates. The species compromise map showed that the spatial distribution of diatoms (Figure 2(c)) was clearly determined by salinity. On the other hand, the stable part of species–environment relationship was presented (mostly) as a combination of the phosphate (correlated with the distance from the mouth of the estuary) and the salinity gradient (Figure 2(d)). The

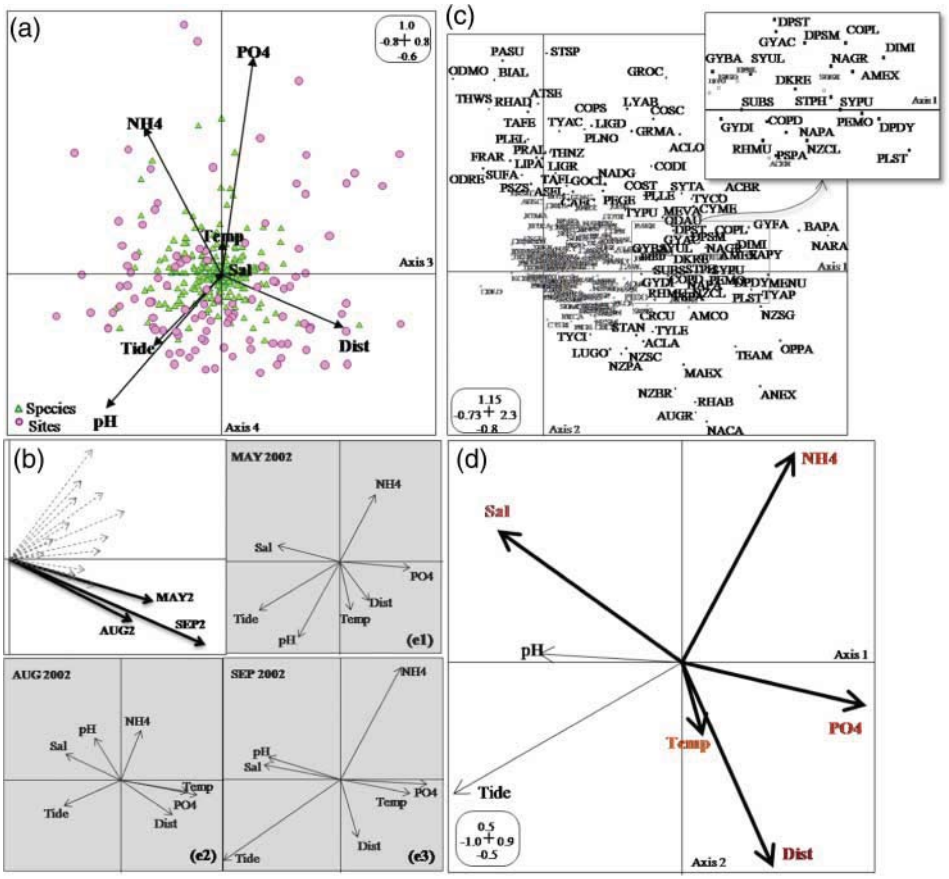


Figure 2. (a) CCA triplot plan 3-4, (b) STATICO inter-structure plot, (c) STATICO compromise plan for environmental variables, (d) STATICO compromise plan for species. The labels of most abundance diatom species and thus of importance in these graphics were with big size letter for clarity. Different gray levels are used to distinguish marine species from freshwater species. See Table 1 for species codes and (e) STATICO trajectories plan for environmental variables. Sal, salinity; Temp, water temperature; NH4, ammonium; PO4, phosphate; Dist, distance to the mouth of the estuary.

Table 1. List of the codes of diatoms species. Codes in bold represent species that stood out in the compromise factor map (Figure 2(c)).

Code	Taxa	Code	Taxa
ACBR	*Achnanthes brevipes*	**NACY**	*Navicula cryptotenella*
ACCO	*Achnanthes coarctata*	**NADE**	*Navicula delawarensis*
ACDE	*Achnanthes delicatula*	**NADG**	*Navicula digitoradiata*
ACLA	*Achnanthes lanceolata*	**NADI**	*Navicula directa*
ACLO	***Achnanthes longipes***	NAGR	*Navicula gregaria*
ACMI	*Achnanthes minutissima*	NAIN	*Navicula integra*
AMCO	***Amphora commutata***	NALA	*Navicula lanceolata*
AMEX	*Amphora exigua*	NALY	*Navicula lyra*
AMHO	*Amphora holsatica*	NAPA	*Navicula palpebralis*
AMOV	*Amphora ovalis*	NAPE	*Navicula peregrina*
ANEX	***Anorthoneis excentrica***	**NARA**	***Navicula radiosa***
ANTU	*Aneumastus tusculus*	NARY	*Navicula rhynchocephala*

(*Continued*).

Table 1. Continued.

Code	Taxa	Code	Taxa
ASFL	*Asteromphalus flabellatus*	NAVI	*Navicula viridula*
ATSE	***Actinoptychus senarius***	NEAM	*Neidium ampliatum*
ATSP	*Actinoptychus splendens*	NEAP	*Neidium apiculatum*
AUGR	***Aulacoseira granulata***	NEDU	*Neidium dubium*
AUSC	*Auliscus sculptus*	NZAC	*Nitzschia acicularis*
BAPA	***Bacillaria paxillifer***	NZAM	*Nitzschia amphibia*
BCHY	*Bacteriastrum hyalinum*	**NZBR**	***Nitzschia brevissima***
BIAL	***Biddulphia alternans***	**NZCL**	***Nitzschia clausii***
BIBD	*Biddulphia biddulphiana*	NZCM	*Nitzschia commutata*
BRBO	*Brebissonia boeckii*	NZCO	*Nitzschia closterium*
CAAM	*Caloneis amphisbaena*	NZIN	*Nitzschia insignis*
CACR	*Caloneis crassa*	NZLI	*Nitzschia linearis*
CAEC	*Campylodiscus echeneis*	NZPA	*Nitzschia palea*
CAIN	*Campylodiscus innominatus*	NZSC	*Nitzschia scalpelliformis*
CALA	*Caloneis latiuscula*	**NZSG**	***Nitzschia sigma***
CASI	*Caloneis silicula*	NZSM	*Nitzschia sigmoidea*
CAVL	*Cavinula lacustris*	NZVE	*Nitzschia vermicularis*
CAWE	*Caloneis westii*	NZVT	*Nitzschia vitrea*
CERA	*Cerataulus radiatus*	ODAU	*Odontella aurita*
CETU	*Cerataulus turgidus*	**ODMO**	***Odontella mobiliensis***
CHCU	*Chaetoceros curvisetus*	ODRE	*Odontella regia*
CHDE	*Chaetoceros decipiens*	**OPPA**	***Opephora pacifica***
CHDS	*Chaetoceros densus*	PABE	*Parlibellus berkeleyi*
CHEI	*Chaetoceros eibenii*	PAPL	*Parlibellus plicatus*
CHLO	*Chaetoceros lorenzianus*	**PASU**	***Paralia sulcata***
CHPE	*Chaetoceros peruvianus*	PEGE	*Petrodictyon gemma*
CHSU	*Chaetoceros subsecundus*	**PEMO**	***Petroneis monilifera***
CMPU	*Cosmioneis pusilla*	PIBO	*Pinnularia borealis*
CODI	***Cocconeis disculus***	PIBR	*Pinnularia braunii*
COPD	*Cocconeis pediculus*	PIBS	*Pinnularia brebissonii*
COPE	*Cocconeis peltoides*	PIDV	*Pinnularia divergens*
COPL	*Cocconeis placentula*	PIGB	*Pinnularia gibba*
COPS	***Cocconeis pseudomarginata***	**PIMA**	*Pinnularia maior*
COSC	***Cocconeis scutellum***	**PIMC**	*Pinnularia microstauron*
COST	*Cocconeis stauroneiformis*	**PISD**	*Pinnularia sudetica*
CRCU	*Craticula cuspidata*	**PISU**	*Pinnularia subcapitata*
CRHA	*Craticula halophila*	**PIVD**	*Pinnularia viridis*
CSCO	*Coscinodiscus concinnus*	**PIVE**	*Pinnuavis elegans*
CYAF	*Cymbella affinis*	**PLAN**	*Pleurosigma angulatum*
CYGR	*Cumbella gracilis*	**PLCL**	*Placoneis clementis*
CYLA	*Cymbella lanceolata*	**PLEL**	***Pleurosigma elongatum***
CYME	***Cyclotella meneghiniana***	**PLGA**	*Placoneis gastrum*
CYSO	*Cymatopleura solea*	**PLLE**	*Plagiotropis lepidoptera*
DESU	*Delphineis surirella*	**PLLV**	*Pleurosira laevis*
DIMI	***Dimeregramma minor***	**PLNO**	***Pleurosigma normanii***
DIVU	*Diatoma vulgare*	**PLST**	***Plagiogramma staurophorum***
DKRE	*Donkinia rectum*	**PRAL**	*Probostia alata*
DPDY	***Diploneis didyma***	**PSPA**	*Psammodictyon panduriforme*
DPEL	*Diploneis elliptica*	**PSZP**	*Pseudo-nitzschia pungens*
DPSM	*Diploneis smithii*	**PSZS**	*Pseudo-nitzschia seriata*
DPST	*Diploneis stroemii*	**RHAB**	***Rhoicosphenia abbreviata***
DTBR	*Ditylum brightwellii*	**RHAD**	***Rhabdonema adriaticum***
ENAL	*Entomoneis alata*	**RHAM**	*Rhaphoneis amphiceros*
ENEX	*Eunotia exigua*	**RHGB**	*Rhopalodia gibberula*
ENPA	*Entomoneis paludosa*	**RHGI**	*Rhopalodia gibba*
ENPE	*Eunotia pectinalis*	**RHMU**	*Rhopalodia musculus*
EPAD	*Epithemia adnata*	**RHRU**	*Rhopalodia rupestris*
EPAR	*Epithemia argus*	**RHZH**	*Rhizosolenia hebetata*
EPTU	*Epithemia turgida*	**RHZS**	*Rhizosolenia stolterfothii*

(*Continued*).

Table 1. Continued.

Code	Taxa	Code	Taxa
ETMA	*Eunotogramma marinum*	**RHZY**	*Rhizosolenia styliformis*
EUZO	*Eucampia zodiacus*	**SEPU**	*Sellaphora pupula*
FAFO	*Fallacia forcipata*	**SKCO**	*Skeletonema costatum*
FAPY	***Fallacia pygmaea***	**SPAL**	*Stephanodiscus alpinus*
FRAR	*Fragilaria arcus*	**SPRO**	*Stephanodiscus rotula*
FRCA	*Fragilaria capucina*	**STAN**	*Stauroneis anceps*
FRCO	*Fragilaria construens*	**STPH**	***Stauroneis phoenicenteron***
FRCR	*Fragilaria crotonensis*	**STSM**	*Stauroneis smithii*
FRCS	*Fragilariforma constricta*	**STSP**	***Stauroneis specula***
FRVA	*Fragilaria vaucheriae*	**SUAN**	*Surirella angusta*
FURH	*Frustulia rhomboides*	**SUBI**	*Surirella brightwellii*
FUVU	*Frustulia vulgaris*	**SUBR**	***Surirella brebissonii***
GOAC	*Gomphonema acuminatum*	**SUBS**	*Surirella biseriata*
GOCL	*Gomphonema clevei*	SUCA	*Surirella capronii*
GOPA	*Gomphonema parvulum*	SUCN	*Surirella constricta*
GOTR	*Gomphonema truncatum*	SUCO	*Surirella comis*
GRMA	***Grammatophora marina***	SUCR	*Surirella crumena*
GROC	***Grammatophora oceanica***	SUFA	*Surirella fastuosa*
GRSE	*Grammatophora serpentina*	SULI	*Surirella linearis*
GYAC	*Gyrosigma acuminatum*	SUMO	*Surirella moelleriana*
GYBA	*Gyrosigma balticum*	SUOV	*Surirella ovalis*
GYDI	*Gyrosigma distortum*	SURO	*Surirella robusta*
GYFA	***Gyrosigma fasciola***	SUST	*Surirella striatula*
HYMA	*Hyalodiscus maculatus*	SYCA	*Synedra capitata*
HYRA	*Hyalodiscus radiatus*	**SYPU**	***Synedra pulchella***
HYSC	*Hyalodiscus scoticus*	SYRU	*Synedra rumpens*
LIAB	*Licmophora abbreviata*	SYTA	*Synedra tabulata*
LIDA	*Licmophora dalmatica*	SYUL	*Synedra ulna*
LIFL	*Licmophora flabellata*	**TAFE**	***Tabellaria fenestrata***
LIGD	***Licmophora grandis***	TAFL	*Tabellaria flocculosa*
LIGR	*Licmophora gracilis*	**TEAM**	***Terpsino¸ ammericana***
LIPA	*Licmophora paradoxa*	THLE	*Thalassiosira leptopa*
LUGO	*Luticola goeppertiana*	THNZ	*Thalassionema nitzschioides*
LUMU	*Luticola mutica*	THTE	*Thalassiosira tenera*
LUNI	*Luticola nivaloides*	**THWS**	***Thalassiosira weissflogii***
LYAB	***Lyrella abrupta***	TRAS	*Trachyneis aspera*
LYAT	*Lyrella atlantica*	TRCL	*Trachyneis clepsydra*
MAEX	*Mastogloia exigua*	TRFA	*Triceratium favus*
MASM	*Mastogloia smithii*	**TYAC**	***Tryblionella acuminata***
MECC	*Meridion circulare*	**TYAP**	***Tryblionella apiculata***
MENU	***Melosira numuloides***	TYCI	*Tryblionella circumsuta*
MEVA	*Melosira varians*	TYCO	*Tryblionella coarctata*
NABR	*Navicula brasiliana*	TYLE	*Tryblionella levidensis*
NACA	***Navicula capitata***	TYPU	*Tryblionella punctata*
NACR	*Navicula cryptocephala*		

first two axes of compromise, referring to species–environmental variables, showed a goodness-of-fit of 38%. By projecting the columns of the tables (species–environmental parameters) at each site on the compromise PCA factor map was possible to observe the characteristics of each table that were shared with the compromise and which were found to have individual features. The most complete and stronger gradients were observed in the months of September, May and August 2002 (in order of decreasing importance) (Figure 2(e1)–(e3)). This means that these dates (which precisely match the diatom blooms) had a greater contribution to the co-structure between environmental variables and the abundance of diatoms (this confirms the interpretations made by the inter-structure result). Finally, the co-structure graphics showed the dynamics of diatom species–environment relationships and highlighted differences between sites (Figure 3(d)). The

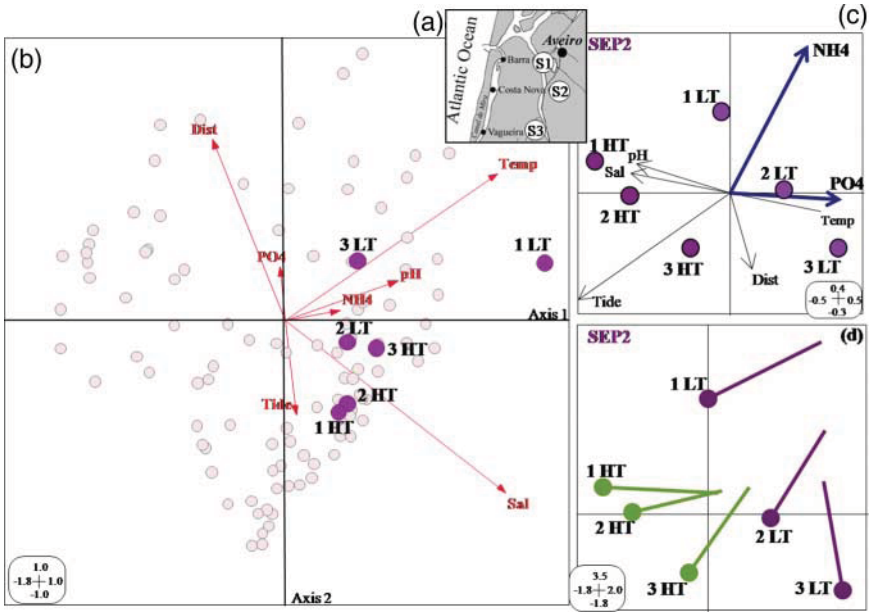


Figure 3. (a) Samples map; (b) CCA plan 1-2 September 2002; (c) STATICO trajectories; (d) STATICO co-structure: projection of the samples with respect to tide condition and site on the first factorial plan of the compromise analysis, for September. Each sample is represented by two points: one is the projection of the row of the species table (circle: origin of arrows), and the other is the projection of the row of the environmental table (end of arrows). The length of the connecting line reveals the disagreement or the consensus between the two profiles (species–environment), i.e. the length of the line is proportional to the divergence between the data sets. When the data sets agree very strongly, the arrows will be short. Likewise, a long arrow demonstrates a locally weak relationship between the environment and diatoms features for that case. Each site is identified by the tide condition (low tide, LT and high tide, HT) followed by the site code (1 for S1, 2 for S2 and 3 for S3) (e.g. LT1, site S1 in low tide). See Figure 2 for environmental labels.

best co-structures were those that correspond to periods where diatoms had a bloom (September, May and August 2002), i.e. a strong structure and good fit between species and the environment. In particular, the co-structure was quite strong in some months of study, as the arrowheads and the positions of the circles in each site-tide were in general quite close (with local exceptions).

5.3 *Comparison of results*

As for the salinity and distance to the mouth of the estuary, the results of the two approaches were consistent (Figures 3(b) and 2(d)). They were inversely related, i.e. the greater the distance from the mouth, the lower the salinity. This is coherent because in the stations farthest from the mouth of the estuary, there is a predominance of freshwater, while in the nearest stations, the waters are more saline. In relation to the temperature, it is noted how this variable was one of the most important in the CCA (Figure 3(b)). However, in the compromise space of STATICO analysis, it appeared with a minor relevance (Figure 2(d)). The reason to this is the fact that temperature is a variable that generally has a high variability (lower in the colder months, higher during warmer months), which, incidentally, is something previously known and that is really what the CCA put in evidence. This happened because CCA is a method that performs a global analysis (setting in the first principal plan, the large variation between months). On the other hand, the temperature variation over a month is low (although in a cold month temperature is consistently low and in a warm month, the

temperature is always high, no relevant variations inside each month is observed). This variability (or lack of it) was precisely what the STATICO has demonstrated through the graphical results obtained. This was clearly visible through the environmental variable trajectories, over the months studied (see Figure 2(e1)–(e3) for the months, where the structures were the strongest). As was possible to observe, the temperature in each month shows some variability, but does not stand out compared with the variability presented by the other variables. Through the CCA method, it was also noted that the variables temperature and salinity were nearly independent (Figure 3(b)). This contradicted the results obtained with the compromise factor map of STATICO, whose correlation was found negative (Figure 2(d)). Indeed, observing the detailed diagram of the trajectories for the environmental variables (Figure 2(e1)–(e3)), a negative correlation was found only in a few months (precisely those that most contributed to the co-structure, i.e. May, August and September 2002). For other dates, this relationship was direct or *quasi*-independent. This is precisely what was reported by the CCA, which is reported what happens "on average". In fact, this means that the data are reported in a widespread manner reduces (or mask) the individual relationships within each month. This resulted in "an average" relationship, designed in the first principal plan as an independent relationship between temperature and salinity. Another difference can be observed when analyzing the principal plan resulting from the CCA with the compromise of STATICO (Figures 3(b) and 2(d)): the ammonium, phosphate and pH do not appear as relevant variables through the CCA (Figure 3(b)), but appeared as important variables in the factor map of STATICO compromise (especially ammonium and phosphate) (Figure 2(d)). Looking at the trajectories that represent the variables in each month (Figure 2(e1)–(e3)), it was found that these three variables have always been important in each month. This importance becomes even more relevant in the months whose co-structure is more pronounced, corresponding to a higher contribution to the compromise, and larger variability compared with other variables in those months (e.g. September 2002, which was the largest contributor to the compromise). In the remaining months, it was found that the vectors of ammonium and phosphate were always the longest (in their ranges). This showed that within each month, these variables were very important to discriminate between the sampling stations. Moreover, in most months, ammonium and phosphate appeared positively or were close to independent. By observing the month which most contributed to the compromise, i.e. the month whose co-structure of the species–environment was stronger (Figures 3(c) and (d)), it seems that only the station nearest the mouth of the estuary (Figure 3(a)) might actually be differentiated from others. This is clearly highlighted by the information provided by the principal plan of the CCA (Figure 3(b)). Moreover, this difference is probably due to the high temperature (whose vector is longer and presented high correlation with S1 at low tide, Figure 3(b)). Regarding the other variables, they showed no significant differences. Nevertheless, this interpretation must be made with caution, since the position of points near the origin may be only apparent (and the CCA does not provide, through their results, information on quality of representation). However, by observing the trajectories of environmental variables (Figure 3(c)), it was possible to observe a much clearer differentiation between the sampling stations. So, it should be noted that the three stations at high tide have low values of ammonia and phosphate and, additionally, that two of these stations (S1 and S2, i.e. those near the mouth of the estuary, Figure 3(a)) have patterns of high values of salinity and pH. These two stations at low tide have higher phosphate content. However, the information provided by the plan 3-4 (Figure 2(a)) on the CCA (this information is normally ignored by most authors, because only these limit themselves to interpreting the information contained in the factorial plan 1-2) shows similar results to those provided by the STATICO. The three stations at low tide presented a higher content of ammonium and phosphate. These variables, as well as pH, were strictly determined by axis 4, although they correlate positively with each other, but negatively with pH. Even though the information provided by the plan 3-4 is important, the authors usually do not gather it. The authors simply gather information from two or three axes (maximum), and they do not analyze the information from other axes that are usually regarded

by many as residual axes. However, through STATICO diagrams, it was possible to observe that the information gathered by this method was much more emphasized and clear.

6. Discussion and conclusions

The main problems in ecological studies are usually answered by classical coupling methods as CCA. The use of more sophisticated methods provides additional results and useful information of great interest to ecologists [3]. The works recently published by Carassou and Ponton [1], Mendes *et al.* [11] and Simier *et al.* [15] are good examples. However, characterized by the properties of collected data and the objectives of the study, the choice of the "better" statistical method can be a very difficult task. In fact, there are no better or worse methods. There are rather theoretical considerations that must be taken into account, as well as the purpose to which it intends to answer. As demonstrated by this work, the two different approaches allow for meaningful and relevant conclusions. They have different (dis)advantages, whose drawbacks or positive aspects can be often bypassed by researchers. Therefore, although at first glance, the obtained conclusions appear to be different, they are entirely related to the objective of the study and how the data are centered in each of the methods. Whereas, CCA works with the entire matrix at once and each value is centered with respect to the average of all data in all matrices or tables (here months), in STATICO, each matrix is centered separately. The conclusions obtained for the temperature and salinity reflect this difference. For instance, temperature assumes low values in some of the data matrices (here in winter month's tables) in contrast to the others (here in the summer month's tables). So this is one of the variables that showed larger variability and therefore come out clearly revealed in the CCA first principal plan, as in this example. On the other hand, given that the CCA is a global analysis, the variability within groups may be masked, or not to be expressed in the first principal plan, also shown in this example. So, the variability of the most significant effects (in this case, the temporal and the distance to the mouth) masks the variability within the seasons and the conditions found within each month (seasonal variations are mostly linked to water temperature, and the corresponding between-season structures are trivial, summer–winter opposition). Compared with how the STATICO operates, in winter the temperature is centered with respect to the average of the winter. Despite being colder in winter, the temperature is similar in all sampling stations. In summer, the temperature is higher, but the data are centered with respect to the average of the summer month. As a result, the temperature is also similar in all sampling stations. Therefore, given the lack of variability from month to month, the temperature appears here as a less important variable to discriminate the sampling stations, as it is within a small range of variability (in both seasons). Conversely, and in relation to salinity, this varies greatly (and always) among the sampling sites. Moreover, this is always observed in each month and so salinity is a variable that turns out to be important for both methods (Figures 2(d), 2(e1)–(e3), 3(a) and 3(c)).

The main advantage of STATICO is the optimality of the compromise (maximization of the similarity with all the initial tables). It gives a compromise of co-structures, which means that it displays the stable part of species–environment relationship variations. One of the important points that must be taken into account when using these methods, concerns the structure of the data (and its strength). When species–environment relationships are very strong, chronological structures may disappear with the STATICO method [21]. Conversely, when the structure is less strong (or some tables present different contributions to the co-structure), STATICO will be the better option to describe the species–environment relationship. Besides being a good option for the description of the stable part of these relationships, STATICO also enables the description of seasonal variations or long-term changes in the species–environment relationships. This advantage comes from the fact that STATICO computes a consensus of species–environment relationships at each date. The aim of a particular study should define the way data are organized

(site samples × variables × time or time samples × variables × sites). Another important point in the STATICO analysis is related to the numerical constraints on the parameters of the *k*-table, i.e. the number of species, of environmental variables, of sampling sites, and of dates. STATICO presents constraints on the species–environmental variables, which should always be identical: same species and same environmental variables for all dates. That is, the two series of tables must have the same dates, but sites can differ among dates (although they must be equal for the two tables of a pair). In relation to this restriction, CCA also has limitations in this sense. It requires that the number of samples must be high compared with the number of environmental variables. CCA already proved to be quite efficient (the number of existing publications is a good proof). A good example of this efficiency is achieved when the uncorrelated variables are few (conversely, it will become quite unstable when the variables are correlated). For most studies, it is possible to recognize that the STATICO analysis presents information that can be much richer than that offered by the CCA. It allows visualizing the variability within groups (in this case, within the month). STATICO graphic results can be very detailed and easier to interpret. In the example presented here, through the complementary results provided by the graphical outputs, it was possible to know not only what are the months in which the co-structure between variables and species is stronger, but also whether there is a common structure, and to what extent each one of the months are "separated" from the common pattern and what are the reasons for moving away from it (which can be done simply by analyzing the different graphs of trajectories that can be generated).

The two methods presented here reveal important features in the data set and their relevance from the ecological point of view is undeniable. A better ecological interpretation is possible using either STATICO or CCA. Both methods can be considered as they are equally suitable for the purpose, being none better when compared with each other. What is noteworthy is that depending on the type of variability that occurs within and between each data matrix, the variables that are emphasized by the first principal plan of the CCA approach, may or may not coincide with what STATICO emphasizes. Thus, multivariate data analyses (widely used to identify and understand the structure of ecological communities) should always be chosen according to the purpose of the study and the responses that the researcher wants to see answered.

Acknowledgements

The statistical analyses were run using the ADE-4 package [23]. The authors are very grateful to the contributors who have made such a valuable tool available. Our sincere thanks to the Department of Statistics (University of Salamanca) for their inputs and constructive discussions.

References

[1] L. Carassou and D. Ponton, *Spatio-temporal structure of pelagic larval and juvenile fish assemblages in coastal areas of New Caledonia, southwest Pacific*, Mar. Biol. 150 (2007), pp. 697–711.

[2] S. Dolédec and D. Chessel, *Co-inertia analysis: An alternative method for studying species–environment relationships*, Freshwater Biol. 31 (1994), pp. 277–294.

[3] S. Dray, D. Chessel, and J. Thioulouse, *Co-inertia analysis and the linking of ecological data tables*, Ecology 84 (2003), pp. 3078–3089.

[4] H.G.J. Gauch, *Noise reduction by eigenvector ordinations*, Ecology 63 (1982), pp. 1643–1649.

[5] R. Gittins, *Canonical Analysis, a Review with Applications in Ecology*, Springer-Verlag, Berlin, 1985.

[6] M.O. Hill, *Correspondence analysis: A neglected multivariate method*, J. R. Stat. Soc. Ser. B Stat. Methodol. 23 (1974), pp. 340–354.

[7] K.W. Johnson and N.S. Altman, *Canonical correspondence analysis as an approximation to Gaussian ordination*, Environmetrics 10 (1999), pp. 39–52.

[8] R.H. Jongman, C.J.F. ter Braak, and O.F.R. van Tongeren, *Data Analysis in Community and Landscape Ecology*, PUDOC, Wageningen, The Netherlands, 1987.

[9] C. Lavit, *Analyse Conjointe De Tableaux Quantitatifs*, Masson, Paris, 1988.
[10] C. Lavit, Y. Escoufier, R. Sabatier, and P. Traissac, *The act (statis) method*, Comput. Statist. Data Anal. 18 (1994), pp. 97–119.
[11] S. Mendes, M.J. Fernández-Gómez, P. Resende, M.J. Pereira, M.P. Galindo-Villardón, and U.M. Azeiteiro, *Spatio-temporal structure of diatom assemblages in a temperate estuary. A statico analysis*, Est. Coast Shelf Sci. 84 (2009), pp. 637–664.
[12] M.W. Palmer, *Putting things in even better order: The advantages of canonical correspondence analysis*, Ecology 74 (1993), pp. 2215–2230.
[13] P. Resende, U.M. Azeiteiro, and M.J. Pereira, *Diatom ecological preferences in a shallow temperate estuary (Ria De Aveiro, Western Portugal)*, Hydrobiologia 544 (2005), pp. 77–88.
[14] M. Simier, L. Blanc, F. Pellegrin, and D. Nandris, *Approche Simultanée De K Couples De Tableaux: Application À L'étude Des Relations Pathologie Végétale – Environnement*, Rev. Statist. Appl. 47 (1999), pp. 31–46.
[15] M. Simier, C. Laurent, J.M. Ecoutin, and J.J. Albaret, *The Gambia River estuary: A reference point for estuarine fish assemblages studies in West Africa*, Est. Coast Shelf Sci. 69 (2006), pp. 615–628.
[16] C.J.F. ter Braak, *Canonical correspondence analysis: A new eigenvector technique for multivariate direct gradient analysis*, Ecology 67 (1986), pp. 1167–1179.
[17] C.J.F. ter Braak, *The analysis of vegetation–environment relationships by canonical correspondence analysis*, Vegetatio 69 (1987), pp. 69–77.
[18] C.J.F. ter Braak and I.C. Prentice, *A theory of gradient analysis*, Adv. Ecol. Res. 18 (1988), pp. 271–317.
[19] C.J.F. ter Braak and P. Smilauer, *Canoco Reference Manual and Canodraw for Windows User's Guide: Software for Canonical Community Ordination (Version 4.5)*, Microcomputer Power, Ithaca, 2002.
[20] C.J.F. ter Braak and P.F.M. Verdonschot, *Canonical correspondence analysis and related multivariate methods in aquatic ecology*, Aquat. Sci. 57 (1995), pp. 255–289.
[21] J. Thioulouse, *Simultaneous analysis of a sequence of paired ecological tables: A comparison of several methods*, Ann. Appl. Stat. (2010), pp. 1–28.
[22] J. Thioulouse and D. Chessel, *Les Analyses Multitableaux En Écologie Factorielle. I. De La Typologie D'état À La Typologie De Fonctionnement Par L'analyse Triadique*, Acta Oecol. Oec. Gen. 8 (1987), pp. 463–480.
[23] J. Thioulouse, D. Chessel, S. Dolédec, and J.M. Olivier, *Ade-4: A multivariate analysis and graphical display software*, Stat. Comput. 7 (1997), pp. 75–83.
[24] J. Thioulouse, M. Simier, and D. Chessel, *Simultaneous analysis of a sequence of paired ecological tables*, Ecology 85 (2004), pp. 272–283.
[25] P. Williams and M.P. Galindo-Villardón, *Canonical non-symmetrical correspondence analysis: An alternative in restricted ordination*, SORT 32 (2008), pp. 93–112.
[26] A.L. Wollenberg, *Redundancy analysis, an alternative for canonical analysis*, Psychometrika 42 (1977), pp. 207–219.

PAPER VI

Diel vertical behavior of Copepoda community (naupliar, copepodites and adults) at the boundary of a temperate estuary and coastal waters

Gonçalves AMM, Pardal MA, Marques SC, Mendes S, Fernández-Gómez MJ, Galindo-Villardón MP, Azeiteiro UM

Estuarine, Coastal and Shelf Science, 2011 (in press) (#)

(#) Additional information:

- First online: available online: November 2011 doi:10.1016/j.ecss.2011.11.018
- Publ. London: Academic Press / Elsevier Science
- ISSN: 0272-7714
- **Indexed in:** Journal Citation Report - Science Citation Index
- **Impact Factors:** IF 2010 (2 years): 1.887; IF 2010 (5 years): 2.369
- **Category:**
  MARINE & FRESHWATER BIOLOGY. Ranking: 31/93 – Q. 2 (Quartile 2)
  OCEANOGRAPHY. Ranking: 19/59 – Q. 2 (Quartile 2)
- **Also indexed in**: BIOBASE, BIOSIS databases/Zoological Records, CAB Internacional, Chemical Abstracts Service, Chemical Abstracts Service, Current Awareness in Biological Sciences, Current Awareness in Biological Sciences, Current Contents ASCA/Engineering Technology & Applied Science/Science Citation Index/SCISEARCH Data, Current Contents/Agriculture, Biology & Environmental Sciences, Current Contents/Physics, Chemical, & Earth Sciences, Engineering Index, Environmental Periodicals Bibliography, Geo Bib & Index, INSPEC Data/Cam Sci Abstr, Marine Literature Review, Meteorological and Geoastrophysical Abstracts, Oceanbase, Oceanographic Literature Review, Research Alert. Scisearch Scopus.
-

Estuarine, Coastal and Shelf Science xxx (2011) 1–15

Contents lists available at SciVerse ScienceDirect

Estuarine, Coastal and Shelf Science

journal homepage: www.elsevier.com/locate/ecss

Diel vertical behavior of Copepoda community (naupliar, copepodites and adults) at the boundary of a temperate estuary and coastal waters

A.M.M. Gonçalves [a,*], M.A. Pardal [a], S.C. Marques [a], S. Mendes [b,c], M.J. Fernández-Gómez [c], M.P. Galindo-Villardón [c], U.M. Azeiteiro [a,d]

[a] CFE - Centre for Functional Ecology, Department of Life Sciences, University of Coimbra, P.O. box 3046, 3001-401 Coimbra, Portugal
[b] GIRM-Marine Resources Research Group, School of Tourism and Maritime Technology, Polytechnic Institute of Leiria, Campus 4, 2520-641 Peniche, Portugal
[c] University of Salamanca, Department of Statistics, 37007 Salamanca, Spain
[d] Universidade Aberta, Rua do Ameal 752, 4200-055 Porto, Portugal

ARTICLE INFO

Article history:
Received 4 April 2011
Accepted 3 November 2011
Available online xxx

Keywords:
Copepoda
life history stages
vertical migration
horizontal movements
temperate estuarine system

ABSTRACT

Despite a growing interest in diel vertical migration as a research topic, there are few studies in southern European marine coastal systems. This study determined the main structuring hydrological and physical factors at different temporal scales in copepod assemblage distribution patterns. Seasonal, tidal, lunar and diel vertical migrations accomplished by horizontal movements were examined on the main copepod fraction of the Mondego estuary, Portugal. Seasonal samples were conducted hourly at the mouth of the estuary, during diel cycles (25 h), both over neap and spring tides, at the bottom and surface, using a 63 μm and 335 μm mesh size nets. Simultaneously, four sites inside the estuary were sampled during flood tide to evaluate and compare copepods species' distribution along the estuary. Species life cycles were also categorized. Spring–spring tide best expresses the stable part of copepod-environment dynamics. *Acartia tonsa* and *Oithona nana* were distributed mainly at the bottom during ebb tides. A clear resident estuarine performance was noticeable in *O. nana* proving the estuarine preferences of the species. Neritic species showed preferences by saline waters, whereas the resident species were found mainly at estuarine areas. Copepodites stages showed a similar distribution pattern as estuarine species, avoiding leaving the estuary. In contrast nauplii and *Oithona plumifera* showed higher densities at surface flood tides. Indeed, vertical migrations accomplished by horizontal movements were mainly influenced by depth and tidal cycles, whereas day and night were not ecologically significant.

1. Introduction

Diel vertical migration accomplished by horizontal transport in invertebrates and fishes has been recognized for several decades but only recently has it been reported over tidal timescales (Forward and Tankersley, 2001). In estuarine and coastal areas the transport by tidal currents is an efficient method for fast horizontal movements, mainly to earlier stages due to their limited swimming abilities. Even so, it also reduces the energy necessary in organisms reported as good swimmers (Metcalfe et al., 1990). Moreover, tidal-stream transport is used mainly during the night when predation by visual predators is greatly reduced, and thus losses of individuals (Zeng and Naylor, 1997). Furthermore, a similar mechanism is associated with the larval releasing activity. Ovigerous females release their larvae during the crepuscular phase which is probably an adaptation to avoid predation by visual feeders (Paula, 1989; Zeng and Naylor, 1997). Indeed, several studies focused on the synchronism between the rhythmic cycles of decapod larval export and the light and lunar phases, as an export strategy of their life cycle (Paula, 1989; Queiroga et al., 1994; 1997; Queiroga, 1996; Gonçalves et al., 2003). Another less studied factor is the migratory behavior of the different life history stages. To date this was mainly reported on fish (e.g. the Atlantic salmon *Salmo salar*, Rommel and McCleave, 1973 and the European ell *Anguilla anguilla*, Berg, 1979) and crabs (e.g. *Carcinus maenas*, Linnaeus 1758 and *Callinectes sapidus*, M. J. Rathbun, 1896) (Forward and Tankersley, 2001). Consequently there is a lack of knowledge on copepods that are undoubtedly one of the most abundant zooplankton groups.

In recent decades, the importance of planktonic copepods in the dynamic of aquatic food webs has been highlighted (McLusky and Elliott, 2004; Damotharan et al., 2010). They play a pivotal

* Corresponding author.
E-mail addresses: ammendes@student.zoo.uc.pt, anamartagoncalves@gmail.com (A.M.M. Gonçalves).

0272-7714/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ecss.2011.11.018

Please cite this article in press as: Gonçalves, A.M.M., et al., Diel vertical behavior of Copepoda community (naupliar, copepodites and adults) at the boundary of a temperate estuary and coastal waters, Estuarine, Coastal and Shelf Science (2011), doi:10.1016/j.ecss.2011.11.018

ecological role in terms of biomass and energy fluxes transference between primary producers and higher trophic levels, indicating the importance of this group in ecological studies (Kršinić et al., 2007; Damotharan et al., 2010).

Several works in coastal, oceanic and estuarine waters have studied the distribution and vertical heterogeneity of zooplankton (Morgado et al., 2003; Marques et al., 2009; Hsieh et al., 2010). It is now assumed that the vertical distribution patterns of copepods throughout the year are a dynamic complex response structured by tidal, diel and lunar cycles, by gradients of environmental variables (e.g. temperature, salinity, hydrostatic pressure) and lastly by the reproductive cycles of predators (e.g. fish) and prey (e.g. phytoplankton) (Forward and Tankersley, 2001). Despite this, studies focusing seasonal responses of copepod communities, and respective life history stages, to changes of environmental factors linked to tidal, lunar and diel cycles at different depths are scarce or lacking in southern European latitudes. This is the first attempt conducted in the Mondego estuary on this small-sized zooplankton fraction. In fact, naupliar stages are the numerical bulk of copepods, whilst smaller copepods normally dominate in terms of abundance, biomass and productivity. However, these small fractions are significantly under-represented in many current and historical data sets, limiting the knowledge of copepod populations (Gallienne and Robins, 2001; Titelman and Fiksen, 2004).

Mechanisms of entrance and retention in nursery areas are the main focus of several studies across North America (Epifanio and Garvine, 2001; Rooper et al., 2006), Japan (Islam et al., 2007), or Europe (Jager, 1999; Jessopp and McAllen, 2008). However little is known of the implications of endogenous rhythms associated with lunar-tide cycle of larval entrance and consequent up-estuary transport and horizontal distribution inside these nursery areas. The present study gathers data from seasonal diel cycles sampled both at neap and spring tides and data from monthly spatial sampling inside the estuary.

The present work furthers knowledge of the distribution and life history stages of the main copepod community. The study of the associated environmental parameters (e.g. lunar phase, diel cycles, depth, tidal currents) will allow the identification of the factors contributing to the vertical migratory behavior promoting retention or displacement inside the estuarine system. Thus, this work aims to study the most dominant marine copepod fraction and the resident estuarine species, and respective life history stages (nauplii, copepodites and adults), in order: 1) to understand the correlation between copepod distribution and the physical properties of water column over seasonal, tidal, lunar and nycthemeral cycles; 2) to evaluate diel vertical migration and horizontal movements in the distribution of each species and respective life stage at the mouth of the estuary; 3) to compare and evaluate the distribution of the copepod species in upstream areas and at the mouth of the Mondego estuary. This gives a first characterization of life cycles of the main copepod species in relation to the distribution of the different life history stages of each species.

2. Materials and methods

2.1. Study area

The Mondego River estuary is a small mesotidal system located in the western Atlantic coast of Portugal (40° 08′ N, 8° 50′ W) (Fig. 1). The hydrological basin of the Mondego, with an area of 6 670 km^2, provides an average freshwater flow rate of 8.5×10^9 m^3 s^{-1} [Author - check this number and units at proof stage – it is probably wrong]. Its terminal part is divided into two arms (north and south), separated by the Murraceira Island at about 7 km, converging again near the mouth. Here the influence of both river flow and neritic waters is strong and the depth is around 6–13 m. Tides in this system are semi-diurnal, and at the inlet the tidal range is 0.35–3.3 m.

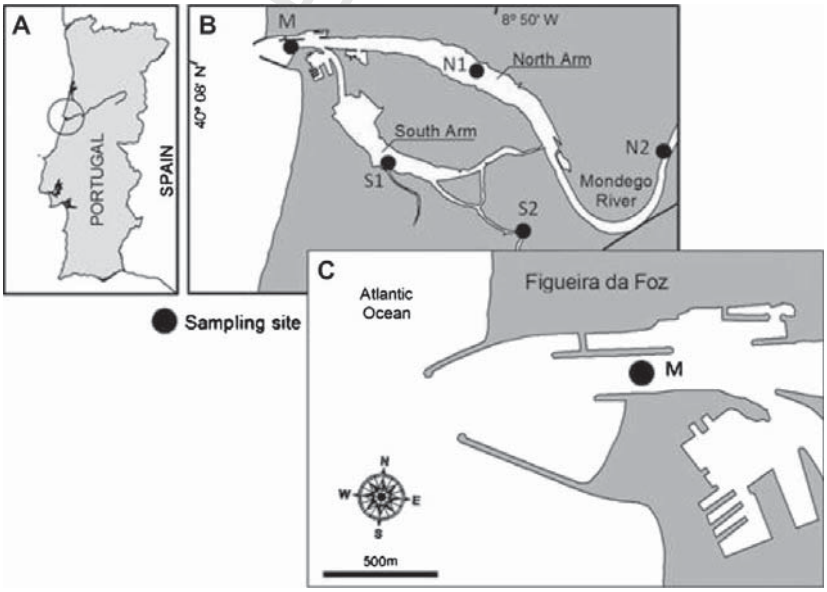


Fig. 1. Location of the Mondego estuary on the western coast of Portugal and the five sampling stations within the estuary: (M – mouth station (sampling over diel cycles – 25 h), N1 and N2 – northern arm stations, S1 and S2 – southern arm stations).

Please cite this article in press as: Gonçalves, A.M.M., et al., Diel vertical behavior of Copepoda community (naupliar, copepodites and adults) at the boundary of a temperate estuary and coastal waters, Estuarine, Coastal and Shelf Science (2011), doi:10.1016/j.ecss.2011.11.018

2.2. Field sampling and laboratory analysis

In the diel cycles (25 h) zooplankton was collected hourly from two depths ranges – sub-surface and near the bottom (6–13 m) at a fixed station located at the mouth of the Mondego estuary (Fig. 1). Samples were classified as day or night, where the day phase was taken from sunrise to sunset, and the night phase, the complementary period. Seasonal sampling took place in June 2005 (summer), September and October 2005 (autumn); December 2005 (winter) and March and April 2006 (spring), over neap and spring tides. Simultaneously, monthly spatial sampling was performed during flood tide inside the estuary at four stations: N1 and N2 in the north arm; S1 and S2 in the south arm (Fig. 1). The samples were collected with a 335 μm mesh Bongo net (diameter: 0.5 m) and a 63 μm mesh net (diameter: 0.30 m) to prevent underestimation of the smaller zooplankton species and the early developmental stages. Both nets were equipped with a Hydro-Bios flow meter in the mouth (to estimate the volume of water filtered by the nets). Samples were fixed and preserved in 4% buffered formaldehyde in seawater and returned to the laboratory for analysis. The contents of each sample were examined and identified to the lowest possible taxonomic level under a stereoscopic microscope. The samples with high numbers of organisms were sub-sampled using a Folsom plankton splitter. In all samples a minimum of 500 individuals were counted and identified, with densities calculated and expressed as individuals/m^3. In parallel to each sampling event, several hydrological parameters were measured *in situ*: water temperature (°C), salinity (S), dissolved oxygen concentration (DO, mg L^{-1}) and pH were recorded with appropriate sensors (WTW) at both depths; transparency (m) was measured with a Secchi disc. Additionally, water samples were collected on each sampling station to determine nutrient concentrations (Strickland and Parsons, 1972 for nitrates and nitrites, mg L^{-1}; Limnologisk Metodik, 1992 for phosphates and ammonia, mg L^{-1}). Chlorophyll *a* concentration (Chl *a*) and suspended particulate matter (SPM) were also determined by filtering 500–1000 mL of water.

2.3. Statistical analysis

From a total of 115 species identified belonging to 18 taxonomic groups, were selected the 5 most regular and abundant copepod species and stages (4 neritic species and a resident estuarine species), plus nauplii of copepod and other copepodites (belong to other copepod species identified), to investigate diel vertical distribution throughout the seasonal cycle. Despite the presence of some copepod species being sampled by both mesh size nets, in accordance with the literature (Antacli et al., 2010; Riccardi, 2010; Tseng et al., 2011) and previous works (Primo et al., 2009; Gonçalves et al., 2010b; Falcão et al., 2011), densities of smaller copepod species such as *Oithona*, copepod nauplii and copepodites were quantified using the 63 μm mesh net, whereas adult copepods of larger species (*Acartia clausi*, *Acartia tonsa* and *Temora longicornis*) were quantified using the 335 μm mesh net to avoid underestimation of species abundance.

Copepods density and environmental factors for each season, lunar phase and tide were combined to generate two series of tables (Fig. 2): one for 12 environmental parameters and the other for 5 species densities (including nauplii, juveniles and adults), plus nauplii and other copepodites (the most common species were selected in order to decrease the number of zero in the analyses). Each pair of tables share the same sampling occurrence (in rows) for the 4 seasons (spring, summer, autumn and winter), 2 lunar conditions (neap and spring), 2 depths (sub-surface and bottom), the diel cycle (daylight and dark period) and 2 tidal conditions (ebb and flood). Therefore, each series of tables (biological and environmental) was composed of 16 matrices. Prior to calculations species density was log ($x + 1$) transformed, to minimize the dominant effect of exceptional values and environmental data were normalized.

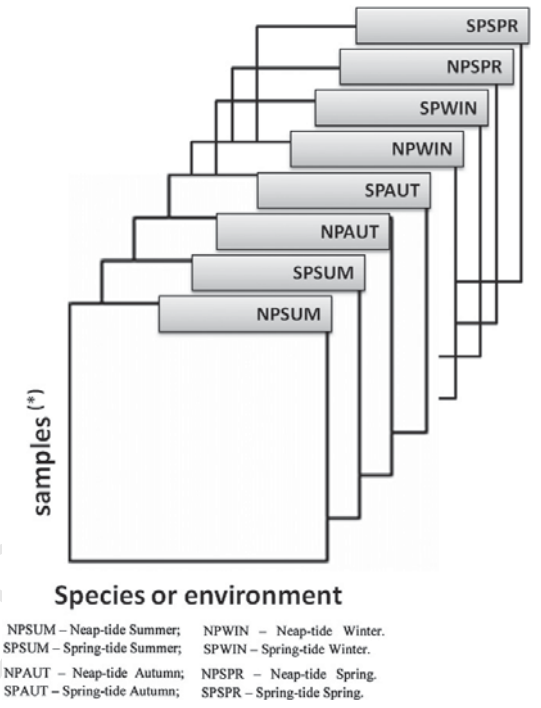


Fig. 2. Three – way data structure: for environmental parameters and species densities at the mouth of the estuary.

The STATICO method (Simier et al., 1999; Thioulouse et al., 2004) was carried out to analyze the two series of tables coupled two by two. In this study the common structure between environmental and species density tables and the stability of this structure over the sampling period were assessed. The samples, which must be the same for both paired tables, but may vary between the pairs, correspond to a diel cycle reflecting the semi-diurnal tidal cycle. The STATICO method proceeds in three stages: (1) the first stage consists in analyzing each table by a one-table method (normalised PCA of the environmental variables and centered PCA of the species data); (2) after that, each pair of tables is linked by the Co-inertia analysis (Dolédec and Chessel, 1994) which provides an average image of the co-structure (species-variables cross-table); (3) Partial Triadic Analysis (Thioulouse and Chessel, 1987) is finally used to analyze this sequence. It is a three step procedure, namely the interstructure, the compromise and the intrastructure (or trajectories) analyses. STATICO also enables plotting the projection of the sampling data in respect to the depth, diel and tidal cycle, of each original table on the compromise axes, in terms of species density and environmental factors (Mendes et al., 2009; Marques et al., 2011). Hence, it is possible to discuss the correlation between species distribution and environmental factors. Calculations and graphs presented were performed using ADE-4 software (Thioulouse et al., 1997). This software is available at http://pbil.univ-lyon1.fr/ADE-4.

Please cite this article in press as: Gonçalves, A.M.M., et al., Diel vertical behavior of Copepoda community (naupliar, copepodites and adults) at the boundary of a temperate estuary and coastal waters, Estuarine, Coastal and Shelf Science (2011), doi:10.1016/j.ecss.2011.11.018

3. Results

3.1. Environmental factors

In Table 1 the hydrological conditions recorded during the study period are listed. The water temperature varied between 20.77 ± 1.47 °C, in summer (neap-surface), and 10.75 ± 0.70 °C, in winter (spring-surface). In general, the higher values of salinity (close to 35) were recorded at all seasons at the bottom at both lunar phases, with exception in summer (Table 1). The lowest values were registered in winter, at the surface (4.96 ± 1.72, neap tide; 9.50 ± 4.61, spring tide). Considering Chl *a*, the concentrations were higher in summer (7.05 ± 6.21 mg m^{-3}) and winter (6.85 ± 2.06 mg m^{-3}), showing slightly differences between lunar and tidal cycle. Concerning nutrients concentrations, phosphates and nitrites showed the lowest values, followed by ammonia, nitrates and silica. Mean silica values were generally higher at the surface at both lunar phases. Suspended particulate matter (SPM) and particulate organic matter (POM) showed slightly seasonal differences, with lower values in winter and spring. pH varied from 6.79 ± 0.98 (in winter at spring tide) to 8.25 ± 0.59 (in spring at spring tide). Considering dissolved oxygen (DO), values ranged from 7.96 ± 0.32 mg L^{-1} (in autumn) to 10.65 ± 0.52 mg L^{-1} (in winter).

3.2. Copepod assemblages' distribution and dynamics over seasonal, lunar, tidal and day–night cycles

Copepod assemblages were dominated by the marine cyclopoid species (*Oithona plumifera* Baird, 1843), the estuarine cyclopoid species (*Oithona nana* Giesbrecht, 1892 and), the marine calanoid species (*A. clausi* Giesbrecht, 1889 and *T. longicornis* O. F. Müller, 1785), the estuarine calanoid species (*A. tonsa* Dana, 1849), and copepod nauplii. A high number of copepodites from other species were also present.

All species showed the highest densities in the bottom and almost all, with the exception of *A. tonsa* (adults and juveniles) and *O. nana* (adults and juveniles), occurred mainly at neap tide (Fig. 3). Moreover, at this lunar phase, nauplii of copepods, *A. clausi* (juveniles and adults), copepodites of *T. longicornis* and other copepodites showed higher densities in summer at flood tide during night (FN) (Fig. 3A, C, E, K and W). In winter and spring, at both lunar phases, these species showed lower densities or did not occur (Fig. 3A–F, K, L, W and X). *O. plumifera* (adults and juveniles) and *T. longicornis* showed their peaks of abundance in autumn (neap tide), at ebb tide, with the adults of *O. plumifera* occurring mainly during the day (ED), while the juveniles of *O. plumifera* and *T. longicornis* appeared at higher densities during night (Fig. 3G, I and M). Considering the estuarine species *A. tonsa* (adults and juveniles) and *O. nana* (adults and juveniles) their occurrence in neap tide was very low or null, showing the highest densities in spring tide (Fig. 3O–RV). Adults of *A. tonsa* only occurred in summer, with their peak of density at flood tide during night (FN) while the juveniles were observed in winter mainly at ebb tide (ED and EN) (Fig. 3T and V). *O. nana* (adults and juveniles) were presented in autumn (mainly the adults) and winter showing higher densities in winter at ebb tide during the night (EN) (Fig. 3P and R).

3.3. Relationships between environmental factors and copepod assemblages

The factor plots of the first two axes of the compromise analysis are shown for the species assemblages (Fig. 4A – see Table 2 for species codes) and for the environmental variables (Fig. 4B).

Table 1
Environmental parameters (mean ± SD) measured during the studied period, at spring-neap tide and at surface-bottom deepness.

Season	Lunar phase	Depth	NH_4 (mg L^{-1})	PO_4 (mg L^{-1})	Si (mg L^{-1})	NO_2 (mg L^{-1})	NO_3 (mg L^{-1})	Chl *a* (mg L^{-1})	pH	O_2 (mg L^{-1})	Temp (°C)	Sal (‰)	SPM (g L^{-1})	POM (g L^{-1})
Summer	Neap	S	0.03 ± 0.02	0.01 ± 0.01	0.27 ± 0.13	0.00 ± 0.00	0.04 ± 0.03	7.05 ± 6.21	8.10 ± 0.18	9.47 ± 0.70	20.77 ± 1.47	29.87 ± 3.05	0.05 ± 0.03	0.01 ± 0.01
		B	0.04 ± 0.02	0.01 ± 0.00	0.14 ± 0.07	0.00 ± 0.00	0.02 ± 0.03	4.25 ± 2.75	8.03 ± 0.19	9.60 ± 1.03	18.64 ± 0.80	33.37 ± 0.61	0.07 ± 0.05	0.01 ± 0.01
	Spring	S	0.05 ± 0.02	0.03 ± 0.01	0.24 ± 0.12	0.01 ± 0.01	0.03 ± 0.03	3.59 ± 1.77	8.14 ± 0.08	9.14 ± 0.78	18.99 ± 0.85	26.08 ± 7.12	0.03 ± 0.01	0.01 ± 0.00
		B	0.05 ± 0.02	0.03 ± 0.01	0.24 ± 0.14	0.01 ± 0.00	0.03 ± 0.03	3.99 ± 1.03	8.18 ± 0.07	9.12 ± 0.48	18.28 ± 0.65	26.58 ± 8.08	0.04 ± 0.01	0.01 ± 0.00
	Autumn	Neap	S	0.06 ± 0.02	0.05 ± 0.01	0.61 ± 0.18	0.04 ± 0.01	0.22 ± 0.09	2.50 ± 1.52	7.72 ± 0.32	7.96 ± 0.32	17.85 ± 0.94	28.81 ± 3.24	0.03 ± 0.01	0.01 ± 0.00
			B	0.03 ± 0.01	0.03 ± 0.01	0.28 ± 0.07	0.03 ± 0.01	0.08 ± 0.03	2.84 ± 0.99	7.84 ± 0.27	8.77 ± 0.96	15.98 ± 0.39	35.19 ± 0.64	0.03 ± 0.01	0.01 ± 0.00
		Spring	S	0.03 ± 0.02	0.03 ± 0.01	0.44 ± 0.18	0.00 ± 0.00	0.26 ± 0.15	3.79 ± 2.40	7.95 ± 0.40	9.12 ± 0.84	15.71 ± 0.74	32.05 ± 4.06	0.03 ± 0.01	0.01 ± 0.00
			B	0.02 ± 0.02	0.03 ± 0.01	0.30 ± 0.11	0.00 ± 0.00	0.16 ± 0.08	5.56 ± 4.13	7.86 ± 0.23	9.65 ± 0.70	15.35 ± 0.37	34.64 ± 1.93	0.03 ± 0.00	0.01 ± 0.00
	Winter	Neap	S	0.10 ± 0.04	0.03 ± 0.01	1.63 ± 0.55	0.01 ± 0.00	0.93 ± 0.23	6.85 ± 2.06	7.38 ± 0.38	9.81 ± 0.53	11.97 ± 0.37	4.96 ± 1.72	0.01 ± 0.00	0.00 ± 0.00
			B	0.04 ± 0.02	0.01 ± 0.01	0.42 ± 0.32	0.01 ± 0.00	0.21 ± 0.17	3.68 ± 1.40	7.37 ± 0.43	9.04 ± 0.82	13.64 ± 0.52	31.62 ± 6.33	0.03 ± 0.01	0.00 ± 0.00
		Spring	S	0.11 ± 0.06	0.02 ± 0.01	1.53 ± 0.20	0.01 ± 0.00	0.89 ± 0.22	4.62 ± 2.07	6.79 ± 0.98	10.65 ± 0.52	10.75 ± 0.70	9.50 ± 4.61	0.01 ± 0.00	0.00 ± 0.00
			B	0.04 ± 0.01	0.01 ± 0.01	0.33 ± 0.13	0.01 ± 0.00	0.22 ± 0.07	3.79 ± 3.75	6.86 ± 0.59	10.08 ± 0.47	12.90 ± 0.39	34.19 ± 1.61	0.03 ± 0.00	0.00 ± 0.00
	Spring	Neap	S	0.91 ± 0.24	0.04 ± 0.01	2.13 ± 0.43	0.02 ± 0.01	1.25 ± 0.31	3.54 ± 0.73	7.38 ± 0.35	9.96 ± 0.76	12.43 ± 0.41	20.93 ± 13.26	0.01 ± 0.01	0.00 ± 0.00
			B	0.07 ± 0.08	0.02 ± 0.01	0.53 ± 0.16	0.01 ± 0.00	0.37 ± 0.14	3.15 ± 1.26	7.49 ± 0.26	9.34 ± 0.40	12.38 ± 0.74	17.20 ± 13.37	0.04 ± 0.02	0.01 ± 0.00
		Spring	S	0.08 ± 0.03	0.03 ± 0.00	1.62 ± 0.42	0.01 ± 0.00	0.69 ± 0.27	2.63 ± 1.11	8.25 ± 0.59	8.64 ± 0.58	15.38 ± 1.64	10.35 ± 5.20	0.01 ± 0.00	0.00 ± 0.00
			B	0.02 ± 0.02	0.01 ± 0.01	0.64 ± 0.34	0.01 ± 0.01	0.31 ± 0.17	1.97 ± 0.64	8.21 ± 0.29	9.21 ± 0.91	15.34 ± 0.65	29.15 ± 4.82	0.02 ± 0.00	0.00 ± 0.00

Please cite this article in press as: Gonçalves, A.M.M., et al., Diel vertical behavior of Copepoda community (naupliar, copepodites and adults) at the boundary of a temperate estuary and coastal waters, Estuarine, Coastal and Shelf Science (2011), doi:10.1016/j.ecss.2011.11.018

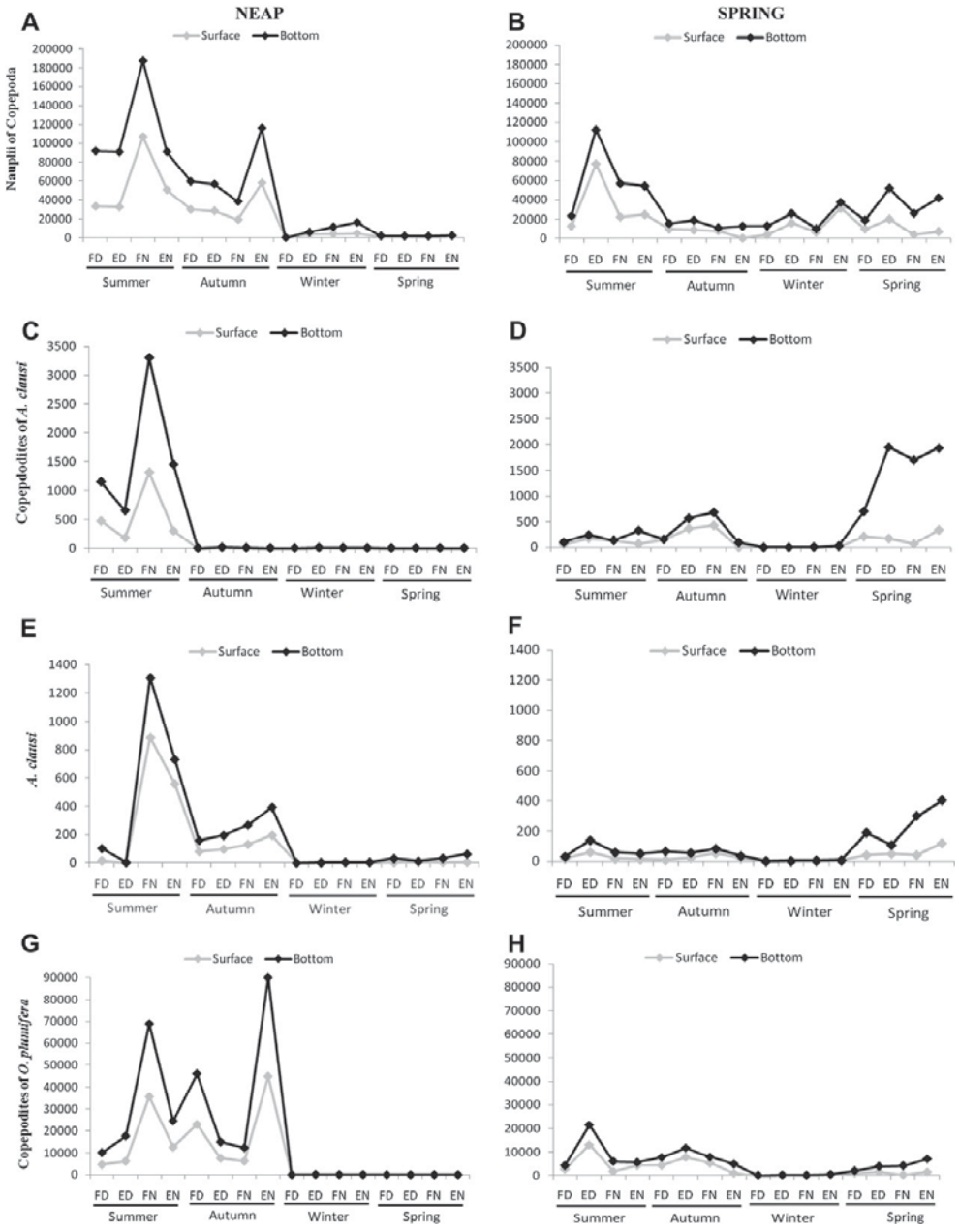


Fig. 3. Seasonal cycle of the dominant marine copepod and the resident estuarine species' densities, and respective life history stages, over neap-spring tides, for each diel and tidal cycles (FD – Flood day; ED – Ebb day; FN – Flood night; EN – Ebb night).

The eigenvalues diagram (Fig. 4C) shows that the first axis was clearly dominant (accounted for 93% of the explained variance) in contrast with the second axis (3% of the explained variance). From the species compromise analysis (Fig. 4A), three groups, belonging to two species, were correlated with axis 1: from left to right copepodites of *T. longicornis*, *T. longicornis* and copepodites of *O. nana*. All the other species adult stages and copepodites showed an intermediate position between axis 1 and 2, but may be attached to the horizontal axis. The axis 2 was characterized by *O. nana*, copepodites and adults of *A. tonsa*, nauplii of Copepoda and

Please cite this article in press as: Gonçalves, A.M.M., et al., Diel vertical behavior of Copepoda community (naupliar, copepodites and adults) at the boundary of a temperate estuary and coastal waters, Estuarine, Coastal and Shelf Science (2011), doi:10.1016/j.ecss.2011.11.018

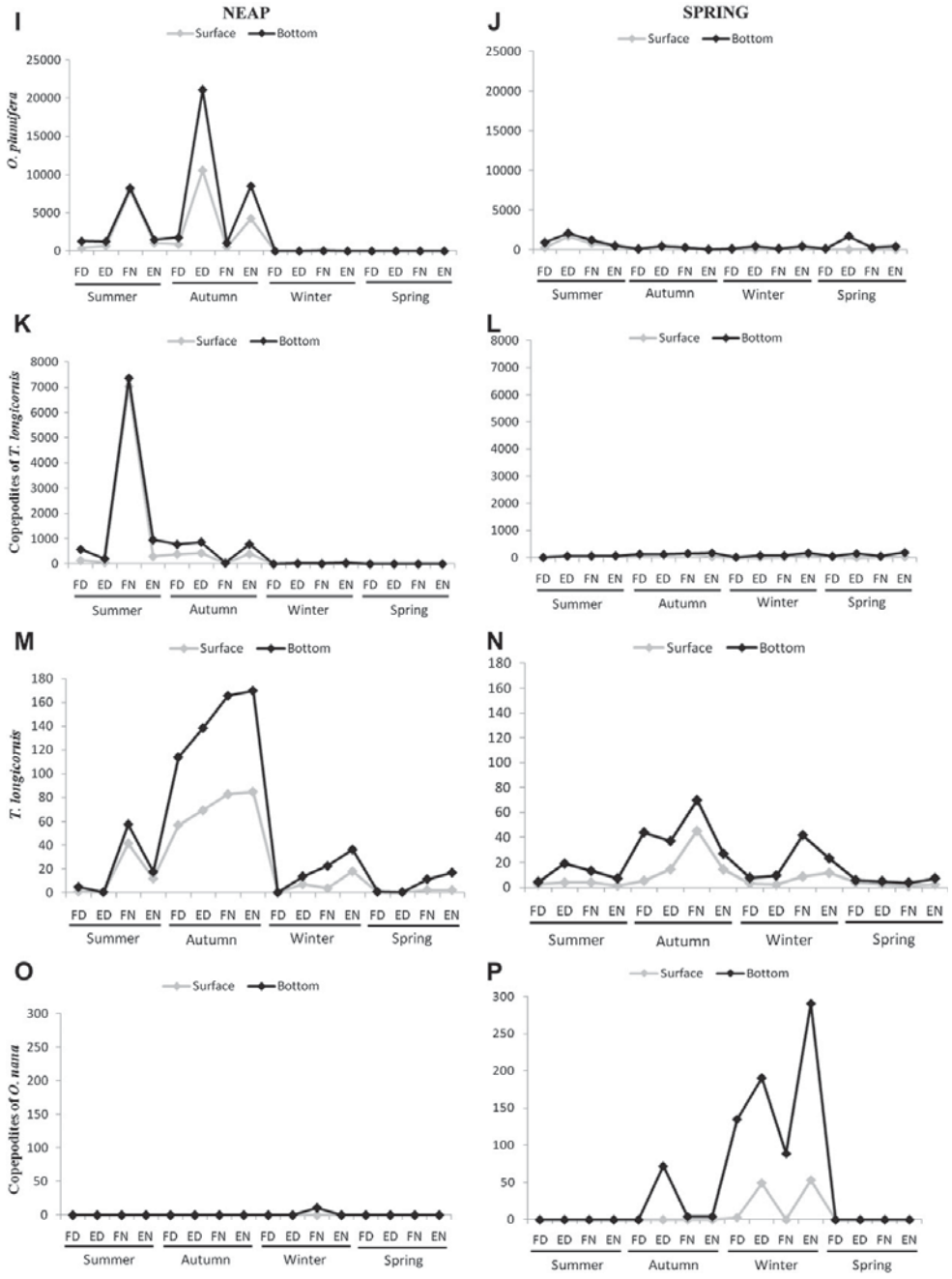


Fig. 3. *Continued*

Please cite this article in press as: Gonçalves, A.M.M., et al., Diel vertical behavior of Copepoda community (naupliar, copepodites and adults) at the boundary of a temperate estuary and coastal waters, Estuarine, Coastal and Shelf Science (2011), doi:10.1016/j.ecss.2011.11.018

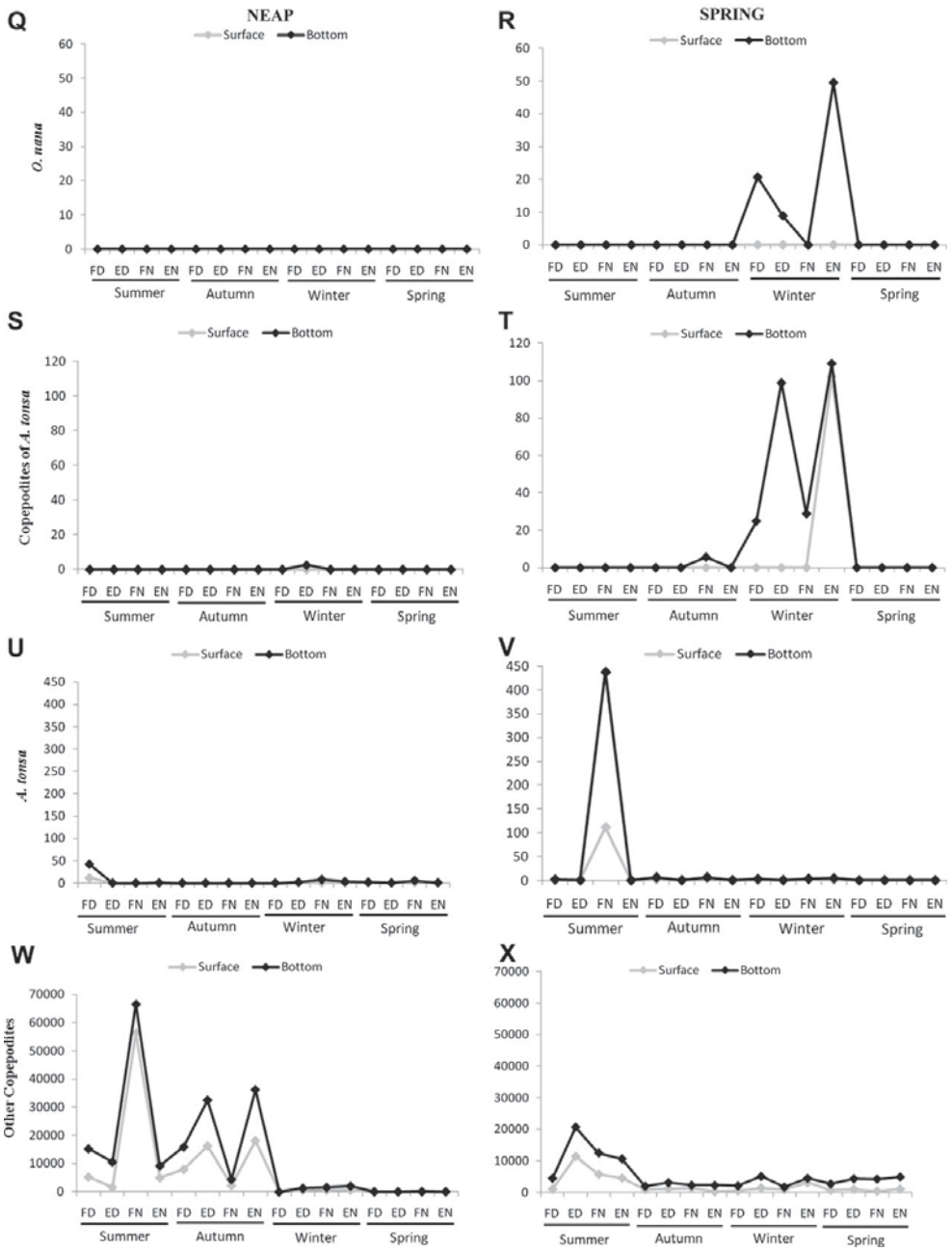


Fig. 3. *Continued*

copepodites of *O. plumifera*. From the environmental compromise analysis (Fig. 4B), the horizontal axis expresses an opposition between POM, salinity and SPM (in the left) and Si, NH_4 and PO_4 (in the right). The axis 2 was characterized by pH. NO_3, and chlorophyll *a* (Chl *a*), have an intermediate position between axis 1 and 2. However, NO_3 can be associated to the horizontal axis. Temperature, NO_2 and dissolved oxygen (O_2) presented short arrows meaning that they have an average performance in all sampling

Please cite this article in press as: Gonçalves, A.M.M., et al., Diel vertical behavior of Copepoda community (naupliar, copepodites and adults) at the boundary of a temperate estuary and coastal waters, Estuarine, Coastal and Shelf Science (2011), doi:10.1016/j.ecss.2011.11.018

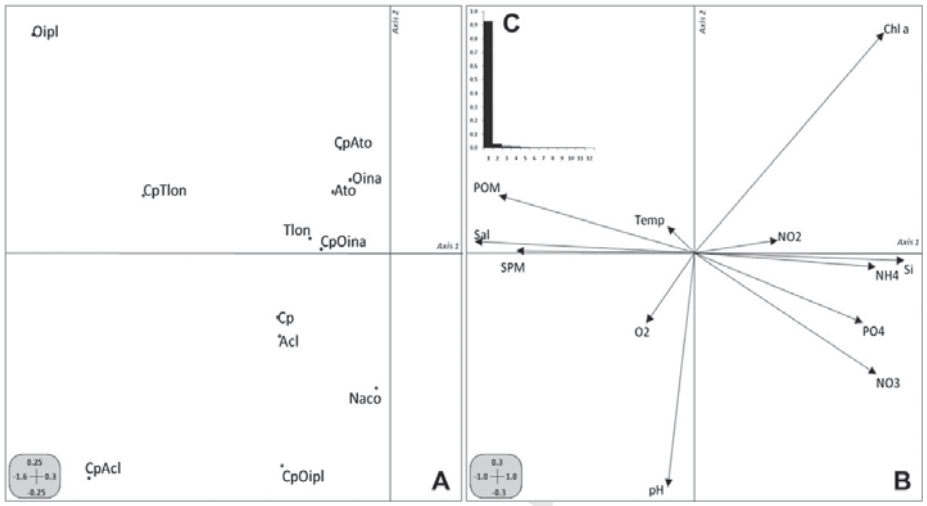


Fig. 4. Compromise factor map of the STATICO analysis of the copepod species variables (A) and environmental parameters (B). (C) Eigenvalues diagram. The scales for axes are given in the boxes. Temp – water temperature, Sal – salinity, O_2- dissolved oxygen, PO_4 – phosphates, NO_2 – nitrites, NO_3 – nitrates, Si – silica, NH_4 – ammonia, Chl *a* – chlorophyll *a*, SPM – suspended particulate matter, POM – particulate organic matter. See Table 2 for species codes.

period. So, they do not play a major role in the average species-environment relationship. Furthermore, Chl *a* concentration is opposed to pH sand O_2.

The most abundant species in the samples were mainly associated with the same environmental variables: salinity (Sal), SPM and POM (Fig. 4A and B). *O. plumifera* and copepodites of *A. clausi* and *T. longicornis* were more abundant at higher concentrations of Sal, SPM and POM (and when the nutrients concentrations NO_2, Si, NH_4, PO_4 and NO_3 were low). Copepodites of *O. plumifera*, *A. clausi* and copepod nauplii were clearly influenced by high values of pH and O_2 and lower values of Chl *a* concentration. Copepodites and adults of *A. tonsa*, copepodites and adults of *O. nana* and *T. longicornis* were influenced by warmer waters. Moreover, *A. tonsa* (adults and juveniles) and *O. nana* (adults and juveniles) were the species with lowest densities while *O. plumifera* (adults and juveniles), *A. clausi* (adults and juveniles), *T. longicornis* (adults and juveniles) and other copepodites showed the highest densities (Fig. 4A and B).

Table 2
Code abbreviations of species (including life stages) used in multivariate analysis (STATICO).

Taxa		Abbreviation	Habitat
	Nauplii Copepoda	Naco	
	Other Copepodites	Cp	
Calanoida			
Acartiidae	Copepodite *Acartia clausi*	CpAcl	Marine
	A. clausi	Acl	Marine
	Copepodite *Acartia tonsa*	CpAto	Estuarine
	A. tonsa	Ato	Estuarine
Temoridae	Copepodite *Temora longicornis*	CpTlon	Marine
	T. longicornis	Tlon	Marine
Cyclopoida			
Oithonidae	Copepodite *Oithona nana*	CpOina	Estuarine
	O. nana	Oina	Estuarine
	Copepodite *Oithona plumifera*	CpOipl	Marine
	O. plumifera	Oipl	Marine

3.4. Dynamics of copepod species

Fig. 5 shows the environmental variables and species (see Table 2 for species codes) of the original tables on the compromise axes. The graphics were separated according to seasons and lunar phases. The stable part of the species-environment dynamics revealed by the compromise analysis, was evident in spring at spring tide (SPSPR). Winter at neap (NPWIN) and spring (SPWIN) tides followed the same tendency. To highlight the lack of representativeness in the spring neap tide (NPSPR), most species rise very close to the origin axes, representing the lowest densities. Considering SPSPR, and similarly to the pattern found in the compromise analysis, the estuarine cyclopoid *O. nana* (adults and copepodites) and the estuarine calanoid *A. tonsa* (adults and copepodites) showed the lowest densities. Additionally, the species-environmental relationship was very similar to the representation of the compromise analysis. Indeed, at this season (SPSPR), *A. clausi* (adults and juveniles), *O. plumifera* (adults and juveniles), juveniles of *T. longicornis* and other copepodites were influenced by saline waters, higher values of POM and SPM. Furthermore, copepodites of *A. clausi* and of *O. plumifera* were positively correlated with higher values of pH while *O. plumifera* and copepodites of *T. longicornis* were influenced by lower values of pH (Fig. 5A and B). A similar trend was observed in *O. plumifera* (copepodites and adults) in winter at neap tide (NPWIN). This cyclopoid showed a positive correlation with saline and warmer waters and higher values of SPM and POM. In this season case (NPWIN), nauplii of copepods was positively correlated with pH and negatively correlated with all other environmental parameters. Despite this, at the spring tide (SPWIN) nauplii of copepods were influenced by colder waters and lower values of salinity, SPM and POM, similarly to copepodites of *O. plumifera*. On the other hand, copepodites of *T. longicornis*, copepodites of *A. tonsa* and *O. nana* (adults and juveniles) were influenced by warmer and saline waters, higher values of SPM and POM.

The projection of the samples of the original tables on the compromise axes, in terms of both environmental parameters and

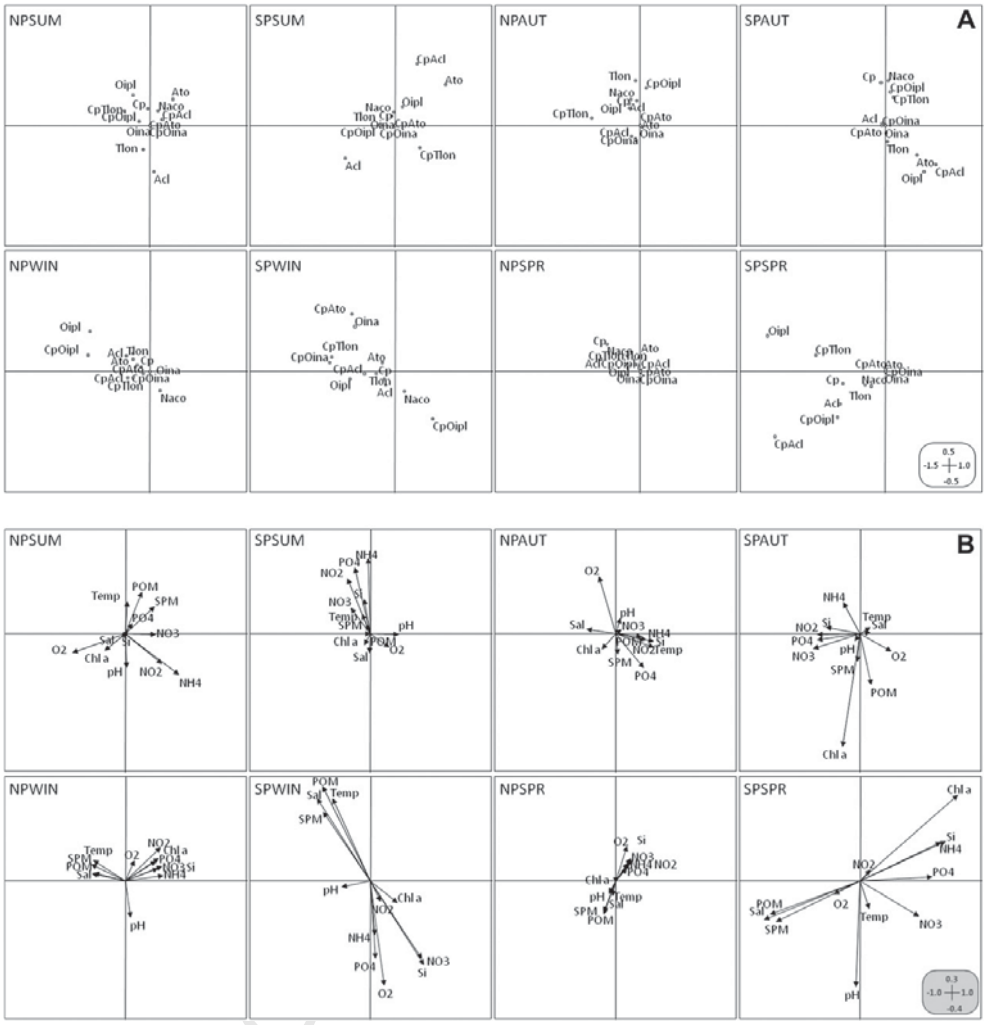


Fig. 5. Trajectories factor maps of the STATICO analysis: seasonal projections (A) of the average position of copepod species and (B) of the environmental variables for each lunar phase (neap and spring tides) at the mouth of the estuary. The scales for axes are given in the boxes. For species codes see Table 2. Environmental variables codes as in Fig. 4.

species structure (Fig. 6) shows the samples grouped in the central part of the axes. Exceptions are at neap tide for summer and at spring tide for winter and spring. Spring - spring tide (SPSPR) was the sample that expresses the best consensus (short arrows) between environmental and species structure (which confirms that previously described, since this is the situation that best represents the compromise image). An opposition (separated by axis 2) between surface-ebb tide (right) and bottom-ebb tide (left) (SE and BE, respectively) was observed. Despite this, there was no clear distinction between day and night. Additionally, the samples surface-flood tide (SF) and bottom-flood tide (BF) were mainly projected in the upper horizontal axis. Therefore, BE samples were mainly associated to the pattern representing the correlation between POM, salinity and SPM with *A. clausi* (adults and juveniles), *O. plumifera* (adults and juveniles), copepodites of *T. longicornis* and other copepodites (as previously described; see Fig. 5A and B). Moreover, the SE samples were characterized by a positive correlation with Si, NH_4, PO_4, NO_3 and low densities of adults and juveniles of *A. clausi* (and *O. plumifera*, copepodites of *T. longicornis* and other copepodites. In summer (neap tide) and autumn (spring tide), day and night samples were spatially opposed. In summer (NPSUM), night was mainly projected at the left side and in autumn (SPAUT) night was projected mainly at the right side. This means that the diel cycle was also an important factor structuring species assemblages. In all other samples, this contrast was present but not so prominent. In winter spring tide (SPWIN), BE and BF samples were mainly projected on the second quadrant, meaning that these samples were essentially associated to juveniles of *A. tonsa*, *O. nana* (adults and juveniles), juveniles of *T. longicornis* and associated with the environmental parameters: salinity, POM, Temp and SPM (see Fig. 5A and B). The SF (surface flood tide) samples were mainly projected on the right side of axis 1 being characterized by high densities of nauplii and copepodites of *O. plumifera* (see Fig. 5A and B).

Please cite this article in press as: Gonçalves, A.M.M., et al., Diel vertical behavior of Copepoda community (naupliar, copepodites and adults) at the boundary of a temperate estuary and coastal waters, Estuarine, Coastal and Shelf Science (2011), doi:10.1016/j.ecss.2011.11.018

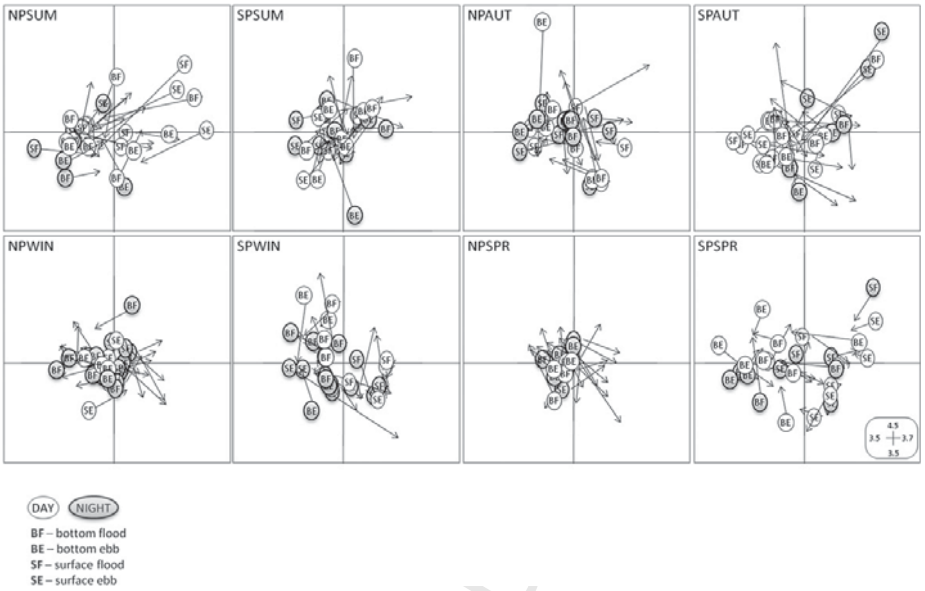


Fig. 6. Trajectories factor plots of the STATICO analysis: projection of the samples in response to two depths ranges (surface and bottom), tidal and diel cycles on the first factorial plan of the compromise analysis, for each season and neap-spring tides. Each sample is represented by two points: one is the projection of the row of the species table (circle: origin of arrows), and the other is the projection of the row of the environmental table (end of arrows). The length of the connecting line reveals the disagreement or the consensus between the two profiles (species–environment), i.e., the length of the line is proportional to the divergence between the datasets. When the datasets agree very strongly, the arrows will be short. Likewise, a long arrow demonstrates a locally weak relationship between the environment and copepod features for that case. The scales for axes are given in the boxes.

3.5. Seasonal, lunar and spatial patterns of copepods distribution

The distribution inside the estuary showed that the studied species occurred mainly at downstream stations of both arms of the estuary (St S1 and St N1), with the exception of the resident *A. tonsa* that occurred at upstream stations, more precisely at the south arm station near the Pranto River (St S2) (Fig. 7). A similar pattern was observed in *O. nana* (Fig. 7Q–S). This species revealed an estuarine behavior with the copepodites occurring along the estuary and the adults clearly distributed in the inner stations, mainly at upstream St N2, during autumn. Nauplii of copepods were observed at neap tide, in autumn and spring, at downstream stations (St S1 and N1) and at St N2, whereas at spring tide naupliar stages were predominant in winter at St S2 (Fig. 7A and B). A similar distribution was observed in juveniles of *A. tonsa*, at spring tide (Fig. 7M). Adults and juveniles of this species were observed in autumn, at St S2 in neap tide (Fig. 7N and P).The highest densities of *A. clausi* (adults) were present in autumn and spring, at the downstream stations, while juveniles were dominant in summer and winter at St S1 (Fig. 7C–F). Juveniles and adults of *T. longicornis* were typically abundant in summer, in the downstream stations (St N1 and St S1), at both lunar phases (Fig. 7G–J). Moreover, adults of *T. longicornis* were also abundant in autumn, at the same sampling stations and lunar phases described above (Fig. 7I and J).

4. Discussion

Estuarine systems are unstable habitats characterized by large scale seasonal fluctuations and small scale variability (e.g. tidal, diel) of the biological and environmental factors. Previous works in the study area examine the relationship between environmental factors and small-sized zooplankton distribution (Vieira et al., 2003; Gonçalves et al., 2010a, b). These first studies with the small-sized zooplankton fractions proved that the studies with the mesozooplankton fractions (200 and 335 μm mesh net sizes) (Marques et al., 2007a,b, 2008; 2009; Primo et al., 2009; Falcão et al., 2011) either under-sampled or missed the overall contribution of this small-sized component, a fact already pointed out by several studies (Gallienne and Robins, 2001; Satapoomin et al., 2004; Hopcroft et al., 2005; Williams and Muxagata, 2006). Moreover, this is the first study encompassing species' life cycles distribution, including smaller specimens, with seasonal, lunar and diel cycles.

All of the main copepod species identified in the Mondego estuary occurred mainly at neap tide, with the exception of winter that presented higher densities at the spring tide, mainly related to *A. tonsa* and *O. nana*. The pattern of densities observed in this study shows a clear dependence between the structure of copepod community and the environmental factors in spring, whereas the vertical migration patterns show a closely dependence of the physical parameters: depth and tide currents. Moreover, the light cycle was also preponderant at summer samples. Similarly, Kennish (1990) stated that the structure of zooplankton community along the estuary depends of both neap-spring and ebb-flood cycles. Several works (Paula, 1989; Queiroga et al., 1994, 1997; Queiroga, 1996; Gonçalves et al., 2003) performed in Portuguese estuaries (Mondego estuary, Mira estuary and Canal de Mira) reported the synchronism between the larval-releasing activity of several decapod species with neap high tide during night-time, rather than tidal amplitude. The semi-lunar rhythm of larval released during nocturnal maximum amplitude tides suggests an export strategy in species' life cycles using strong ebb currents to a rapid export to the ocean (Zeng and Naylor, 1997; Papadopoulos et al., 2002). This synchronised behavior between light intensity, and lunar phase,

Please cite this article in press as: Gonçalves, A.M.M., et al., Diel vertical behavior of Copepoda community (naupliar, copepodites and adults) at the boundary of a temperate estuary and coastal waters, Estuarine, Coastal and Shelf Science (2011), doi:10.1016/j.ecss.2011.11.018

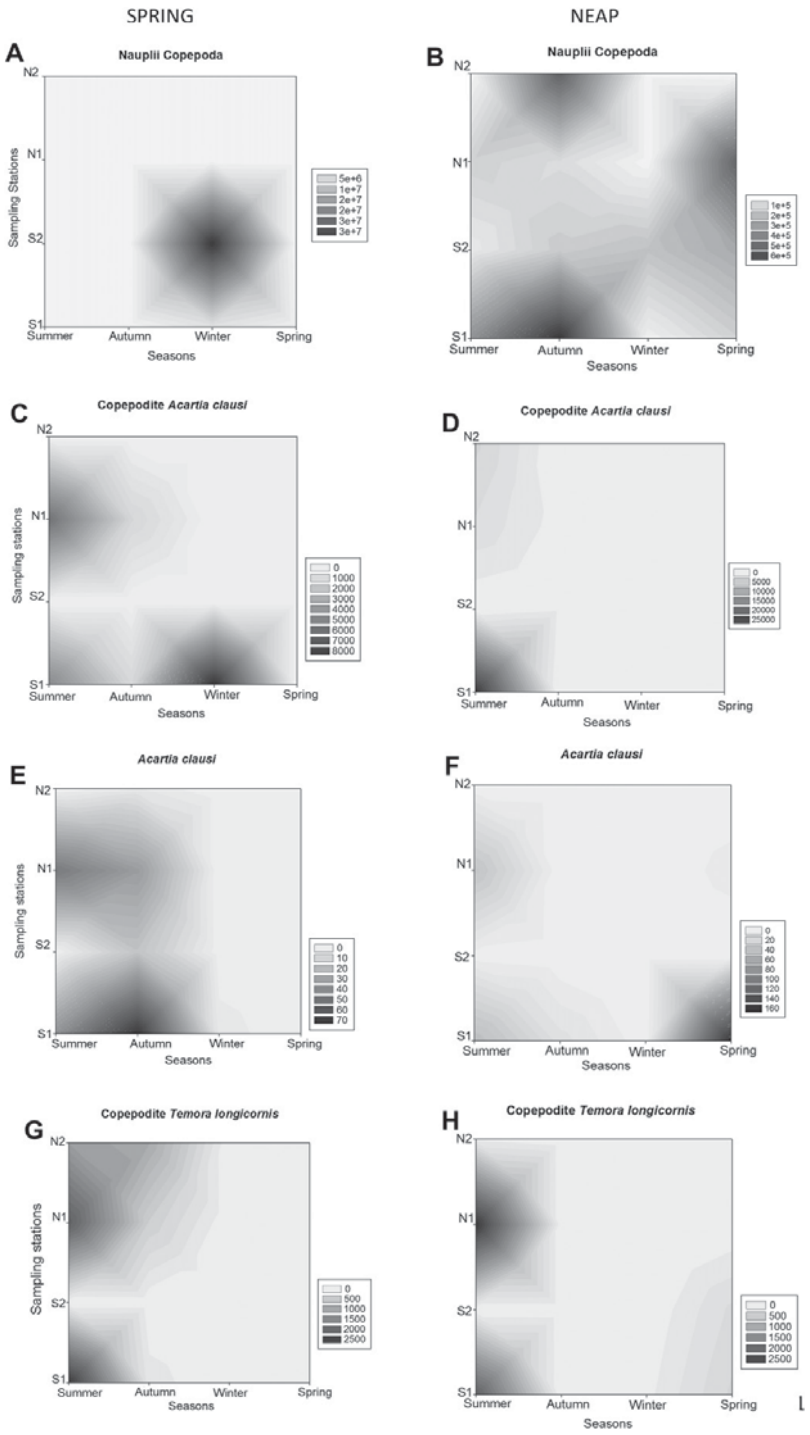


Fig. 7. Seasonal and spatial densities of the copepod species, and their respective life stages, over neap-spring tides, in Mondego estuary.

Please cite this article in press as: Gonçalves, A.M.M., et al., Diel vertical behavior of Copepoda community (naupliar, copepodites and adults) at the boundary of a temperate estuary and coastal waters, Estuarine, Coastal and Shelf Science (2011), doi:10.1016/j.ecss.2011.11.018

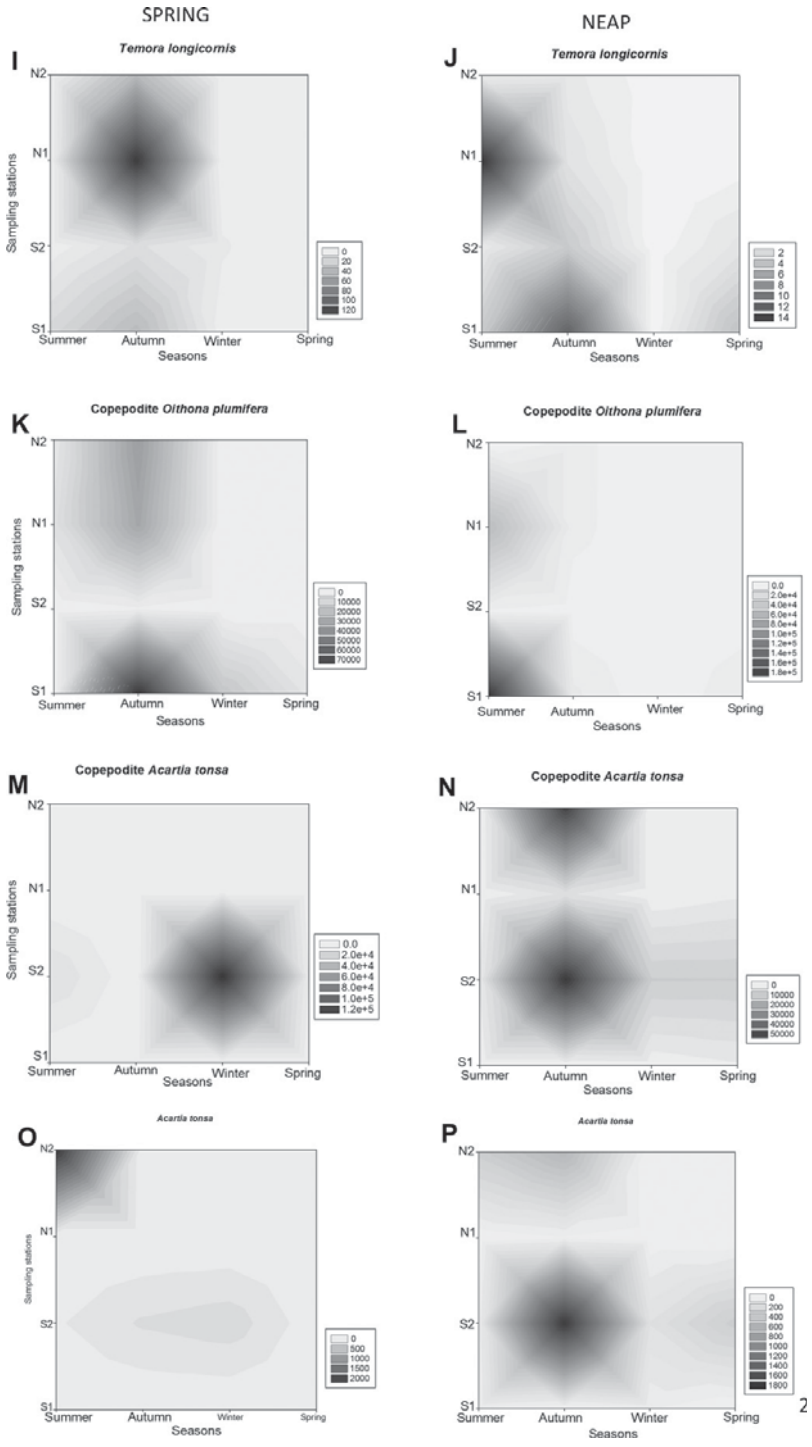


Fig. 7. *Continued*

Please cite this article in press as: Gonçalves, A.M.M., et al., Diel vertical behavior of Copepoda community (naupliar, copepodites and adults) at the boundary of a temperate estuary and coastal waters, Estuarine, Coastal and Shelf Science (2011), doi:10.1016/j.ecss.2011.11.018

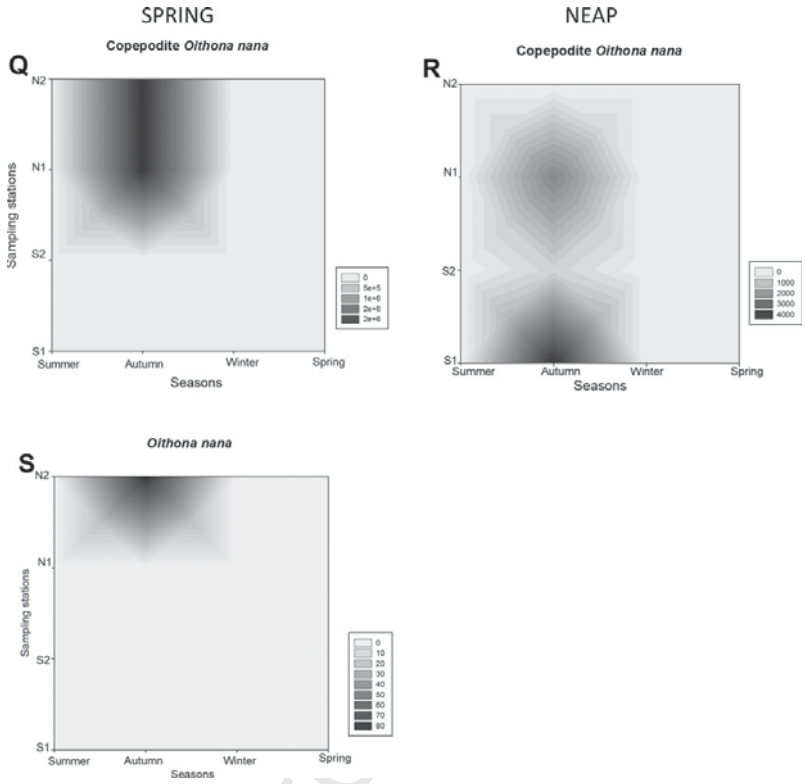


Fig. 7. *Continued*

minimizes visual predation, hence reducing vulnerability to predators and consequently individual losses. Indeed, in the present work spring tide samples were performed at full moon, corresponding to less dark nights, where the lower densities of copepod species were found. Finally, higher densities of copepodites stages were found mostly near the bottom. This pattern may be explained by: 1) the ability of the organisms to adjust their diel and tidal rythms in order to avoid the surface, mainly during the day; and 2) the development of a mechanism to avoid physical stressors such as the tidal currents that transport them outside the system and also to control their horizontal transport, maintaining themselves inside the estuary (Hill, 1991; Morgado et al., 2003; McLusky and Elliott, 2004). An opposite trend was observed to copepod nauplii and copepodites of *O. plumifera* that presented higher densities at surface flood tides. Our findings suggest that nauplii and juveniles of *O. plumifera* do not remain inside the estuary, being carried out to coastal areas by flood tides. In fact, small sized copepods such as *Oithona* and *Oncaea* are referred as the mainly prey of larger predators, including important commercial fishes (e.g. sardines and anchovies larvae stages), having an important role in the dynamics of marine food webs (Porri et al., 2007; Takahashi and Uchiyama, 2008).

Titelman and Fiksen (2004) claimed that nauplii of most species live near the surface and suggested that the distribution of small copepods, especially the nauplii and small species such as *Oithona* spp., differ from the better studied large species and later life history stages. Nauplii and small oithonids are less sensitive to visual predators and more sensitive to invertebrate predators such as larger copepods. Generally, both calanoid and cyclopoid nauplii were found near the surface, while copepodites stayed deeper in the water column (Titelman and Fiksen, 2004).

At the Mondego estuary, neritic species showed preferences by saline waters, distributing mainly at downstream stations and at the mouth of the estuary, whereas resident species were mainly at upstream stations. In this way, earlier stages were transported to downstream stations and to the mouth of the estuary, whereas adults returned to upstream stations, the area of adult populations. *O. nana*, *A. tonsa* and other copepodites showed a similar behavior presenting a clear estuarine distribution. Our findings showed new behavioral aspects of this Oithonidae species and about the copepodites stages that had not yet been explored. In previous works this species (as other smaller specimens and earlier developmental stages) were underestimated due to the use of larger mesh size nets (Azeiteiro et al., 2000; Marques et al., 2009; Primo et al., 2009). Our findings demonstrated that *O. nana* remains in the estuary, reproducing and performing the development stages inside the estuarine system. Copepodites of *O. nana* showed a wide distribution along the estuary, mainly at downstream stations, whereas the adults prefer brackish waters, occurring at upstream stations. Indeed, some authors (Williams and Muxagata, 2006; Porri et al., 2007) suggest the well-defined spatial distribution of juveniles and adults of *O. nana* to be related to respiratory and feeding rates patterns added to a wide tolerance to salinity and temperature of these stages. Considering *A. tonsa*, this is an important resident estuarine species in Mondego estuary, being a common and dominant species of the upper reaches of the estuary (Marques

Please cite this article in press as: Gonçalves, A.M.M., et al., Diel vertical behavior of Copepoda community (naupliar, copepodites and adults) at the boundary of a temperate estuary and coastal waters, Estuarine, Coastal and Shelf Science (2011), doi:10.1016/j.ecss.2011.11.018

et al., 2008; Gonçalves et al., 2010b; Falcão et al., 2011). The ecology of this species is well documented by several authors being considered an important and abundant taxon in European estuaries (Cervetto et al., 1999; Tackx et al., 2004; David et al., 2007).

5. Conclusions

There is now a growing interest on both the numerical importance of cyclopoid copepods, particularly the genus *Oithona*, and also their relative importance compared with calanoid copepods. This study clarifies the ecological significance of this small-sized cyclopoid fraction, namely the numerically significant presence of the estuarine *O. nana*, within the upper estuary, with a progressive decrease in abundance toward the mouth of the estuary, complementing the published mesozooplankton studies on the Mondego estuary. This was an already verified distributional pattern for the species in Southampton Water (Williams and Muxagata, 2006), that in the Mondego estuary was clarified with an obvious residence behavior of the species and the more saline waters preference of the copepodites. *O. nana* is a widespread neritic species. It has been extensively reported in estuarine 'small' mesh sampling (<200 μm) although, because of its small size, it is consistently underestimated.

The structure of Copepod species' distribution in the Mondego estuary proved to be best correlated with environmental factors registered in spring. It is also of note that depth and tide currents were the main physical conditions under which the vertical migration patterns resulted at horizontal displacement or retention of copepod species, while day and night were not ecologically significant, unless in summer.

Acknowledgments

The Portuguese Foundation for Science and Technology (FCT, Portugal) financed A. M. M. Gonçalves through a PhD research grant (SFRH/BD/30475/2006).

A special thanks to all colleagues that helped during this fieldwork.

References

Antacli, J.C., Hernández, D., Sabatini, M.E., 2010. Estimating copepods' abundance with paired nets: implications of mesh size for population studies. Journal of Sea Research 63, 71–77.

Azeiteiro, U.M., Marques, J.C., Ré, P., 2000. Zooplankton assemblages in a shallow seasonally tidal estuary in temperate Atlantic Ocean (Western Portugal: Mondego estuary). Arquivos do Museu Bocage 3, 357–376.

Cervetto, G., Gaudy, R., Pagano, M., 1999. Influence of salinity on distribution of *Acartia tonsa* (copepod calanoida). Journal of Experimental Marine Biology and Ecology 235, 33–45.

Damotharan, P., Perumal, N.V., Arumugam, M., Perumal, P., Vijayalakshmi, S., Balasubramanian, T., 2010. Studies on zooplankton ecology from Kodiakkarai (Point Calimere) coastal waters (South East coast of India). Research Journal of Biological Sciences 5, 187–198.

David, V., Sautour, B., Chardy, P., 2007. Successful colonization of the calanoid copepod *Acartia tonsa* in the oligo-mesohaline area of the Gironde estuary (SW France) – Natural or anthropogenic forcing? Estuarine, Coastal and Shelf Science 71, 429–442.

Dolédec, S., Chessel, D., 1994. Co-inertia analysis: an alternative method for studying species-environment relationships. Freshwater Biology 31, 277–294.

Epifanio, C.E., Garvine, R.W., 2001. Larval transport on the Atlantic continental shelf of North America: a review. Estuarine, Coastal and Shelf Science 52, 51–77.

Falcão, J., Marques, S.C., Pardal, M.A., Marques, J.C., Primo, A.L., Azeiteiro, U.M., 2011. Mesozooplankton structural responses in a shallow temperate estuary following restoration measures. Estuarine, Coastal and Shelf Science. doi:10.1016/j.ecss.2011.06.007.

Forward Jr., R.B., Tankersley, R.A., 2001. Selective tidal-stream transport of marine animals. Oceanography and Marine Biology: an Annual Review 39, 305–353.

Gallienne, C.P., Robins, D.B., 2001. Is *Oithona* the most important copepod in the world's oceans? Journal of Plankton Research 12, 1421–1432.

Gonçalves, A.M.M., De Troch, M., Marques, S.C., Pardal, M.A., Azeiteiro, U.M., 2010a. Spatial and temporal distribution of harpacticoid copepods in Mondego estuary. Journal of the Marine Biological Association of the United Kingdom 90, 1279–1290.

Gonçalves, A.M.M., Pardal, M.A., Marques, S.C., De Troch, M., Azeiteiro, U.M., 2010b. Distribution and composition of small-size zooplankton fraction in a temperate shallow estuary (western Portugal). Fresenius Environmental Bulletin 19 (12b), 3160–3176.

Gonçalves, F., Ribeiro, R., Soares, A.M.V.M., 2003. Comparison between two lunar situations on emission and larval transport of decapod larvae in the Mondego estuary (Portugal). Acta Oecologica 24, S183–S190.

Hill, A.E., 1991. A mechanism for horizontal zooplankton transport by vertical migration in tidal currents. Marine Biology 111, 485–492.

Hopcroft, R.R., Clarke, C., Nelson, R.J., Raskoff, K.A., 2005. Zooplankton communities of the Arctic's Canada Basin: the contribution by smaller taxa. Polar Biology 28, 197–206.

Hsieh, H.-L., Fan, L.-F., Chen, C.-P., Wu, J.-T., Liu, W.-C., 2010. Effects of semidiurnal tidal circulation on the distribution of holo- and meroplankton in a subtropical estuary. Journal of Plankton Research 32, 829–841.

Islam, S., Hibino, M., Tanaka, M., 2007. Tidal and diurnal variations in larval fish abundance in an estuarine inlet in Ariake Bay, Japan: implication for selective tidal stream transport. Ecological Research 22, 165–171.

Jager, Z., 1999. Selective tidal stream transport of Flounder larvae (*Platichthys flesus* L.) in the Dollard (Ems estuary). Estuarine, Coastal and Shelf Science 49, 347–362.

Jessopp, M.J., McAllen, R.J., 2008. Go with the flow: tidal import and export of larvae from semi-enclosed bays. Hydrobiologia 606, 81–92.

Kennish, M.J., 1990. Ecology of Estuaries: Biological Aspects, vol. II. CRC Press, Boca Raton, Florida, p. 391.

Kršinić, F., Dubravka, B., Precali, R., Kraus, R., 2007. Quantitative variability of the copepod assemblages in the northern Adriatic Sea from 1993 to 1997. Estuarine, Coastal and Shelf Science 74, 528–538.

Limnologisk Metodik, 1992. In: Ferskvandsbiologisk Laboratorium. Københavns Universitet. Akademisk Forlag, København.

Marques, S.C., Azeiteiro, U.M., Martinho, F., Pardal, M.A., 2007a. Climate variability and planktonic communities: the effects of an extreme event (severe drought) in a southern European estuary. Estuarine, Coastal and Shelf Science 73, 725–734.

Marques, S.C., Pardal, M.A., Pereira, M.J., Gonçalves, F., Marques, J.C., Azeiteiro, U.M., 2007b. Zooplankton distribution and dynamics in a temperate shallow estuary. Hydrobiologia 587, 213–223.

Marques, S.C., Azeiteiro, U.M., Leandro, S., Queiroga, H., Primo, L., Martinho, F., Viegas, I., Pardal, M.A., 2008. Predicting zooplankton response to environmental changes in a temperate estuarine ecosystem. Marine Biology 155, 531–541.

Marques, S.C., Azeiteiro, U.M., Martinho, F., Viegas, I., Pardal, M.A., 2009. Evaluation of estuarine mesozooplankton dynamics at a fine temporal scale: the role of seasonal, lunar, and diel cycles. Journal of Plankton Research 31, 1249–1263.

Marques, S.C., Pardal, M.A., Mendes, S., Azeiteiro, U.M., 2011. Using multitable techniques for assessing the temporal variability of species–environment relationship in a copepod community from a temperate estuarine ecosystem. Journal of Experimental Marine Biology and Ecology 405, 59–67. doi:10.1016/j.jembe.2011.05.015.

McLusky, D.S., Elliott, M., 2004. The Estuarine Ecosystem. Ecology, Threats, and Management, third ed. Oxford University Press Inc., New York, 214 pp.

Mendes, S., Gómez, M.J.F., Resende, P., Pereira, M.J., Galindo-Villardón, P., Azeiteiro, U.M., 2009. Spatio-temporal structure of diatom assemblages in a temperate estuary. A STATICO analysis. Estuarine, Coastal and Shelf Science 84, 637–644. doi:10.1016/j.ecss.2009.08.003.

Metcalfe, J.D., Arnold, G.P., Webb, P.W., 1990. The energetics of migration by selective tidal-stream transport: an analysis for plaice tracked in the southern North Sea. Journal of the Marine Biological Association of the United Kingdom 70, 149–162.

Morgado, F., Antunes, C., Pastorinho, R., 2003. Distribution and patterns of emergence of suprabenthic and pelagic crustaceans in a shallow temperate estuary (Ria de Aveiro, Portugal). Acta Oecologica 24, S205–S217.

Papadopoulos, I., Wooldridge, T.H., Newman, B.K., 2002. Larval life history strategies of sub-tropical southern African estuarine brachyuran crabs and implications for tidal inlet management. Wetlands Ecology and Management 10, 249–256.

Paula, J., 1989. Rhythms of larval release of decapod crustaceans in the Mira Estuary, Portugal. Marine Biology 100, 309–312.

Porri, F., McQuaid, C.D., Froneman, P., 2007. Spatio-temporal variability of small copepods (especially *Oithona plumifera*) in shallow nearshore waters off the south coast of South Africa. Estuarine, Coastal and Shelf Science 72, 711–720.

Primo, A.L., Azeiteiro, U.M., Marques, S.C., Martinho, F., Pardal, M.A., 2009. Changes in zooplankton diversity and distribution pattern under varying precipitation regimes in a southern temperate estuary. Estuarine, Coastal and Shelf Science 82, 341–347.

Queiroga, H., Costlow Jr., J.D., Moreira, M.H., 1994. Larval abundance patterns of *Carcinus maenas* (Decapoda, Brachyura) in Canal de Mira (Ria de Aveiro, Portugal). Marine Ecology Progress Series 111, 63–72.

Queiroga, H., 1996. Distribution and drift of the crab *Carcinus maenas* (L.) (Decapoda, Portunidae) larvae over the continental shelf off northern Portugal in April of 1991. Journal of Plankton Research 18, 1981–2000.

Queiroga, H., Costlow Jr., J.D., Moreira, M.H., 1997. Vertical migration of the crab *Carcinus maenas* first zoeae: implications for tidal stream transport. Marine Ecology Progress Series 149, 121–132.

Please cite this article in press as: Gonçalves, A.M.M., et al., Diel vertical behavior of Copepoda community (naupliar, copepodites and adults) at the boundary of a temperate estuary and coastal waters, Estuarine, Coastal and Shelf Science (2011), doi:10.1016/j.ecss.2011.11.018

Riccardi, N., 2010. Selectivity of plankton nets over mesozooplankton taxa: implications for abundance, biomass and diversity estimation. Journal of Limnology 69 (2), 287–296.

Rooper, C.N., Gunderson, D.R., Hickey, B.M., 2006. An examination of the feasibility of passive transport from coastal spawning grounds to estuarine nursery areas for English sole. Estuarine. Coastal and Shelf Science 68, 609–618.

Satapoomin, S., Nielsen, T.G., Hansen, P.J., 2004. Andaman Sea copepods: spatio-temporal variations in biomass and production, and role in the pelagic food web. Marine Ecology Progress Series 274, 99–122.

Simier, M., Blanc, L., Pellegrin, F., Nandris, D., 1999. Approche simultanée de K couples de tableaux: apllication à l'étude des relations pathologie végétale-environnement. Revue de statisque appliquée 47, 31–46.

Strickland, J.D.H., Parsons, T.R., 1972. A practical handbook of seawater analysis. Fisheries Research Board of Canada Bulletin 167, 1–311. second ed.

Tackx, M.L.M., Pauw, N., van Mieghem, R., Azémar, F., Hannouti, A., van Damme, S., Fiers, F., Daro, N., Meire, P., 2004. Zooplankton in the Schelde estuary, Belgium and the Netherlands. Spatial and temporal patterns. Journal of Plankton Research 26, 133–141.

Takahashi, T., Uchiyama, I., 2008. Seasonal changes in the density and vertical distribution of nauplii, copepodites and adults of the genera *Oithona* and *Oncaea* (Copepoda) in the surface water of Toyama Bay, southern Sea of Japan. Plankton and Benthos Research 3, 143–151.

Thioulouse, J., Chessel, D., 1987. Les analyses multitableaux en écologie factorielle. I. De la typologie d'état à la typologie de fonctionnement par l'analyse triadique. Acta Oecologica, Oecologia Generalis 8, 463–480.

Thioulouse, J., Chessel, D., Dolédec, S., Olivier, J.M., 1997. ADE-4: a multivariate analysis and graphical display software. Statistics and Computing 7, 75–83.

Thioulouse, J., Simier, M., Chessel, D., 2004. Simultaneous analysis of a sequence of paired ecological tables. Ecology 85, 272–283.

Titelman, J., Fiksen, Ø, 2004. Ontogenic vertical distribution patterns in small copepods: field observations and model predictions. Marine Ecology Progress Series 284, 49–63.

Tseng, L.-C., Dahms, H.-U., Hung, J.-J., Chen, Q.-C., Hwang, J.-S., 2011. Can different mesh sizes affect the results of copepod community studies? Journal of Experimental Marine Biology and Ecology 398 (1–2), 47–55.

Vieira, L., Azeiteiro, U., Ré, P., Pastorinho, R., Marques, J.C., Morgado, F., 2003. Zooplankton distribution in a temperate estuary (Mondego estuary southern arm: western Portugal). Acta Oecologica 24, S163–S173.

Williams, J.A., Muxagata, E., 2006. The seasonal abundance and production of *Oihtona nana* (Copepoda: Cyclopoida) in Southampton water. Journal of Plankton Research 28, 1055–1065.

Zeng, C., Naylor, E., 1997. Rythms of larval release in the shore crab *Carcinus maenas* (Decapoda: Brachyura). Journal of the Marine Biological Association of the United Kingdom 77, 451–461.

Please cite this article in press as: Gonçalves, A.M.M., et al., Diel vertical behavior of Copepoda community (naupliar, copepodites and adults) at the boundary of a temperate estuary and coastal waters, Estuarine, Coastal and Shelf Science (2011), doi:10.1016/j.ecss.2011.11.018

PAPER VII

Response to Climatic variability of Copepoda life history stages in a southern European temperate estuary

Gonçalves AMM, Pardal MA, Marques SC, Mendes S,

Fernández-Gómez MJ, Galindo-Villardón MP, Azeiteiro UM

Zoological Studies, 2011 (in press) (#)

(#) Additional information:

- Publ. Taipei : Biodiversity Research Center Academia Sinica
- ISSN: 1021-5506
- **Indexed in:** Journal Citation Report - Science Citation Index
- **Impact Factors:** IF 2010 (2 years): 1.046; IF 2010 (5 years): 0.931
- **Category:** ZOOLOGY. Ranking: 65/145 – Q. 2 (Quartile 2)
- **Also indexed in**: Current Contents and the Science Citation Index

Zoological Studies

Editorial Office
Biodiversity Research Center
Academia Sinica, Taipei 115, Taiwan
Tel & Fax: 886-2-27899529 E-mail: zoolstud@gate.sinica.edu.tw

Date: October 24, 2011

ZS 000207

Title: Response to Climatic variability of Copepoda life history stages in a southern European temperate estuary

Author: GONÇALVES, A.M.M.[1,*], PARDAL, M.A.[1], MARQUES S.C.[1], MENDES, S.[2,3], FERNÁNDEZ - GÓMEZ, M.J.[3], GALINDO-VILLARDÓN, M.P.[3], AZEITEIRO, U.M.[1,4]

We are glad to inform you that the manuscript described above has been accepted, and will be published in *Zoological Studies*. You will be informed of the volume and issue number as soon as it be decided.

We thank you for considering our journal for the publication of your research work. Now that your paper has been accepted for publication in our journal, may we remind you that it would be very helpful for the increase in Impact Factor of our journal if you yourself, your friends and colleagues would cite your paper in their future publications.

Lastly, in order to increase the world-wide circulation of our journal, which would also help in the increase in Impact Factor if our journal can reach more readers, we would like you to recommend departments, universities, institutions, or organizations which you think would be suitable for their faculty or staffs to be aware of our journal.

Sincerely yours,
Editorial Office
Zoological Studies
Academia Sinica

Date of submission: 4th February 2011

Revised manuscript: 27th September 2011

Response to Climatic variability of Copepoda life history stages in a southern European temperate estuary

GONÇALVES, A.M.M.[1,*], PARDAL, M.A.[1], MARQUES S.C.[1], MENDES, S.[2,3], FERNÁNDEZ - GÓMEZ, M.J.[3], GALINDO-VILLARDÓN, M.P.[3], AZEITEIRO, U.M.[1,4]

[1]CFE - Centre for Functional Ecology, Department of Life Sciences, University of Coimbra, Apartado 3046, 3001-401 Coimbra, Portugal

[2]GIRM-Marine Resources Research Group, School of Tourism and Maritime Technology, Polytechnic Institute of Leiria, Campus 4, 2520-641 Peniche, Portugal

[3]University of Salamanca, Department of Statistics, 37007 Salamanca, Spain

[4]Universidade Aberta, Department of Sciences and Technology, 4200-055 Porto, Portugal

E-mail addresses: ammendes@student.zoo.uc.pt; ana.goncalves@bio.ua.pt (A. M. M. Gonçalves); mpardal@ci.uc.pt (M. A. Pardal); scotrim@ci.uc.pt (S. C. Marques); susana.mendes@ipleiria.pt (S. Mendes); mjfg@usal.es (M. J. F. Gómez); pgalindo@usal.es (M. P. Galindo-Villardón); Ulisses@univ-ab.pt (U. M. Azeiteiro)

*Corresponding author: ammendes@student.zoo.uc.pt; ana.goncalves@bio.ua.pt

Phone: +351239836386; Fax: +351239823603

Running head: Copepoda responses to a severe drought

Keywords: Copepods; life stages; Mondego estuary; Seasonal and inter-annual variability; STATICO.

Abstract

This study aims to investigate the effects of an extreme climate event (severe drought) on Copepoda ecology. Monthly samples were conducted from 2005 to 2007, at five stations, using a 63 and 335µm mesh nets. Calanoida were represented mainly by *Acartia clausi, Temora longicornis* and *Acartia tonsa* and Cyclopoida by *Oithona plumifera* and *Acanthocyclops robusts. A. clausi and T. longicornis* dominated at the mouth and middle estuary; *A. tonsa* and *A. robustus* were associated to the upper estuary while *O. plumifera* showed the highest densities at the downstream section. Nauplii occurred in higher densities at the mouth. The relationship of copepod assemblages and environmental factors was analyzed using the STATICO method which allowed distinguishing the combination factors that mostly contributed to this relationship. Winter was characterized by high concentrations of nutrients, cold waters and low salinities while summer was related, in general, by high values of phosphate, salinity and temperature. Marine and estuarine species (mainly copepodites) showed high densities in summer. Freshwater species occurred at maximal densities in winter, coincidently with higher river flow. Copepoda assemblages showed a clear seasonal pattern that superimposed to the inter-annual variability. Moreover, the severe drought was responsible for the predominantly dominance of marine species.

Introduction

Estuaries are transition zones between the rivers and the sea differing in biotic (e.g. predation, competition) and abiotic conditions (e.g. temperature, salinity, food quantity and quality). Plankton (and mainly zooplankton) is known to be particularly sensitive to these changes because it is strongly influenced by climatic features and changes in hydrological conditions (Ara 2001, Isari et al. 2007, Hwang et al. 2010a b, Kâ and Hwang 2011). Several studies have focused zooplankton ecology and dynamics (Uriarte and Villate 2005, Lam-Hoai et al. 2006, Hafferssas and Seridji 2010, Hwang et al. 2010b) namely in the Mondego estuary (Azeiteiro et al. 2000, Morgado et al. 2007, Marques et al. 2009, Primo et al. 2009, Gonçalves et al. 2010a b). Copepods are usually the dominant group of mesozooplankton, playing an important role in the trophic food web since they are a link between producers and secondary consumers (Richmond et al. 2007, Hwang et al. 2010b, Hsiao et al. 2011, Kâ and Hwang 2011). Still, studies integrating simultaneously copepods' life history stages (nauplii, juveniles and adults) with the aim to examine their ecology over contrasting environmental conditions is scarce in literature, mainly in European southern systems (Kršinić et al. 2007). Studies have been developed to a single species or taxonomic categories (Incze and Ainaire 1994, Hansen et al. 2004) in a spatio-temporal distribution perspective or performing laboratorial tests to investigate organisms' responses to environmental factors (Cook et al. 2007). Moreover, earlier life history stages (mainly naupliar stages) are pivotal for further development and growth of juveniles and to the maintenance of copepod populations (Cook et al. 2007). Temperature, food quality and quantity are the main environmental factors

controlling stage duration of copepods (Morgado et al. 2007, Hafferssas and Seridji 2010), while preferential salinity regions to lay the eggs are chosen in order to obtain the highest hatch success (Chinnery and Williams 2004; Cook et al. 2007). Thus, studies not addressing the early copepod stages have no information about changes in species dynamics and thus cannot infer about seasonal fluctuations due to a lack of information. This frequent undersampling of small copepod species, and mainly earlier stages (Dias et al 2010, Hwang et al. 2010a), may lead to a limited view of the ecology of planktonic systems (namely in estuarine systems). Furthermore, the understanding of how climate change will affect the planet is a key issue worldwide and lately more attention has been given to the global ecological change. However, few studies have focused on the impact of large-scale weather events, such as an extreme drought, in Copepoda community, in order to assess a holistic and integrative view of an ecosystem to global climate change. This change associated to biological long time series may contribute to further knowledge about inter-annual variations in abundance and diversity of copepod community.

The aims of this study are focused on the life-history stages (nauplii, juveniles and adults) of copepods under two contrasting environmental conditions (an extreme dry year (2005) and regular years (2006 and 2007) in order to: (1) determine the major environmental parameters explaining copepods dynamics in a southern European shallow temperate estuary, (2) reveal seasonal variations in the distribution patterns of copepods in dependence to hydrological factors and (3) determine inter-annual variations of copepods assemblages during an extreme drought.

Materials and Methods

Study area and sampling sites

The Mondego estuary is a small mesotidal system with 8.6 km^2, located in the western Atlantic coast of Portugal (40º 08´ N, 8º 50´ W) (Fig. 1). It comprises two channels; north and south, separated by the Murraceira island about 7 km from the shore joining again near the mouth. These two arms present different hydrological characteristics. The north arm is deeper (4-8m during high tide, tidal range about 2-3 m), presents a low residence time (<1day) and is the location of the commercial harbour and the main navigation channel. At neap tides, this arm is characterized by a salt-wedge during low tide, changing to partially mixed water column at high tide. At spring tides, it is characterized by a partially mixed water column at low tide and well mixed at high tide (Cunha and Dinis 2002). The southern arm is shallower (2-4m deep, during high tide), has higher residence times (2-8 days) and the water circulation is mostly dependent on the tides and on the freshwater input from a small tributary system, the Pranto River. Freshwater discharge of this river is controlled by a sluice according to the water needs of the Mondego valley rice fields.

Sample collection and laboratorial procedures

Copepod samples were collected monthly in the Mondego estuary from February 2005 to December 2007, during high tide, at five sampling stations (M, N1, N2, S1 and S2) distributed throughout both arms (Fig. 1). Copepods were collected by subsurface tows with a 335 µm mesh Bongo net (diameter: 0.5 m) and a 63 µm

mesh net (diameter: 0.30 m), equipped with a Hydro-Bios flow-meter in the mouth (to estimate the volume of water filtered by the nets). Samples were fixed and preserved in 4% buffered formaldehyde in seawater. In order to determine the number of *taxa* and abundances, all samples were counted (individuals m^{-3}) and identified to the lowest possible taxonomic level. Water samples were collected on each sampling station to determine nutrient concentrations (Strickland and Parsons 1972 for nitrates and nitrites, $mg.L^{-1}$; Limnologisk Metodik 1992 for phosphates and ammonia, $mg.L^{-1}$), Chlorophyll *a* concentration (Parsons et al. 1985 – Chl *a*, $mg.m^{-3}$) and total suspended solids (APHA 1995 – TSS, $mg.L^{-1}$). Additionally, several hydrological parameters were measured *in situ*: water temperature (ºC) and salinity (WTW Cond 330i), dissolved oxygen concentration (WTW OXI 330i – DO, $mg.L^{-1}$), pH (WTW pH 330i) and transparency with a Secchi disc depth (m). Monthly precipitation and long-term monthly average precipitation (from 1971-2000) were measured at the Soure 13 F/01G station and acquired from INAG – Portuguese Water Institute (http://snirh.inag.pt). Freshwater runoff from Mondego River was obtained from INAG station Açude Ponte Coimbra 12G/01AE, near the city of Coimbra (located 40 km upstream).

Data analysis

Only the most abundant *taxa*, having a minimal mean occurrence of 0.1% of the total density observed in the study area were considered. This cut-off eliminated the species that occurred rarely, some being observed on few or rare occasions. Moreover, well-represented species can be viewed as proxies of copepod dynamics and ecosystem functioning.

In order to investigate the spatial variability in the copepod community structure, the species density and environmental parameters of each sampling site were combined to generate two series of tables (Fig. 2): one for the environmental variables and the other one for species density. Each pair of tables sharing the same sampling period (spring 2005 to autumn 2007) for five sites (M, N1, N2, S1 and S2). Species abundance was transformed using log(x+1) prior to calculations, to minimize the dominant effect of exceptional catches. Environmental data were normalized by the use of scaling to unit standard deviation within the mode to homogenize the table.

The common structure between environmental and species density tables and the stability of this structure across the sampling stations were assessed by STATICO method (Simier et al. 1999; Thioulouse et al. 2004). The STATICO method was used by Simier et al. (2006) and Carassou and Ponton (2007) to study the spatial and seasonal variability of fish assemblages in Gambia estuary and coastal areas of New Caledonia, respectively, and by Mendes et al. (2009) to describe the spatio-temporal structure of diatom assemblages in Ria de Aveiro (Portugal). This method proceeds in three stages: (1) the first stage consists in analyzing each table by a one-table method (normed PCA of the environmental variables and centered PCA of the species data). The calculation of the vectorial correlations (RV) matrix between stations (in terms of the co-structure between environment and species density) allows the comparison of the stations and the representation of the proximity between stations. The function of this step is to attribute a weight to each station sub-matrix; (2) each pair of tables is linked by the Co-inertia analysis (Dolédec and Chessel 1994) which provides an average image of the co-structure (species-variables); (3) Partial Triadic Analysis (Thioulouse and Chessel 1987) is finally used to analyze this sequence. It is a three-step procedure, namely the interstructure, the

compromise and the intrastructure (or trajectories) analyses. In fact, the compromise is the main step of the analysis (Thioulouse et al. 2004). It is based on the compromise table which is computed as the weighted mean of all the tables of the series, using the components of the first eigenvector of the interstructure as weights. This table is called the compromise, and it has the same dimensions and the same structure and meaning as the tables of the series. It is analyzed by a PCA, giving a picture of the structures common to all the tables. STATICO also enables to plot the projection of the sampling seasons of each original table on the compromise axes (of the PCA factor map), in terms of species abundances and environmental factors structures. Hence, it is possible to discuss the correlation between species distribution and environmental factors. Calculations and graphs were done using ADE-4 software (Thioulouse et al. 1997).

Results

Climate – precipitation and environmental background

In the Mondego estuary a clear seasonal and yearly variation of rainfall and freshwater discharges was observed during the three-year period (Fig. 3A). In 2005 an extreme drought was recorded with precipitation and freshwater discharge values much lower than the 1971-2000 average, causing one of the biggest droughts of the 20^{th} century in Portugal. A severe reduction was evidenced in freshwater flow with the lowest value in 2005 almost 48-fold lower the highest in 2006. In 2006 and 2007 precipitation values were closer to average except in October 2006 where was

registered an above-mean precipitation (Fig. 3A). So, the last two years within the study period were considered regular years.

In general, water temperature and salinity showed similar variation patterns during the study period with lower values in winter months (Fig 3B). Nevertheless, salinity was highly variable during the sampling period with the lowest value in 2006 almost 14-fold lower than the highest in 2005 (corresponding to the extreme drought). As expected, the highest values of salinity were observed during 2005. Water temperature showed a typical pattern for temperate regions, ranging from 10.12 ± 0.51 ºC to 20.86 ± 2.55 ºC (Fig. 3B).

Analysis of the interstructure (between-stations analysis)

From the 55 different Copepoda species identified 28 (including nauplii, copepodites and adults stages) were dominant and occurred regularly. The interstructure factor map of the STATICO analysis, based on the 12 environmental variables (pH, DO, temperature, salinity, transparency, Chl a, TSS and nutrients) and on the abundances of the 28 copepod species from different stages (nauplii, copepodites and adults), showed that the relationship between environmental variables and species appeared to be stronger in N1 (with the longest arrow) followed, in decrease order of importance in the compromise, by M, S1, N2 and S2. This means that the compromise will be more influenced by N1, M and S1) (Fig. 4A). The remaining sampling sites (N2 and S2) presented short arrows, meaning that the corresponding tables are less structured being lower its importance in the compromise. The first two axes represented, respectively, 69% and 11% of the total variability (Fig. 4B).

The matrix presenting the RV between the stations sub-matrices (Table 1) showed the strongest correlation (RV = 0.77) observed between the stations S1 and N1 whereas the stations S2 and M pointed out the weakest one (RV = 0.32). From the analysis of the interstructure is possible to analyze the contribution of each sub-matrix in the construction of the compromise (Table 1). They represent the weight of each sub-matrix in the definition of the compromise. It seems that the sub-matrices M, N1 and S1 contributed a larger part in the definition of the compromise suggesting that the other stations had more particular structures leading to a weaker weight. Regarding to cos^2 (Table 1), an indicator of how much the compromise expresses the information contained in each table, the station N1 was the one that fits best (cos^2 = 0.81), followed by stations M and S1 (cos^2 =0.71 and 0.64, respectively). Lastly, the seasonal dynamics at stations N2 and S2 shows the least accuracy with the compromise (cos^2 = 0.54 and 0.24, respectively), in terms of the co-structure between environment and species density.

Spatial Structure

The factor plots of the first two axes of the compromise analysis are shown for the copepod community and the environmental variables (Fig. 5). The first axis was clearly dominant, and accounted for 89% of the explained variance in contrast with the second axis which accounted for 4% of the explained variance being much less significant (Fig. 5C). Therefore, they provided a good summary and typology of the spatial species organization, on the basis of the common structure, for the sampling sites across the 3 years.

The factor map of the compromise for the copepod community in the STATICO analysis indicates that the most abundant species in the samples were mainly

associated with the same environmental variables: temperature and salinity (Fig. 5A, 5B). Copepodites of *Oithona plumifera* Baird, 1843, *Euterpina acutifrons* Dana, 1848, Copepodites of *E. acutifrons*, Copepodites of *Oithona* sp. and Copepodites of *Temora longicornis* O. F. Müller, 1785 were more abundant when salinity and transparency were high and the concentrations of ammonium, nitrites, nitrates and silica were low (Fig. 5A, 5B). In particular Copepodites of *O. plumifera, E. acutifrons* and Copepodites of *E. acutifrons* were clearly influenced by high values of temperature, presenting during the study period abundances of $2.62\times10^{4}\pm 1.33\times10^{5}$ ind.m^{-3}; $3.92\times10^{3}\pm 1.44\times10^{4}$ ind.m^{-3} and $9.59\times10^{3}\pm 4.99\times10^{4}$ ind.m^{-3}, correspondingly. Beyond that, the high density of Copepodites of *O. plumifera* ($2.62\times10^{4} \pm 1.33\times10^{5}$ ind.m^{-3}) was also under the influence of high values of phosphates and low concentrations of dissolved oxygen. In opposition were Copepodites and adults *of Paronychocamptus nanus* Sars, 1980 ($1.10\times10^{3}\pm 5.75\times10^{3}$ ind.m^{-3}; $3.59\times10^{2}\pm 1.52\times10^{3}$ ind.m^{-3} respectively), Copepodites and adults of *Acanthocyclops robustus* G. O. Sars, 1863 ($1.67\times10^{3}\pm 1.12\times10^{4}$ ind.m^{-3}; 1.20±3.76 ind.m^{-3}, respectively) and Copepodites of *Centropages* sp.($6.46\times10^{2}\pm 5.48\times10^{3}$ ind.m^{-3}). These were more abundant in waters of lower salinity and temperature combined with the reduction of transparency and TSS values. Moreover, Copepodites of *Acartia* sp. ($1.94\times10^{3}\pm7.97\times10^{3}$ ind.m^{-3}), Copepodites of *Clausocalanus* sp. (2.19×10^{2} $\pm 1.17\times10^{3}$ ind.m^{-3}), Copepodites of *Acartia clausi* Giesbrecht, 1889 (9.94×10^{2} ±

3.66×10^3 ind.m^{-3}), Copepodites of *T. longicornis* ($1.05\times10^3\pm5.09\times10^3$ ind.m^{-3}), Copepodites of *Oithona* sp. ($1.46\times10^4\pm5.00\times10^4$ ind.m^{-3}) and Copepodites of *Paracalanus*-type form (calanoid) ($2.22\times10^3\pm8.85\times10^3$ ind.m^{-3}) were abundant when salinity and TSS concentrations were elevated and nitrites, ammonium, pH and phosphates were low (Fig. 5A, 5B). Copepodites of *Pseudocalanus* sp. *Paracalanus* sp. and *Clausocalanus* sp. are referred as *Paracalanus*-type form when they do not show the main characteristics of each genus to their identification. Regardless of species, the first axis of the compromise factor map for environmental parameters mainly described a negative association between salinity and nitrates in addition to silica (Fig. 5B). On this factorial plan pH, ammonium and chlorophyll *a* concentration presented a weak representation (Fig. 5B). The second axis opposed phosphates, as well as temperature and dissolved oxygen.

Internal typology of each site in the composition of species and variation of environmental factors

For each survey, the projection on the compromise axes of the 28 species (see table 2 for species codes) and of the 12 environmental variables is shown in figure 6. The stable part of the species-environment dynamics revealed by the compromise analysis was better expressed by the sampling stations M, N1 and S1 surveys (Fig. 6A, 6B). The N2 and S2 surveys, also showed similar oppositions, however with a slight difference in the density of Copepodites of *A. clausi*, *T. longicornis*, *Acartia* sp., *Clausocalanus* sp. and *Paracalanus*-type form combined with the dynamics of

environment over these stations (Fig. 6A, 6B). Species' distribution patterns are more similar among downstream stations, with a high correlation with species abundances and the environment. In upstream stations species presented lower abundances and lower correlation with the environment, mainly in the southern station (S2). At a spatial scale, salinity and temperature were unrelated at downstream stations, whereas at upstream stations these two factors presented a positive correlation (Fig. 6B). Furthermore, nutrients concentration appeared positively correlated on the left side of axis 1, along which the freshwater species *A. robustus* appeared to distribute. Marine species are mostly located on the right-hand side of the first axis, which means higher temperature, salinity and TSS concentration. Moreover, looking at some environmental characteristics, it can be confirmed that the stations M, N1, S1 are mutually more similar, as compared to the other ones.

The major groups of copepods, calanoida, cyclopoida and harpacticoida are represented in the study area. Calanoida were represented mainly by marine species (e.g. *A. clausi* and *T. longicornis*) and by the estuarine species *Acartia tonsa* Dana, 1849. Cyclopoida were mainly represented by euryhaline species *O. plumifera* and the freshwater species *A. robustus. A. clausi* and *T. longicornis* dominated principally at the mouth (M) and middle estuary (N1 and S1) while *A. tonsa* and *A. robustus* were more associated to the upper estuary (N2 and S2). The cyclopoid *O. plumifera*, mostly copepodite stages, occur along the estuary, showing the highest densities at the mouth (M) and middle north arm (N1). This species (mainly juvenile stages) represents one of the most abundant copepod species. Harpacticoids species are represented by the estuarine species *E. acutifrons* and *P. nanus*, presenting always lower densities than the Euterpinidae species. *E. acutifrons* occurs along the whole salinity gradient, showing a widely distribution in the estuary. Although *P. nanus* is

also found at the five sampling stations, this species shows higher densities at middle estuary (N1 and S1). The order Poecilostomatoida is represented by *Sapphirina* sp. This taxon shows the lowest abundance from all the copepods' ($9.10\times10^{-3}\pm4.47\times10^{-2}$ ind.m^{-3}) being considered a rare species. Nauplii of copepods occur along the whole estuary, still were observed at higher densities at the mouth station (St M).

The co-structure graphics (divided according sampling sites) clearly showed the dynamics of species-environment relationships and highlighted differences between seasons (Fig. 7). The projections of the species points (origin of arrows) were located differently according to the site raising their dispersion from the upstream stations (N2 and S2) to the mouth station (St M). These positions were essentially related to the salinity and temperature, with the smallest values observed in upstream stations. Whatever the date, the species points (origin of arrows) were more stable than the environmental points, expressing the steady establishment of the species assemblages in the estuary, in spite of the high environmental variability (principally for salinity, temperature and phosphates). In general, winter was mostly projected on the left-hand side of the first axis, characterized by the high concentrations of nitrates, nitrites, silica and ammonium, in opposition to high salinity and warm waters and, for the most part, with high density of juveniles and adults of *A. robustus* and *P. nanus*. In contrast, nauplii of copepods appeared to be numerically less important during winter months. This season was the most regularly projection and only presented exceptions at S1 and S2. Summer season was characterized by the factors that described axis 2, which means that at stations M, N1, S1 and N2 summer was mainly under the influence of high values of phosphates, salinity, temperature, TSS and dissolved oxygen. By other hand, for S2 summer was influenced by highest phosphates, transparency, salinity, pH and temperature. From a species point of view summer was

principally characterized by high density of Copepodites of *O. plumifera*, *E. acutifrons* (adults and juveniles), Copepodites of *Oithona* sp., Copepodites of *A. clausi*, Copepodites of *Acartia* sp., Copepodites of *Paracalanus*-type form, Copepodites of *T. longicornis* and Copepodites of *Clausocalanus* sp.. Spring and autumn presented a mixed pattern, without a clearly dominant axis. The exceptions were at S1 for autumn (mostly projected on the right-hand side of the first axis) and N2 for spring (mostly projected on the left-hand side of the first axis). The arrows were mostly short for summer, expressing a good coincidence between environmental conditions and plankton structure. The exception occurred at N2, with a poor fit between the copepod species' densities and environment (long arrows) for some surveys in summer 2005. However, at this season, there was a higher correlation between the distribution of copepod densities and the environment under the direct influence of high values of phosphates, salinity, TSS, dissolved oxygen and temperature at the downstream stations (M, N1 and S1). Furthermore, there is a strong co-structure between species densities and environmental factors at upstream station of the southern arm (S2) under the influence of high values of phosphate and salinity, transparency, temperature and pH, in summer months. By the other hand, at S1 all winter surveys presented short arrows expressing a strong co-structure between species and environment, which means that the environmental factors (high values of nitrates, nitrites, silica and ammonia and low values of temperature and salinity concentrations) explained well the distribution of species at this season. At upstream stations (S2 and N2) the arrows were mostly long which means a lower correlation between the distribution of copepod densities and the environment.

Discussion

This study focused on the abundance of key copepod species and their respective life stages in the Mondego estuary. Moreover, it is also examined the relationship between environmental factors and species densities and distribution, over a time course of three years. Indeed, in the Mondego estuary, as in other coastal areas (Uriarte and Villate 2005, Hwang et al. 2010b, Kâ and Hwang 2011) the zooplankton group of Copepoda was the dominating one. The most abundant copepod species (e.g. *A. clausi, T. longicornis, O. plumifera, A. tonsa, A. robustus, P. nanus and E. acutifrons*) were associated with different environmental variables. The marine component of the community was associated with saline waters, high values of TSS and low concentrations of nutrients, while the freshwater component was characteristic of low values of salinity and high nutrients concentration and dissolved oxygen.

Concerning seasonal variations, winter is marked by a higher presence of freshwater species and lower densities of nauplii. This abundance and distributional patterns were also observed in other works (Tackx et al. 2004, Primo et al. 2009, Gonçalves et al. 2010a b). In spring/summer months there are higher abundances of *A. clausi*, *T. longicornis* and *Clausocalanus arcuicornis* Dana, 1849 (mainly juveniles) while in summer/autumn months copepodites of *O. plumifera* were more abundant. Similarly, Villate et al. (2004) reported higher abundances of the calanoids *A. clausi*, *Paracalanus parvus* Claus, 1863, *Clausocalanus* spp., *Pseudocalanus elongatus* Boeck, 1865 and *T. longicornis* and the small marine cyclopoids of the genus *Oithona* and *Oncaea*, related with saline waters, at the estuarine systems of the Basque coast. Moreover, *Centropages typicus* Krøyer, 1849, *Oithona nana* Giesbrecht, 1892 and *P.*

parvus are typically surface species with high temperature and low salinity affinities whilst *O. plumifera*, *Clausocalanus* sp. *P. elongatus* and *T. longicornis* were reported as species preferring deeper waters with lower temperature and higher salinity concentration. Although several authors (e.g. Villate et al. 1993, Holste and Peck 2006) stated that copepod nauplii distribute towards higher salinities. naupliar stages show a high sensitivity to environmental factors. However, this behavior may change from species to species and also from adults to juveniles and to nauplii of the same species, being some of them more sensitive than others (Chinnery and Williams 2004). According to Incze and Ainaire (1994) the peak of occurrence of copepod nauplii is on latter spring related with the greatest concentrations of Chlorophyll *a*. Chinnery and Williams (2004) stated that, differential distribution of different life stages in estuarine systems may be due to the poorly swimming activity of the naupliar stages. The weak ability to swim leads nauplii into areas of higher salinities and swept them towards the mouth of the estuary where they stay and develop to later stages moving afterwards back up the estuary to take up their adult distribution patterns.

At the Mondego estuary, freshwater species show higher densities at upstream stations (N2 and S2) because of the high influence of the freshwater from the Mondego and the Pranto rivers. Marine and estuarine species are found at the middle stations (N1 and S1) due to the intrusion of marine water in both stations, while at the mouth station (M) a higher number of marine species is verified (Primo et al. 2009, Gonçalves et al. 2010b).

Downstream stations (St M, St S1 and St N1) exhibited a strong correlation, together with the most abundant species mainly associated with salinity and temperature, presenting a rich marine fraction. This fact agrees with other studies carried out in a

wide range of European estuaries (e.g. Mouny and Dauvin 2002, Tackx et al. 2004), that emphasized the significant influence that salinity has on zooplankton abundance, composition and distribution. The higher salinities (and low freshwater inflow) registered in 2005 (the Portuguese Weather Institute - http://web.meteo.pt/clima.jsp - classified the drought period of 2005 as the worst drought of the past 60 years) lead to an increase of marine zooplankters in the estuary. The same was reported by Primo et al. (2009) for the mesozooplankton and by Martinho et al. (2007) for fish species. The severe drought in 2005 was responsible for a clear spatial shift of copepod community predominantly dominated by marine species, which remained along the next regular climatic years (2006 and 2007). Together with the marine spatial preferences of naupliar forms, lead to a clear absence of an inter-annual distributional pattern in the Copepoda assemblages. Seasonality was also detected along the study period, driving to zooplankton assemblage's variability in Mondego estuary. This study also contributed to a detail and further knowledge of naupliar and copepodites' stages, and also determined spatial patterns of distribution. Concerning the study area, *A. tonsa* and *A. clausi* are one of the copepods with higher abundance in the Mondego estuary contributing to the increase of copepod dominance in the estuary. The former exhibit higher densities at upstream stations (N2 and S2) and the latter appear in much higher densities at downstream stations (M, N1 and S1) (Primo et al. 2009, Gonçalves et al. 2010b). At the estuaries of Bilbao (polluted estuary) and Urdaibai (undisturbed estuary), located on the Basque coast (Bay of Biscay) with salinities varying between 35 and 31, Uriarte and Villate (2005) stated that responses of copepods may be related with water desalination and their tolerance to pollution. These authors referred *A. clausi* and *P. parvus* as the most abundant species in Bilbao and Urdaibai estuaries. Still, at the latter estuary higher densities of *P. elongatus*, *E.*

acutifrons and harpacticoids were also referred. A significant reduction of copepod species with a decrease of salinity values at the polluted estuary was observed, whilst *E. acutifrons* and *Oithona helgolandica* Claus, 1863 did not showed significantly differences, which is in agreement to their wide spread distribution. Similarly to our results, some authors (e.g. Ara 2001, Uriarte and Villate 2005) stated that *E. acutifrons* appeared without a clear seasonal trend, occurring throughout the year, and tolerate a wide range of salinity and temperature. The species' highest abundance is found in estuaries and coastal waters rather than offshore waters (Björnberg 1963). Uriarte and Villate (2005) reported a diminishment of *A. clausi*, *P. parvus* and *O. plumifera* with the decrease of salinity, with the former species reaching higher abundances in spring. Besides, the distribution of *A. clausi* is affected by the presence of other species - *Acartia bifilosa* Giesbrecht, 1881, - which does not survive so well at higher salinities as do *A. clausi*. Moreover, *A. clausi* species shows higher hatching success at salinities of 33.3 than *A. bifilosa* (Chinnery and Williams 2004; Uriarte and Villate 2005). Indeed, *A. tonsa* seems to have a physiological plasticity in terms of egg hatching being the most tolerant of the *Acartia* congeners, with a great hatching success at a range of salinity concentrations (from 15.5 to 33.3) (Chinnery and Williams 2004). In terms of distribution, *A. tonsa* is a common species in estuaries and European seas, occurring from temperate to subtropical waters, restricting its distribution to habitats with high levels of food (David et al. 2007, Morgado et al. 2007). Chinnery and Williams (2004) classified *A. tonsa* and *A. clausi* as summer species. David et al. (2007) stated that *A. tonsa* shows a seasonal pattern in the north European estuaries characterized by a peak of abundance in late summer and autumn and second spring peak observed in the southern European estuaries. Two assumptions could explain these peaks of abundance: 1.warmer conditions as

temperature is known to be the main factor controlling the biological cycle of copepods (Gaudy 1972, Hsiao et al. 2011) and the world distribution of *A. tonsa* (Conover 1956); 2. diapauses eggs hatch when water temperature exceeds 10 ºC and population rapidly increase above 15ºC (McAlice 1981) explaining the spring peak and high levels of abundance in the autumn. In fact, David et al. (2007) reported that abundances of *A. tonsa* are significantly and negatively correlated with suspended particulate matter concentrations and positively and significantly correlated with water temperature and salinity. *P. elongatus* and p-*Calanus* showed higher peaks of abundance in winter-spring, rising with higher values of salinity, whereas *Oncaea media* and *Temora stylifera* Dana, 1849 occurred at higher densities in summer along the salinity gradient.

In the Adriatic Sea system was regarded inter-annual and seasonal variability for some copepod groups over flooded and warmer events, respectively (Kršinić et al. 2007). Nauplii was the most numerous fraction of all copepods groups, showing the highest abundances during warmer and saltier conditions that were referred by the authors as an atypical distribution which may be related to specific summer currents. Moreover, harpacticoids' highest abundances were also reported at warmer and saltier conditions, with the highest values occurring in summer and autumn months. Calanoids and oithonids highest densities were observed in warmer months, which may be related with the regulation of phytoplankton production levels during this period (Kršinić et al. 2007). Temperature is one of the most important factors controlling the biological cycle of copepods, affecting juveniles and their growth rate (Hsiao et al. 2011). Moreover, adults may be tolerant to a wide range of salinity, but the earlier stages (nauplii and juveniles) may be not. Thus, the vulnerability of nauplii to environmental conditions, mainly salinity and temperature, can threat the

maintenance of its population as they have a narrow survival limits, suffering from a great amount of mortality (Chinnery and Williams 2004). Moreover, since copepods are an important component of the diet of bentho-pelagic invertebrates and peaks or absence of some copepod populations may influence the presence and life cycle of others populations (Carrasco et al. 2007), their dynamics allows us to infer about ecological changes in aquatic systems (Hwang et al. 2010b). Therefore, more studies must be conducted in terms of populations' distribution (and so life-history stages) to understand the relationships among species and their development influenced by environmental factors. These studies in combination with laboratory experiments give us the chance to delineate the life cycle of species, contributing to a further knowledge of populations and their contribution to the trophic food web.

Acknowledgements

The Portuguese Foundation for Science and Technology (FCT, Portugal) financed A. M. M. Gonçalves (SFRH/BD/30475/2006) by the means of individual research grants. A special thanks to all colleagues that helped during this fieldwork. The statistical analyses were run using the ADE-4 package (Thioulouse et al., 1997). The authors are very grateful to the contributors who have made such a valuable tool available.

References

APHA (American Public Health Association). 1995. Standard Methods for the Examination of Water and Wastewater, 19th ed. American Public Health Association.

Ara K. 2001. Temporal variability and production of *Euterpina acutifrons* (Copepoda: Harpacticoida) in the Cananéia Lagoon estuarine system, São Paulo, Brazil. Hydrobiologia **453/454:** 177-187.

Azeiteiro UM, JC Marques, P Ré. 2000. Zooplankton assemblages in a shallow seasonally tidal estuary in temperate Atlantic Ocean (Western Portugal: Mondego estuary). Arq. Mus. Bocage **3:** 357–376.

Björnberg TKS. 1963. On the marine free-living copepods off Brazil. Bol. Inst. Oceanogr., Sao Paulo **13:** 3-142.

Carassou L, D Ponton. 2007. Spatio-temporal structure of pelagic larval and juvenile fish assemblages in coastal areas of New Caledonia, southwest Pacific. Mar. Biol. **150:** 697-711.

Carrasco N, I López-Flores, M. Alcaraz, MD Furones, FCJ Berthe, I Arzul. 2007. First record of a *Marteilia* parasite (Paramyxea) in zooplankton populations from a natural estuarine environment. Aquaculture **269:** 63-70.

Chinnery FE, JA Williams. 2004. The influence of temperature and salinity on *Acartia* (Copepoda: Calanoida) nauplii survival. Mar. Biol. **145:** 733–738.

Conover RJ. 1956. Oceanography and Long Island sound 1952-1954.VI. Biology of *Acartia clausi* and *A. tonsa*. Bull. Bingham Oceanogr. Collect. 156-233.

Cook KB, A Bunker, S. Hay, AG Hirst, DC Speirs. 2007. Naupliar development times and survival of the copepods *Calanus helgolandicus* and *Calanus finmarchicus* in relation to food and temperature. J. Plankton Res. **29:** 757-767.

Cunha PP, J Dinis. 2002. Sedimentary Dynamics of the Mondego Estuary. *In* MA Pardal, JC Marques, M.A. Graça, eds. Aquatic Ecology of the Mondego River Basin. Global importance of local experience.) Coimbra: Imprensa da Universidade: Coimbra pp. 43-62. (.)

David V, B Sautour, P Chardy. 2007. Successful colonization of the calanoid copepod *Acartia tonsa* in the oligo-mesohaline area of the Gironde estuary (SW France) – Natural or anthropogenic forcing? Est. Coast. Shelf Sci. **71:** 429-442.

Dias CO, AV Araujo, R Paranhos, SLC Bonecker. 2010. Vertical Copepod Assemblages (0-2300 m) off Southern Brazil. Zool. Stud. **49 (2):** 230-242.

Dolédec S, D Chessel. 1994. Co-inertia analysis: an alternative method for studying species-environment relationships. Freshw. Biol. **31:** 277-294.

Gaudy R. 1972. Contribution à la connaissance du cycle biologique des copepods du Golfe de Marseille. Etude du cycle biologique de quelques espèces caractéristiques. Tehtys **2:** 175-242.

Gonçalves AMM, M. De Troch, SC Marques, MA Pardal, UM Azeiteiro. 2010a. Spatial and temporal distribution of harpacticoid copepods in Mondego estuary. J. Mar. Biol. Ass. U. K. **90:** 1279-1290.

Gonçalves AMM, MA Pardal, SC Marques, M De Troch, UM Azeiteiro. 2010b. Distribution and composition of small-size zooplankton fraction in a temperate shallow estuary (western Portugal). Fresenius Environ. Bull. **19(12b):** 3160-3176.

Hafferssas A, R Seridji. 2010. Relationships between the hydrodynamics and changes in Copepod structure on the Algerian Coast. Zool. Stud. **49(3):** 353-366.

Hansen FC, C Möllmann, U Schütz, H.-H. Hinrichsen. 2004. Spatio-temporal distribution of *Oithona similis* in the Bornholm Basin (Central Baltic Sea). J. Plankton Res. **26:** 659-668.

Hwang J-S, R Kumar, H-U Dahms, L-C Tseng, Q-C Chen. 2010a. Interannual, seasonal, and diurnal variations in vertical and horizontal distribution patterns of 6 *Oithona* spp. (Copepoda: Cyclopoida) in the South China Sea. Zool. Stud. **49(2):** 220-229.

Hwang J-S, R Kumar, C-W Hsieh, AY Kuo, S Souissi, M-H Hsu, J-T Wu, W-C Liu, C-F Wang, Q-C Chen. 2010b. Patterns of Zooplankton distribution along the marine, estuarine, and riverine portions of the Danshuei ecosystem in northern Taiwan. Zool. Stud. **49(3):** 335-352.

Holste L, MA Peck. 2006. The effects of temperature and salinity on egg production and hatching success of Baltic *Acartia tonsa* (Copepoda: Calanoida): A laboratory investigation. Mar. Biol. **148:** 1061-1070.

Hsiao S-H, T-H Fang, C-T Shih, J-S Hwang. 2011. Effects of the Kuroshio Current on Copepod assemblages in Taiwan. Zool. Stud. **50(4):** 475-490.

Incze LS, T Ainaire. 1994. Distribution and abundance of copepod nauplii and other small (40-300 μm) zooplankton during spring in Shelikof Strait, Alaska. Fish. Bull. **92:** 67-78.

Isari S, S Psarra, P Pitta, P Mara, MO Tomprou, A Ramfos, S Somarakis, A Tselepides, C Koutsikopoulos, N Fragopoulu. 2007. Differential patterns of mesozooplankters' distribution in relation to physical and biological variables of the northeastern Aegean Sea (eastern Mediterranean). Mar. Biol. **151:** 1035-1050.

Kâ S, J-S Hwang. 2011. Mesozooplankton distribution and composition on the Northeastern Coast of Taiwan during Autumn: effects of the Kuroshio Current and Hydrothermal Vents. Zool Stud. **50(2):** 155-163.

Kršinić F, B Dubravka, R Precali, R Kraus. 2007. Quantitative variability of the copepod assemblages in the northern Adriatic Sea from 1993 to 1997. Est. Coast. Shelf Sci. **74:** 528-538.

Lam-Hoai T, D Guiral, C Rougier. 2006. Seasonal change of community structure and size spectra of zooplankton in the Kaw River estuary (French Guiana). Est. Coast. Shelf Sci. **68:** 47-61.

Limnologisk Metodik. 1992. Ferskvandsbiologisk Laboratorium. Københavns Universitet (ed.), Akademisk Forlag, København.

Marques SC, UM Azeiteiro, F Martinho, I Viegas, MA Pardal. 2009. Evaluation of estuarine mesozooplankton dynamics at a fine temporal scale: the role of seasonal, lunar, and diel cycles. J. Plankton Res. **31:** 1249-1263.

Martinho F, R Leitão, I Viegas, M Dolbeth, JM Neto, HN Cabral, MA Pardal. 2007. The influence of an extreme drought event in the fish community of a southern Europe temperate estuary. Est. Coast. Shelf Sci. **75:** 537-546.

McAlice BJ. 1981. On the post-glacial history of *Acartia tonsa* (Copepoda: Calanoida) in the Gulf of Maine and the Gulf of St. Lawrence. Mar. Biol. **64:** 267-272.

Mendes S, MJ Fernández-Gómez, P Resende, MJ Pereira, MP Galindo-Villardón, UM Azeiteiro. 2009. Spatio-temporal structure of diatom assemblages in a temperate estuary. A STATICO analysis. Est. Coast. Shelf Sci. **84:** 637-644.

Morgado F, C Quintaneiro, E. Rodrigues, MR Pastorinho, P Bacelar-Nicolau, L Vieira, UM Azeiteiro. 2007. Composition of the trophic structure of

Zooplankton in a shallow temperate estuary (Mondego estuary, Western Portugal). Zool. Stud. **46(1):** 57:68.

Mouny P, J-C Dauvin. 2002. Environmental control of mesozooplankton community structure in Seine estuary (English Channel). Oceanol. Acta **25:** 13–22.

Parsons TR, Y Maita, CM Lally, eds. 1985. A Manual of Chemical and Biological Methods for Seawater Analysis. Pergamon Press: Oxford.

Portuguese Water Institute (Instituto da Água). 2008. http://snirh.inag.pt [Accessed October 2010].

Portuguese Weather Institute (Instituto de Meteorologia). 2008. http://web.meteo.pt [Accessed October 2010].

Primo AL, UM Azeiteiro, SC Marques, F Martinho, MA Pardal. 2009. Changes in zooplankton diversity and distribution pattern under varying precipitation regimes in a southern temperate estuary. Est. Coast. Shelf Sci. **82:** 341 – 347.

Richmond EC, DS Wethey SA Woodin. 2007. Climate change and increased environmental variability: demographic responses in an estuarine harpacticoid copepod. Ecol. Model. **209:** 189-202.

Simier M, L Blanc, F Pellegrin, D Nandris. 1999. Approche simultanée de K couples de tableaux : apllication à l'étude des relations pathologie végétale-environnement. Rev. de Stat. Appl. **47:** 31-46.

Simier, M., Laurent, C., Ecoutin, J.-M., and Albaret, J.-J. 2006. The Gambia River estuary: a reference point for estuarine fish assemblages studies in West Africa. Est. Coast. Shelf Sci. **69:** 615-628.

Strickland JDH, TR Parsons. 1972. "A Practical Handbook of Seawater Analysis." 2nd edn. Fish. Res. Board Canada Bull. **167:** 1-311.

Tackx MLM, N Pauw, R van Mieghem, F Azémar, A Hannouti, S van Damme, F Fiers, N Daro, P Meire. 2004. Zooplankton in the Schelde estuary, Belgium and the Netherlands. Spatial and temporal patterns. J. Plankton Res. **26:**133-141.

Thioulouse J, D Chessel. 1987. Les analyses multitableaux en écologie factorielle. I. De la typologie d'état à la typologie de fonctionnement par l'analyse triadique. Acta Oecol-Oec Gen **8:** 463-480.

Thioulouse J, D Chessel, S Dolédec, JM Olivier. 1997. ADE-4: a multivariate analysis and graphical display software. Statistics and Computing **7:** 75-83.

Thioulouse J, M Simier, D Chessel. 2004. Simultaneous analysis of a sequence of paired ecological tables. Ecology **85:** 272-283.

Uriarte I, F Villate. 2005. Differences in the abundance and distribution of copepods in two estuaries of the Basque coast (Bay of Biscay) in relation to pollution. J. Plankton Res. **27:** 863-874.

Villate F, A Ruiz, J Franco. 1993. Summer zonation and development of zooplankton populations within a shallow mesotidal system: the estuary of Mundaka. Cah. Biol. Mar. **34:** 131-143.

Villate F, I Uriarte, X Irigoien, G Beaugrand, U Cotano. 2004. *In* Á. Borja, M. Collins, eds. Zooplankton communities.. Oceanography and Marine Environment of the Basque Country. Amsterdam: Elsevier Oceanography Series, pp. 395-423. (.)

ADE-4 software available at http://pbil.univ-lyon1.fr/ADE-4.

Figure captions

Figure 1 – Location of the Mondego estuary on the western coast of Portugal and the five sampling stations within the estuary: (M – mouth station, N1 and N2 – northern arm stations, S1 and S2 – southern arm stations)

Figure 2 – Three –way data structure: for environmental parameters and species densities at 5 sampling stations

Figure 3 – (A) Monthly precipitation (mm) in Mondego estuary during the study period (from 2005 to 2007). Grey filled squares represent monthly average of 1971 to 2000 (http://snirh.inag.pt) and (B) monthly average of salinity and water temperature (ºC) during the study period (2005-2006 and 2006-2007)

Figure 4 – Interstructure factor map of the STATICO analysis on the Mondego estuary data. (A) This map shows the importance of each sampling station in the compromise (M – mouth station; N1 and N2 – stations of the north arm; S1 and S2 – stations of the south arm). The scales are given in the boxes

Figure 5 - Compromise factor map of the STATICO analysis of the copepod species variables (A) and environmental parameters (B). (C) Eigenvalues diagram. The scales for axes are given in the boxes. Chl *a* – chlorophyll *a*, Temp – water temperature, Sal – salinity, O2- dissolved oxygen, TSS – total suspended solids, Transp –

transparency,PO4 – phosphates; NO2 – nitrites, NO3 – nitrates, SiO2 – silica, NH4 – ammonia. See table 2 for species codes

Figure 6 – Trajectories factor maps of the STATICO analysis: (A) projection of the average position of copepod species at each sampling station and (B) projection of the environmental variables at each sampling station (M – mouth station; N1 and N2 – stations of the north arm; S1 and S2 – stations of the south arm). The scales for axes are given in the boxes. For species codes see Table 2. See figure 5 for environmental variables codes

Figure 7 - Trajectories factor plots of the STATICO analysis: projection of the seasons along the study period (2005, 2006 and 2007) in terms of both environmental and copepod structure. Graphs are given for each sampling site (M – mouth station; N1 and N2 – stations of the north arm; S1 and S2 – stations of the south arm) that showed the highest contribution to the co-structure between environmental factors and copepod abundances. Each sample is represented by two points: one is the projection of the row of the species table (circle: origin of arrows), and the other is the projection of the row of the environmental table (end of arrows). The length of the connecting line reveals the disagreement or the consensus between the two profiles (species–environment), i.e., the length of the line is proportional to the divergence between the datasets. When the datasets agree very strongly, the arrows will be short. Likewise, a long arrow demonstrates a locally weak relationship between the environment and copepod features for that case. The scales for axes are given in the boxes

Tables

Table 1 – Typological value indices: RV = Correlation matrix: contains the cosines between tables; Weights = weights of tables in the compromise; Cos^2 = Square cosine between table and approximated compromise

Sampling sites	RV					Weights	Cos^2
M	1					0.55	0.71
N1	0.72	1				0.55	0.81
N2	0.60	0.65	1			0.37	0.54
S1	0.59	0.77	0.49	1		0.46	0.64
S2	0.32	0.41	0.43	0.44	1	0.19	0.24

Table 2 – List of species (including life stages) and their abbreviations used in multivariate analysis (STATICO). Average of copepod densities during the study period is recorded

Taxa		Abbreviatiaon	Habitat	Average
	Nauplii Copepoda	Naco		3.57E+05 (0-5.61E+07)
	Copepodite not identified	Cpni		2.60E+03 (0-4.09E+05)
Calanoida				
Acartiidae	Copepodite *Acartia* sp.	CpA		1.94E+03 (0-3.04E+05)
	Copepodite *Acartia clausi*	CpAcl	Marine	9.94E+02 (0-1.56E+05)
	Acartia clausi	Acl	Marine	8.45E+01 (0-1.33E+04)
	Copepodite *Acartia tonsa*	CpAto	Estuarine	4.30E+03 (0-6.74E+05
	Acartia tonsa	Ato	Estuarine	1.93E+02 (0-3.02E+04)
Temoridae	Copepodite *Temora longicornis*	CpTelo	Marine	1.05E+03 (0-1.64E+05)
	Temora longicornis	Telo	Marine	7.84E+01 (0-1.23E+04)
Centropagidae	Copepodite *Centropages* sp.	CpCen		6.46E+02 (0-1.01E+05)
	Centropages sp.	Cen		9.58E-01 (0-1.50E+02)
Clausocalanidae	Copepodite *Clausocalanus* sp.	CpCl		2.19E+02 (0-3.43E+04)
	Copepodite *Clausocalanus arcuicornis*	CpClar	Marine	5.65E+02 (0-8.88E+04)
	Clausocalanus arcuicornis	Clar	Marine	7.68E+00 (0-1.21E+03)
	Pseudocalanus elongatus	Psel	Marine	3.11E-02 (0-4.88E+00)
Paracalanidae	*Paracalanus parvus*	Ppa	Marine	6.64E+01 (0-1.04E+04)
	Copepodite of *Paracalanus*-type form	CpPspacl		2.22E+03 (0-3.49E+05)
Cyclopoida				
Oithonidae	Copepodite *Oithona* sp.	CpOi		1.46E+04 (0-2.30E+06)
	Copepodite *Oithona plumifera*	CpOipl	Marine	2.62E+04 (0-4.12E+06)
	Oithona plumifera	Oipl	Marine	4.10E+02 (0-6.44E+04)
Cyclopidae	Copepodite *Acanthocyclops robustus*	CpAcro	Freshwater	1.67E+03 (0-2.62E+05)
	Acanthocyclops robustus	Acro	Freshwater	1.20E+00 (0-1.89E+02)
Harpacticoida				
Euterpinidae	Copepodite *Euterpina acutifrons*	CpEuac	Estuarine	9.59E+03 (0-1.51E+06)
	Euterpina acutifrons	Euac	Estuarine	3.92E+03 (0-6.16E+05)
Laophontidae	Copepodite *Paronychocamptus nanus*	CpPana	Marinee	1.10E+03 (0-1.73E+05)
	Paronychocamptus nanus	Pana	Marine	3.59E+02 (0-5.63E+04)
Poecilostomatoida				

Sapphirinidae	*Sapphirina* sp.	Sapp	9.10E-03 (0-1.43E+00)

Figures

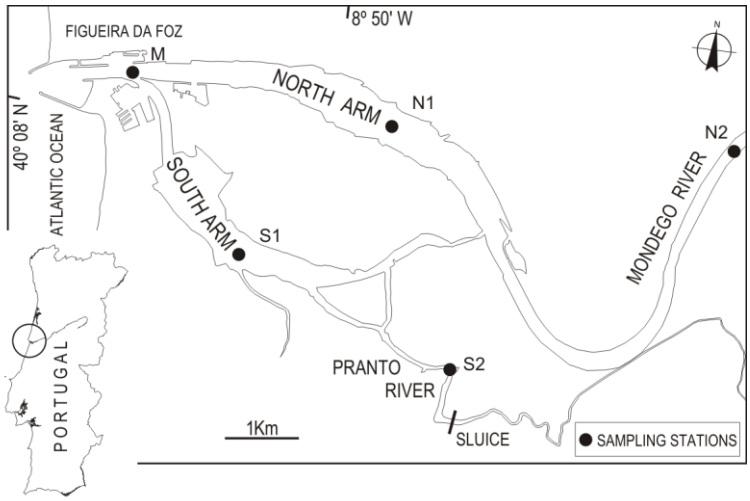


Figure 1

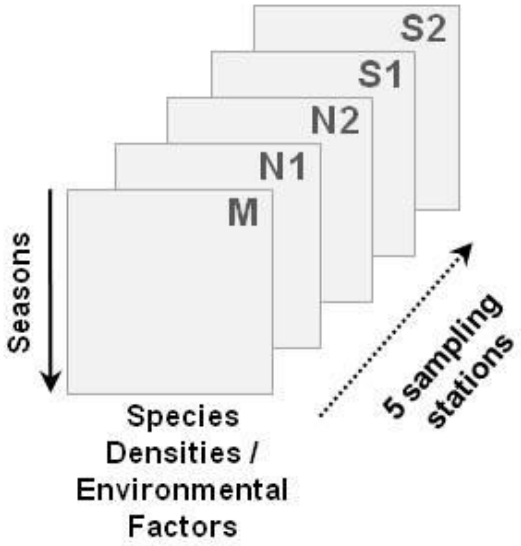

S2
S1
N2
N1
M
Seasons
5 sampling stations
Species Densities / Environmental Factors

Figure 2

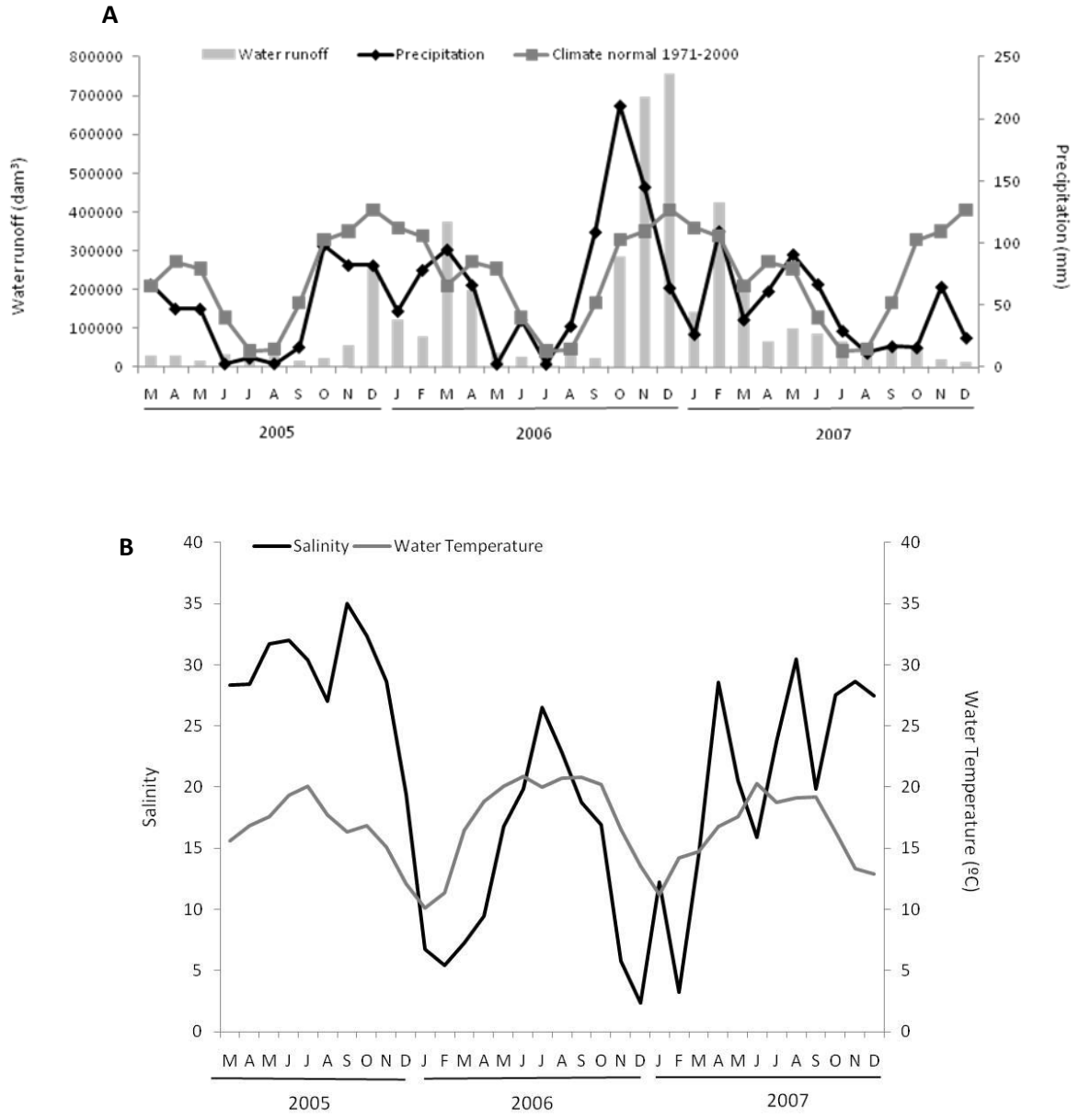

A
Water runoff
Precipitation
Climate normal 1971-2000
Water runoff (dam³)
Precipitation (mm)
800000
700000
600000
500000
400000
300000
200000
100000
0
250
200
150
100
50
0
M A M J J A S O N D J F M A M J J A S O N D J F M A M J J A S O N D
2005
2006
2007
B
Salinity
Water Temperature
Salinity
Water Temperature (ºC)
40
35
30
25
20
15
10
5
0
M A M J J A S O N D J F M A M J J A S O N D J F M A M J J A S O N D
2005
2006
2007

Figure 3

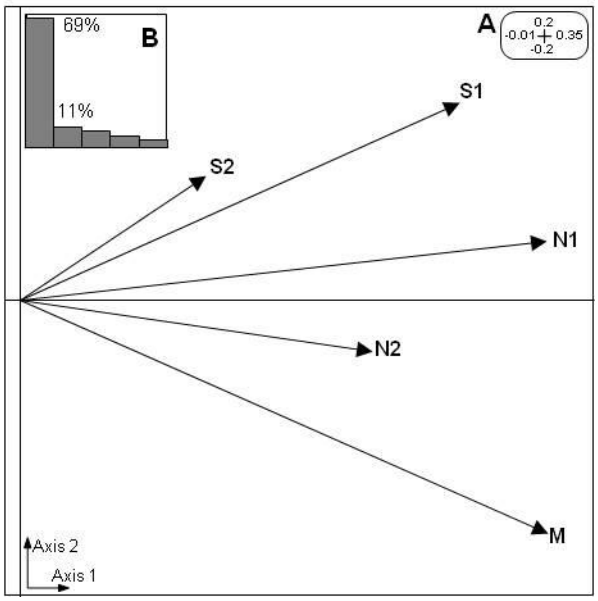


Figure 4

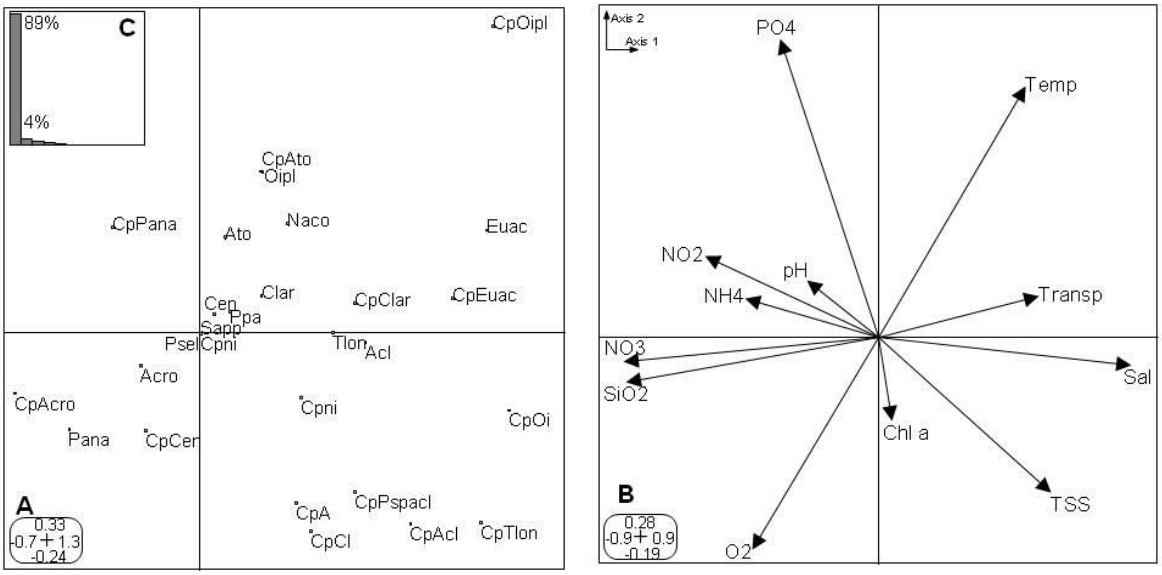

89%
C
4%
CpOipl
CpAto
Oipl
CpPana
Ato
Naco
Euac
Clar
CpClar
CpEuac
Cen
Ppa
Sapp
Psel
Cpni
Tlon
Acl
Acro
CpAcro
Cpni
CpOi
Pana
CpCer
A
0.33
-0.7
1.3
-0.24
CpPspacl
CpA
CpCl
CpAcl
CpTlon
Axis 2
Axis 1
PO4
Temp
NO2
pH
NH4
Transp
NO3
Sal
SiO2
Chl a
TSS
B
0.28
-0.9
0.9
-0.19
O2

Figure 5

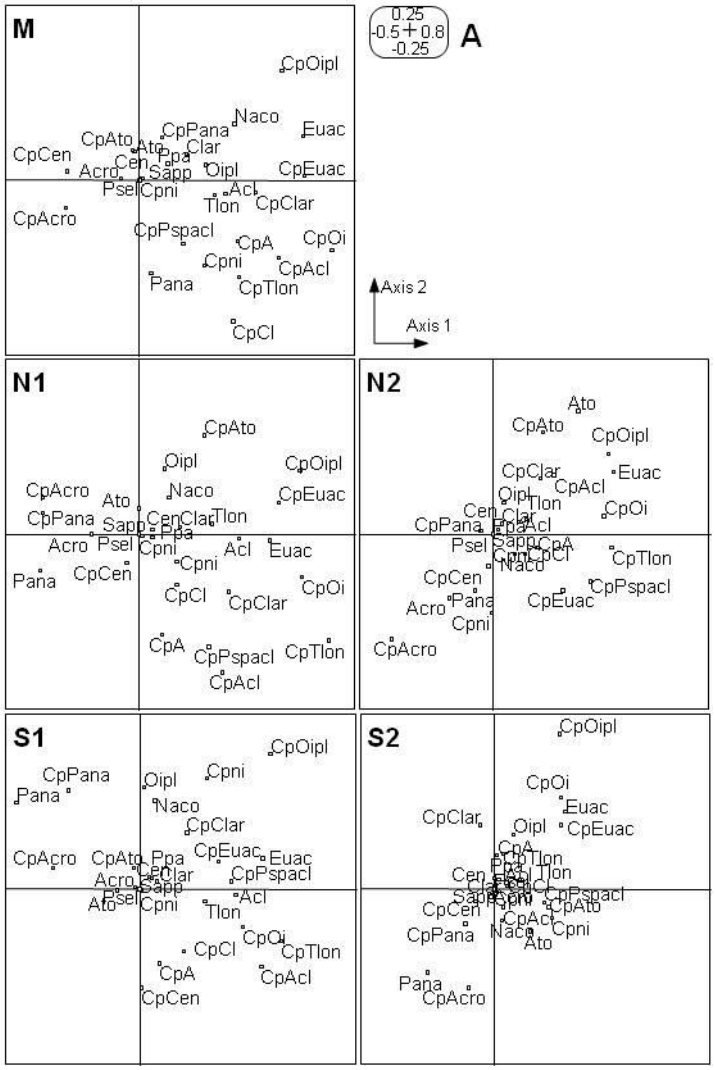

M
0.25
-0.5 0.8
-0.25
A
CpOipl
Naco
CpPana
Euac
CpAto
Ato
Clar
CpCen
Ppa
Cen
Acro
Sapp
Oipl
CpEuac
Psel
Cpni
Acl
Tlon
CpClar
CpAcro
CpPspacl
CpA
CpOi
Cpni
CpAcl
Pana
CpTlon
CpCl
Axis 2
Axis 1
N1
CpAto
Oipl
CpOipl
CpAcro
Naco
CpEuac
Ato
CpPana
Sapp
Cen
Clar
Tlon
Ppa
Acro
Psel
Cpni
Acl
Euac
Cpni
Pana
CpCen
CpCl
CpClar
CpOi
CpA
CpPspacl
CpTlon
CpAcl
N2
Ato
CpAto
CpOipl
CpClar
Euac
CpAcl
Oipl
Tlon
CpOi
Cen
Clar
CpPana
Ppa
Acl
Psel
Sapp
CpA
Cpni
CpCl
CpTlon
Naco
CpCen
CpPspacl
Acro
Pana
CpEuac
Cpni
CpAcro
S1
CpOipl
CpPana
Oipl
Cpni
Pana
Naco
CpClar
CpAcro
CpAto
Ppa
CpEuac
Euac
Acro
Cen
Clar
CpPspacl
Psel
Sapp
Ato
Cpni
Acl
Tlon
CpCl
CpOi
CpTlon
CpA
CpAcl
CpCen
S2
CpOipl
CpOi
Euac
CpClar
Oipl
CpEuac
CpA
CpTlon
Cen
Tlon
CpPspacl
CpCen
CpAto
CpAcl
CpPana
Naco
Cpni
Ato
Pana
CpAcro

Figure 6

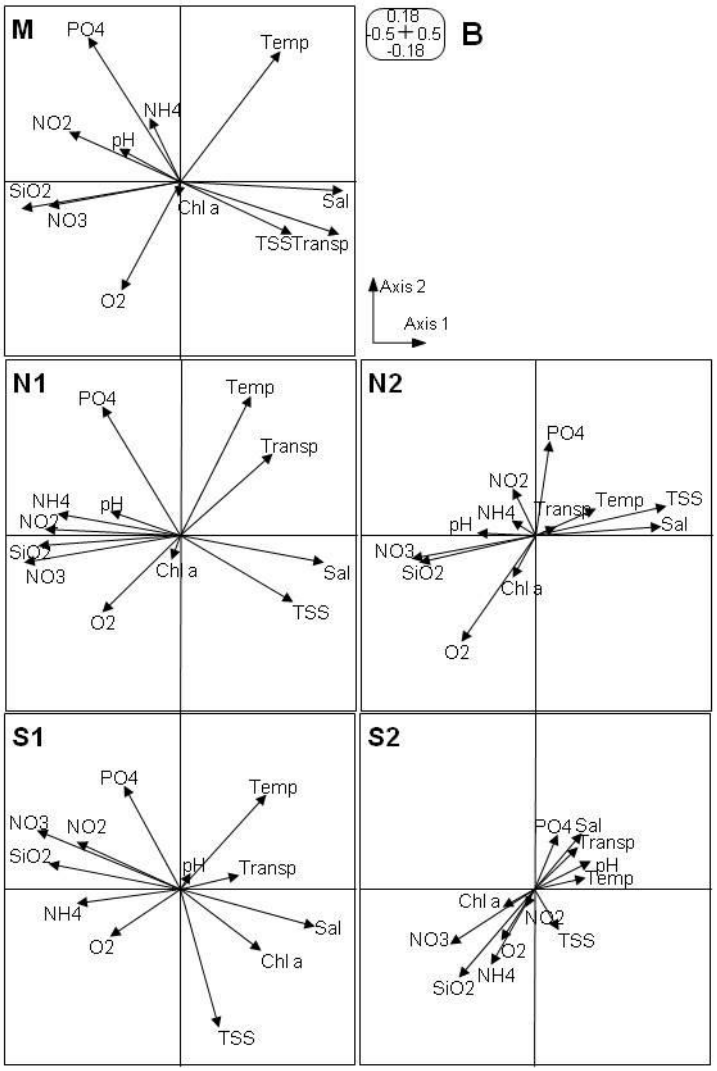


Figure 6 (continued)

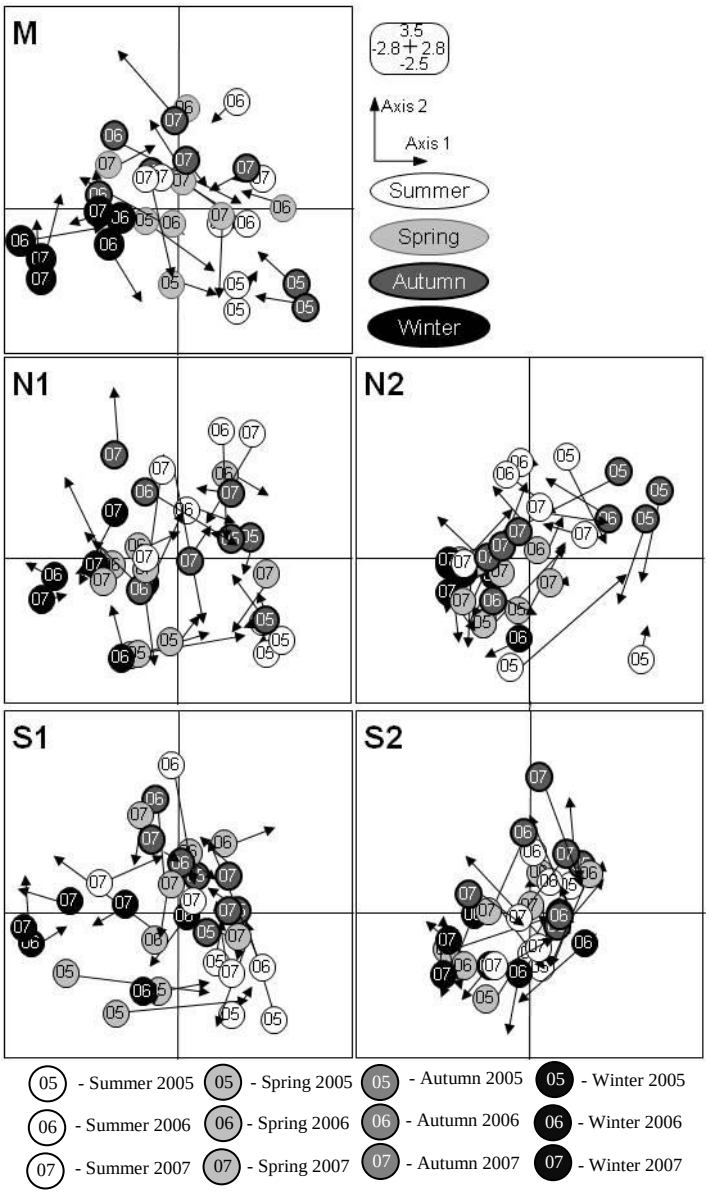

M
N1
N2
S1
S2
3.5
-2.8 + 2.8
-2.5
Axis 2
Axis 1
Summer
Spring
Autumn
Winter
05 - Summer 2005
05 - Spring 2005
05 - Autumn 2005
05 - Winter 2005
06 - Summer 2006
06 - Spring 2006
06 - Autumn 2006
06 - Winter 2006
07 - Summer 2007
07 - Spring 2007
07 - Autumn 2007
07 - Winter 2007

Figure 7

SUMMARY AND CONCLUSIONS IN SPANISH

SUMMARY AND CONCLUSIONS IN SPANISH

RESUMEN

INTRODUCCIÓN

El análisis espacio-temporal de cualquier ecosistema a través de las interacciones "especies-ambiente" conduce a la obtención de un complejo conjunto de datos de naturaleza necesariamente multivariante, ya que mediante un esquema espacial de toma de muestras en diferentes lugares (sitios) a lo largo del tiempo, se obtiene un conjunto de datos tridimensionales: los lugares, las variables medias en dichos lugares (bien sean bióticas como por ejemplo la abundancia o presencia/ausencia de determinadas especies) y/o abióticas (variables ambientales) y el tiempo. El interés creciente, en las últimas décadas, por este tipo de estudios, ha contribuido de notablemente al desarrollo muchos métodos de análisis de datos multivariantes.

En numerosos estudios, los investigadores organizan este tipo de datos en dos tablas[14]: (1) la que contiene información sobre la composición de la comunidad (por ejemplo número de organismos, su presencia/ausencia o un valor de abundancia de especies, en determinados lugares) y (2) la que contiene información sobre las variables ambientales que afectan a la distribución de las especies, registradas en esos mismos lugares, llamadas "par de tablas ecológicas", y abordan el análisis simultáneo de estas dos tablas mediante el Análisis Canónico de Correspondencias (CCA, ter Braak, 1986), que es la técnica más ampliamente utilizada.

[14] A lo largo de este trabajo, los términos "matriz" y "tabla" se utilizarán con el mismo significado.

Pero en realidad, cuando el investigador está interesado en los cambios que se producen en las relaciones entre especies y ambiente y el muestreo se repite a lo largo del tiempo o a través del espacio, se obtiene, en realidad, una secuencia de tablas, también conocida como k-tabla o cubo de datos. Cuando un par de tablas ecológicas se repite, el resultado es un par de k-tablas o dos cubos de datos.

Si el análisis de esos dos cubos de datos se aborda desde la perspectiva de construir dos tablas (una a partir de cada cubo, situando en cada uno de ellos las tablas de la serie una tras otra) y analizarlas mediante un CCA por ejemplo, la variabilidad temporal y la espacial se mezclan. Ésta es, en realidad, una consecuencia que resulta de la forma en cómo se organizan las tablas de datos en el análisis. Por tanto el estudio de los gradientes temporales y espaciales no siempre se realiza, en estos casos, de manera óptima, puesto que serán los efectos preponderantes los que emergerán en el análisis, lo que puede convertirse en una desventaja si no es ese precisamente el efecto en el que el investigador ha fijado su objetivo.

En este sentido, estudiar las escalas temporales y espaciales de los ecosistemas, se convierte en un reto para los investigadores. Esto implica el estudio de tres dimensiones: el tiempo, el espacio y las variables (y o especies), en otras palabras, el estudio de un cubo o de dos cubos de datos. El análisis óptimo de este conjunto de datos es abordado, desde el punto de vista del análisis multivariante, mediante métodos de análisis de tablas de tres vías. Esta potencialidad permite el análisis de la estructura espacial y su estabilidad o variabilidad a lo largo del tiempo, o estudiar la estructura temporal y su estabilidad o variabilidad espacial, así como analizar la estabilidad de las relaciones especies ambiente a lo largo del tiempo o espacio.

En realidad, los métodos de análisis para tablas de tres vías (tiempo, espacio y variables), son una alternativa que va cobrando cada vez mayor interés entre los investigadores pero su uso sigue siendo aún minoritario en comparación con el del CCA, sin duda por la dificultad que presenta el estudio de estos procedimientos para aquellos usuarios que no suelen dominar el lenguaje algebraico subyacente a dichas técnicas. Sin embargo, el poder separar no sólo el efecto de las tres dimensiones, sino el evaluar el efecto de la interacción entre ellas en la dinámica de las relaciones entre especies y variables ambientales, nos demuestra su interés en este campo.

OBJETIVOS

Los objetivos de esta investigación son:

(1) Hacer una revisión de los métodos clásicos para el análisis de datos organizados una o dos matrices de datos: Análisis de Correspondencias, Análisis HJ-Biplot y Análisis Canónico de Correspondencias;

(2) Hacer una revisión exhaustiva de las propiedades de los métodos de la familia STATIS de la escuela francesa, utilizados para estudiar la estructura consenso de varias configuraciones;

(3) Realizar el estudio comparativo de estos métodos con el modelo Tucker3 de la escuela holandesa/inglesa, que es en la actualidad, el más utilizado para estudiar la dinámica de las estructuras (análisis de la parte no estable debida a posibles interacciones);

(4) Realizar una búsqueda exhaustiva de la bibliografía especializada en el campo de la Ecología para poder conocer el impacto de estos métodos en ese campo sobre todo en los últimos años;

(5) Proponer, como parte central de este trabajo, un nuevo método de análisis multivariante que permita resolver el problema de describir no sólo la parte estable de las relaciones entre dos conjuntos de datos (por ejemplo, especies y variables ambientales) en diferentes condiciones (por ejemplo, lugares y/o en diferentes tiempos), sino también el cambio (parte dinámica);

(6) Aplicar la nueva propuesta a un conjunto de datos reales, organizados en tablas de tres vías (objetos[15] × variables × condiciones) y comparar los resultados con los encontrados con las técnicas clásicas;

(7) Crear un marco metodológico general, que incluya las distintas técnicas de estudio de tablas de tres vías desde la perspectiva Biplot, y desarrollar un método específico que facilite el uso de estas técnicas a los investigadores aplicados.

ESTRUCTURA DE LOS CONTENIDOS

La parte principal de la tesis se divide en seis capítulos.

El **CAPÍTULO UNO** se inicia con una breve revisión y discusión de algunas cuestiones relacionadas con el uso de análisis indirecto del gradiente, en particular nos centramos en una de las técnicas más comúnmente utilizadas cuando sólo se tiene información sobre la composición de la comunidad: el Análisis de Correspondencias (CA, Benzécri,

[15] A lo largo de este trabajo, los términos "sujeto" y "objeto" se entenderá que tienen el mismo significado.

1973). A continuación, una breve introducción a la teoría de los métodos Biplot (Gabriel, 1971), en particular centrándonos en el análisis HJ-Biplot (Galindo, 1986). Se pone de manifiesto cómo el HJ-Biplot es una herramienta útil para encontrar patrones subyacentes en una matriz de datos, con ventajas similares a las que presenta el CA, pero mucho más generales, puesto que no presenta la exigencia del CA de trabajar con matrices de datos positivos, con lo cual puede utilizarse para analizar no sólo los datos de la composición de la comunidad sino también matrices de medidas asociadas al hábitat.

En el apéndice de la tesis se incluye el artículo publicado en 2009 en la revista internacional ARQUIPELAGO LIFE AND MARINE SCIENCES revisada por pares (también se presentan los indicios de calidad de la misma), y que forma parte del trabajo realizado para la elaboración de esta memoria: *"Bacterioplankton dynamics in the Berlengas Archipelago (West coast of Portugal) using the HJ-biplot method"*. Se trata de una aplicación a datos reales en el cual se enfatiza tanto la utilidad del HJ-Biplot como herramienta de análisis de datos ecológicos como la correcta interpretación de sus resultados.

El **CAPÍTULO DOS** describe uno de los métodos más eficaces para el estudio de las relaciones especies-ambiente, el Análisis Canónico de Correspondencias (CCA, ter Braak, 1986). Además, se describen las principales ventajas y desventajas. También se presenta una descripción concisa del Análisis de la Co-Inercia (Dolédec y Chessel, 1994), teniendo en cuenta el hecho de que es una alternativa simple y robusta al CCA cuando el número de lugares de muestreo es bajo en comparación con el número de variables explicativas

(variables ambientales). Con el fin de entender mejor todas las interpretaciones en el uso de estos métodos, se presentan algunos ejemplos con la respectiva explicación.

El **CAPÍTULO TRES** comienza con una revisión de los métodos multi-tablas procedentes de la escuela francesa, es decir, se estudian los métodos de la familia STATIS (Structuration des Tableaux a Trois Indices de la Statistique, L´Hermier des Plantes, 1976; Lavit, 1988, 1994). En particular se centra en describir el Análisis Triádico Parcial (PTA, Jaffrenou, 1978; Thioulouse y Chessel, 1987; Blanc et al., 1998) que nos permite el análisis de un cubo de datos (bien esté formado por varias matrices de datos de lugares x especies (o variables ambientales) en diferentes tiempos, o bien tiempos × especies (o variables ambientales en distintos lugares), y el método STATICO (Simier et al. 1999; Thioulouse et al. 2004) que nos permite el análisis de dos cubos de datos: uno relativo a la composición de la comunidad y otro de variables ambientales, y que nos va a permitir conocer la estabilidad (bien temporal o espacial) de las relaciones especies-ambiente, ya que las bases metodológicas y objetivos de ambos métodos están relacionados con el nuevo método propuesto en el Capítulo 5.

Con respeto al método PTA, se presentan en el apéndice de la memoria dos artículos ya publicados y que han sido también resultado del trabajo de elaboración de esta tesis: *"The efficiency of Partial Triadic Analysis method: an ecological application"* y "*Zooplankton distribution in a Marine Protected Area: the Berlengas Natural Reserve (Western coast of Portugal"*) publicados en 2010 y 2011 en BIOMETRICAL LETTERS y FRESENIUS ENVIRONMENTAL BULLETIN**,** respetivamente, ambas revisadas por pares, y cuyos indicios de calidad también se acompañan como información.

De igual modo, y como resultado de elaboración de esta tesis, en referencia al estudio y aplicación del método STATICO, se presentan en el apéndice cuatro publicaciones, con los respectivos indicios de calidad de las revistas en las que se han publicado:

"Spatio-temporal structure of diatom assemblages in a temperate estuary", publicada en 2009 en ESTUARINE, COASTAL AND SHELF SCIENCE,

"Diel vertical behavior of Copepoda community (naupliar, copepodites and adults) at the boundary of a temperate estuary and coastal waters" publicado "First on line" en noviembre 2011 en ESTUARINE, COASTAL AND SHELF SCIENCE, doi: 10.1016/j.ecss.2011.11.018,

"Response to Climatic variability of Copepoda life history stages in a southern European temperate estuary" aceptada para su publicación en octubre de 2011 en la revista ZOOLOGICAL STUDIES,

"*An empirical comparison of Canonical Correspondence Analysis and STATICO in identification of spatio-temporal ecological relationships"*) publicado "First on line" en noviembre de 2011 el JOURNAL APPLIED STATISTICS, http://dx.doi.org/10.1080/02664763.2011.634393.

En el **CAPÍTULO CUATRO** se presenta una revisión de los métodos de tres vías con especial énfasis en la terminología que caracteriza a estas técnicas. A continuación, se presentan consideraciones generales sobre el modelo Tucker3 (Tucker, 1966) (de la escuela holandesa/inglesa) para el tratamiento de tablas de tres vías/tres modos.

En el **CAPÍTULO CINCO** se presenta la nueva propuesta, la cual captura las ventajas del STATICO y del modelo Tucker3 y permite el análisis integrado de ambos métodos para el estudio, no solo de las configuraciones consenso, sino también de las interacciones y los cambios asociados a la dinámica de los ecosistemas.

A lo largo del capítulo, se presta especial atención a las ventajas del modelo propuesto CO-TUCKER con respecto a la posibilidad de extraer estructuras ocultas y la captura de las correlaciones entre las variables subyacentes en una matriz de tres vías. Además, se resaltan aquellos aspectos relacionados con el procesamiento de datos, las estrategias en la selección de componentes y la evaluación de la calidad del ajuste de los niveles de un modo. Se realiza un detallado análisis de la matriz *core* y de su flexibilidad a la hora de explicar la interacción entre una componente de un modo con cualquier componente de los otros modos. Además, se exploran, prestando especial atención a su interpretación, los biplots conjuntos, cuyo objetivo consiste en captar las estructuras multilineales en las bases de datos (de orden superior).

La nueva técnica propuesta, CO-TUCKER, se utiliza para analizar un conjunto de datos reales con el objetivo de presentar y discutir de forma pormenorizada los resultados y las ventajas inherentes a la misma. La parte final de este capítulo compara dichos resultados con los obtenidos del análisis del mismo conjunto de datos reales mediante el método STATICO, en particular analizando y discutiendo las ventajas y desventajas de ambos.

Finalmente se presentan las conclusiones del trabajo de tesis realizado así como las perspectivas del trabajo futuro y las referencias bibliográficas.

ASPECTOS MÁS RELEVANTES DEL TRABAJO

La forma típica de presentar los datos para el análisis estadístico es formando una tabla de dos vías; es decir, en una matriz X de dimensión $(I \times J)$ donde I representa las unidades de muestreo/objetos/filas (individuos, estaciones, lugares, etc.) y J el número de variables/columnas (o especies, variables ambientales, etc.). Existen muchas técnicas en la literatura para capturar la información recogida en la matriz entre las que cabe destacar por su uso el CA cuando los datos son datos de frecuencias o de cobertura y los Métodos Biplot cuando los datos son cuantitativos. En aquellos casos en los que el investigador está interesado en conocer de forma directa la influencia de las variables ambientales en la distribución de las especies, la técnica más utilizada, a pesar de presentar algunas limitaciones, es el CCA. Otra forma de evaluar la relación entre las dos matrices (la de composición de la comunidad y la de medidas asociadas al hábitat) es mediante el Análisis de la Co-Inercia. Esta técnica busca las direcciones de máxima variabilidad compartida por ambas matrices de datos, en lugar de buscar las direcciones de máxima inercia en cada matriz.

Sin embargo, si en lugar de integrar la información de dos matrices de datos necesitamos integrar tres o más (por ejemplo, cada una cruza lugares por especies, o por variables, y se dispone de k matrices correspondientes a k tiempos o a k situaciones experimentales), es posible utilizar técnicas de la familia STATIS. El método STATIS (así

como el STATIS DUAL) consiste en un procedimiento de tres etapas que permite representar cada matriz como un punto en un espacio vectorial de dimensión reducida. La proximidad de dos puntos (basándose en aproximaciones geométricas a espacios euclídeos) se traduce en la existencia de una estructura de los individuos común a las matrices. Esas estructuras comunes son reflejo de los patrones de similitud entre los lugares a lo largo del tiempo, y/o el patrón temporal de covariación entre las variables. Los métodos STATIS presentan limitaciones importantes entre las que podemos destacar las siguientes: no contemplan estructura de grupos, comparan solamente direcciones de máxima inercia y trabajan sólo con variables cuantitativas.

El PTA, que si bien es menos general que el STATIS o el STATIS DUAL, permite trabajar con datos de tres vías estructurados según cubos de datos, en los que las entidades en filas son las mismas en todas las tablas y las entidades en columnas también son las mismas en todas ellas, es decir, las dos dimensiones de cada tabla son comunes a todas ellas. A pesar de presentar la condición/limitación de que verdaderamente se trate de un "cubo de datos", tiene la ventaja de trabajar sobre los datos originales directamente, y no sobre operadores como hace el STATIS (STATIS DUAL) lo que permite interpretaciones más directas.

Con el Análisis Factorial Múltiple (MFA) propuesto por Escofier y Pagès (1984), es posible trabajar con variables de varios tipos.

Simier et al. (1999), proponen, con el objetivo de encontrar la parte estable en la dinámica de las relaciones entre las especies y su ambiente, el método STATICO que consiste en analizar mediante un STATIS (o un PTA) el cubo de operadores de co-inercia

resultante de los análisis de Co-Inercia para cada "par de tablas ecológicas". Esta técnica fue introducida por vez primera en Ecología por Thiolouse y Chessel (1987), quienes utilizan el PTA para el estudio simultáneo de una secuencia de pares de tablas ecológicas.

Kroonenberg propone en 1989 el Análisis Triádico Completo, basado en el análisis de componentes principales de tres vías o tres modos, constituyendo una generalización del modelo de Tucker (1966).

Tucker resume los datos a través de componentes para los tres modos, es decir, propone un modelo para el análisis de componentes principales de tres modos, en el que se contempla la reducción de la dimensionalidad en cada uno de ellos. Estos métodos ofrecen marcadores para los niveles de los factores o modos, lo que facilita la interpretación de los resultados en términos de representaciones Biplot. Los elementos y observaciones son clasificados de acuerdo a las categorías del modo individuo (o lugar, o estación, etc.), del modo variable y del modo ocasión (o condición experimental). Cada dato está relacionado con una categoría de cada uno de los modos. Con estos modelos es posible explicar las interacciones de tercer orden. Además, es posible obtener la matriz *core* que relaciona las componentes para los tres modos y trata las inter y intra relaciones con la misma atención. La estimación del modelo se realiza mediante un proceso iterativo alternado (o condicional).

Esta solución es propuesta por Kroonenberg y De Leeuw (1980) a través del algoritmo de TUCKALS3, es aplicable a cualquier conjunto de datos de tres vías, puesto que considera diferentes ejes en cada modo. Se basa en proyecciones simultáneas de tres

nubes de puntos. La ventaja fundamental del modelo es la de considerar la relación entre las componentes de los diferentes modos, a diferencia de la mayor parte de los métodos de integración de matrices que tratan de encontrar una configuración consenso, y que en muchos casos es imposible debido a la estructura de covariación tan diferente presente en las distintas matrices que deseamos integrar.

Como consecuencia de la solución propuesta para encontrar los estimadores de las matrices del modelo, es posible investigar la variabilidad para cada individuo (o lugar) y variable (o especie) a lo largo del tiempo. En resumen, podemos citar los métodos de TUCKALS (Kroonenberg, 1994) por el ajuste que hace en los modelos. La matriz *core* contiene las relaciones entre los factores o componentes de cada modo. Pondera la combinación de componentes caracterizada por la p, q y r categoría del primer modo, segundo y tercer modo, respectivamente. Así, a través de cada elemento de la matriz *core* es posible cuantificar el porcentaje de la inercia explicada por el análisis que es atribuido a una combinación concreta de componentes.

Las comparaciones entre los métodos para análisis de componentes principales de tres modos, STATIS y otros métodos para analizar datos de tres vías se pueden encontrar en los trabajos ten Berge (1987), Lavit (1988), Kiers (1988), Carlier et al. (1989) y Kroonenberg (2008).

MÉTODO PROPUESTO: CO-TUCKER

- **LOS DATOS: TERMINOLOGÍA Y NOTACIÓN**

En esta investigación se trabaja con matrices/datos de tres vías debido a que el objetivo es desarrollar una técnica basada en modelos de tres modos. En estos modelos los elementos u observaciones son clasificados de acuerdo a las categorías de tres modos: individuo u objetos (I), variable (J) y ocasión o condición (K). Cada dato está relacionado con una categoría del primer modo A (individuo u objetos), una categoría del segundo modo B (variable) y una categoría del tercer modo C (ocasión o condición). Así, los datos de tres vías se pueden definir cuando existe un solo conjunto de individuos (u objetos), un solo conjunto de variables y un solo conjunto de ocasiones (o condiciones). Es decir, la información queda recogida en K matrices de orden $I \times J$; siendo I, J y K la dimensión de cada modo.

- CO-TUCKER

El método propuesto CO-TUCKER combina el método Tucker3 y los k Análisis de la Co-Inercia de los k-pares de tablas ecológicas. De hecho, se beneficia de las ventajas de los métodos STATICO (basado en el Análisis de la Co-Inercia) y el modelo Tucker3.

En la figura 1, se presenta un esquema general del método propuesto, CO-TUCKER.

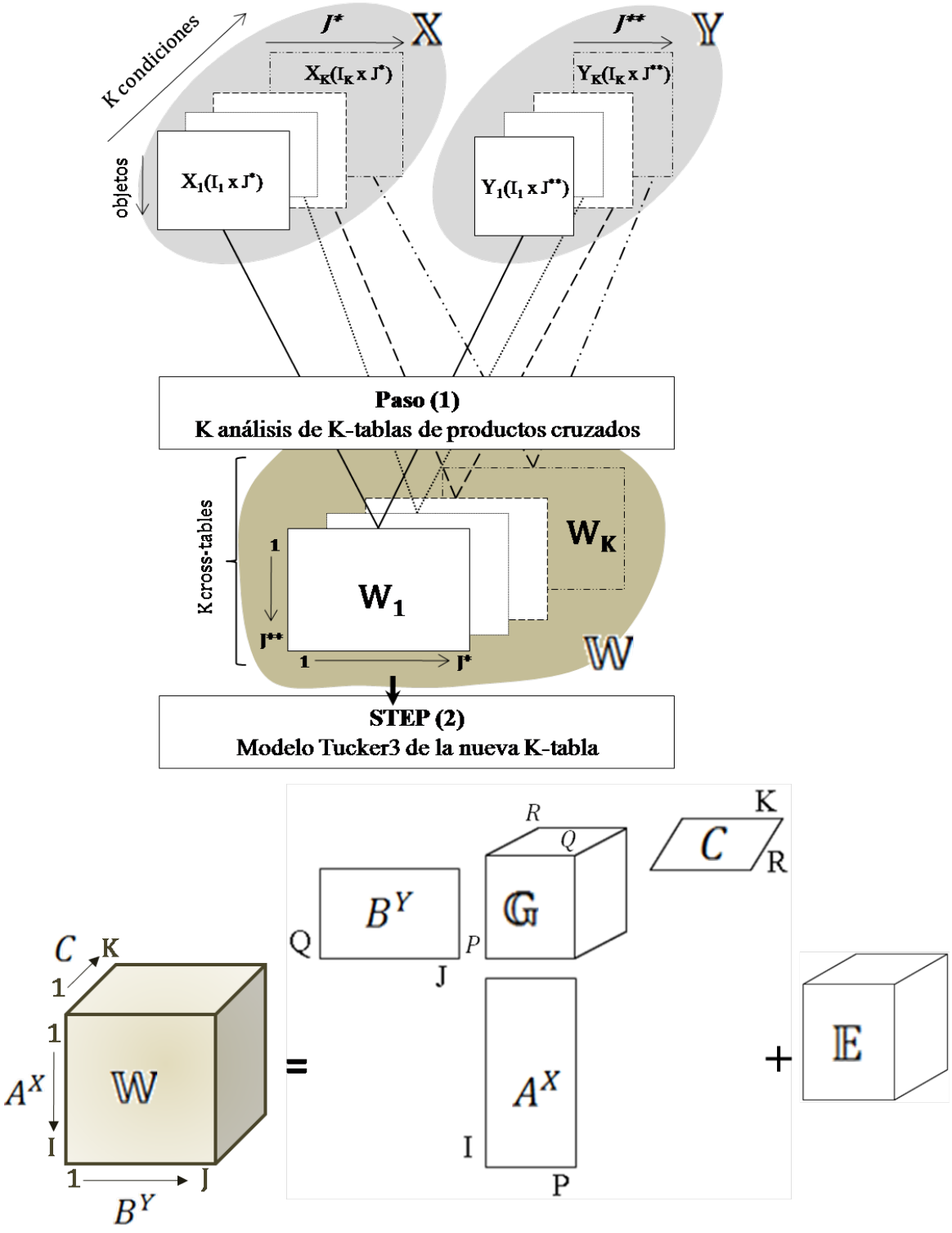


Figura 1. Esquema general del modelo CO-TUCKER. La estructura de datos es una secuencia de K tablas pareadas. Las tablas X_k y Y_k son, respectivamente, el primer grupo (dimensión $I_K \times J^*$) y el segundo grupo (dimensión $I_K \times J^{**}$) de datos. La tabla W_K es la matriz de los productos cruzados en las k condiciones; J^*es el numero de variables de $\mathbb{X}$, J^{**} es numero de variables de $\mathbb{Y}$, I_K es el numero de filas al largo de las k condiciones. (1) K Análisis de Co-Inercia para calcular la secuencia de las k tablas de covarianzas cruzada (W_K); (2) la K-tabla $\mathbb{W}$ es analizada mediante un modelo Tucker3.

En el método propuesto es un método en dos pasos principales:

(1) En un primer paso, tras el análisis individualizado de cada matriz de cada uno de los cubos (mediante un PCA, CA, o análisis de correspondencias múltiples (MCA), según corresponda) se realizan k Análisis de Co-Inercia para calcular la secuencia de las k tablas de covarianzas cruzada (W_k), es decir, consiste en realiza k análisis de k-tablas de productos cruzados ($\left(W_k, D_I, D_J\right)$ donde $W_k = Y_k^T D_{I_k} X_k$, siendo X_k el primer conjunto de datos, Y_k es el segundo conjunto de datos y D_I y D_J las matrices diagonales y D_{I_k} la matriz de pesos);

(2) En el segundo paso, el cubo de operadores de co-inercia, dicho de otro modo, el cubo o la k-tabla $\mathbb{W}$ obtenida en el primer paso, es analizada mediante un modelo Tucker3. Es decir, se realiza un análisis de estas matrices generadas en el primer paso del análisis por medio del modelo Tucker3. Esto significa que las k tablas (o matrices, que representan las k condiciones) tienen I filas y J columnas. Estas k matrices, W_k, se descomponen y se pueden formular como una factorización de la matriz de tres vías, $\mathbb{W}(I \times J \times K)$, de tal manera que:

$$w_{ijk} = \sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{r=1}^{R} a_{ip}^{X} b_{jq}^{Y} c_{kr} g_{pqr} + e_{ijk}\,, i = 1, \dots, I;\ j = 1, \dots, J;\ k = 1, \dots, K$$

donde:

- w_{ijk} se corresponde con valor observado en la combinación de niveles ijk;
- P, Q y R representan el número de componentes retenidas en cada modo;
- a_{ip}^{X} representa el valor que toma para el individuo i la componente p del primer modo $A^{X}(I \times P)$;
- b_{jq}^{Y} representa el valor que toma para la variable j la componente q del segundo modo $B^{Y}(J \times Q)$;
- c_{kr} representa el valor que toma para la ocasión k la componente r del tercer modo $C(K \times R)$;
- g_{pqr} es una medida de la relación entre la componente p del primer modo, la componente q del segundo modo y la componente r del tercer modo;
- e_{ijk} es un elemento de la matriz residual de tres vías $\mathbb{E}$ (y que denota un término de error asociado con la descripción de w_{ijk}).

En el arreglo de tres vías $\mathbb{G}$, (con elemento genérico g_{pqr}), denominado matriz *core, se* encuentran las interrelaciones entre las respectivas direcciones de inercia de cada modo. Al igual que ocurría con $\mathbb{W}$, a partir de $\mathbb{G}$ se construyen tres matrices o arreglos de dos vías, $\mathbb{G}(I \times JK)$, $\mathbb{G}(J \times KI)$ y $\mathbb{G}(K \times IJ)$, denominadas matrices de enlace; las cuales pueden ser entendidas como una generalización de la matriz de valores propios asociada a la descomposición en dos vías.

Por otra parte, la matriz *core* $\mathbb{G}$ se deriva de las tres matrices de componentes $A^{X}(I \times P)$, $B^{Y}(J \times Q)$ y $C(K \times R)$ de la siguiente manera:

$$g_{pqr} = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} a_{ip}^{X}\, b_{jq}^{Y} c_{kr} w_{ijk}$$

En el modelo general CO-TUCKER, los elementos g_{pqr} son una medida de la relación entre la componente p de la primera vía, la componente q de la segunda vía y la componente r de la tercera vía. El hecho de que g_{pqr} represente relaciones entre componentes, es decir, entre variables continuas, hace que su interpretación dependa a su vez de la interpretación que fuera deducida para cada componente.

A diferencia de la matriz de valores propios asociada a la descomposición de valores singulares para datos de dos vías, cuando se generaliza a datos de tres vías, el arreglo $\mathbb{G}(P \times Q \times R)$ puede contener valores g_{pqr} positivos o negativos. El valor absoluto de cada g_{pqr} nos da una medida de la magnitud de la relación existente entre las componentes p, q y r.

Supongamos que se desea interpretar un valor g_{pqr} positivo. Cuando ocurre esto, las posibles combinaciones de cada vía en el modelo CO-TUCKER, en las ternas formadas por el signo de las componentes $(a_{ip}^{X}, b_{jq}^{Y}, c_{kr})$ pueden dar como resultado resultar cuatro situaciones de forma simultánea:

$$[(+)P,(+)Q,(+)R],[(+)P,(-)Q,(-)R],[(-)P,(+)Q,(-)R] \text{ y } [(-)P,(-)Q,(+)R]$$

$[(+)P,(+)Q,(+)R]$: significa que combinaciones en los niveles de los modos i y j con altos valores en a_{ip}^{X} y b_{jq}^{Y}, tienden a interactuar positivamente en cada nivel del modo k con altos valores en c_{kr}.

$[(+)P,(-)Q,(-)R]$: significa que combinaciones en los niveles de los modos i y j con altos valores en a_{ip}^{X} y bajos valores en b_{jq}^{Y}, tienden a interactuar positivamente en cada nivel del modo k con bajos valores en c_{kr}.

$[(-)P,(+)Q,(-)R]$: significa que combinaciones en los niveles de los modos i y j con bajos valores en a_{ip}^{X} y altos valores en b_{jq}^{Y}, tienden a interactuar positivamente en cada nivel del modo k con bajos valores en c_{kr}.

$[(-)P,(-)Q,(+)R]$: significa que combinaciones en los niveles de los modos i y j con bajos valores en a_{ip}^{X} y b_{jq}^{Y}, tienden a interactuar positivamente en cada nivel del modo k con altos valores en c_{kr}.

Las combinaciones de signos para la terna indicada que no cumplan con ninguna de las cuatro combinaciones anteriores, tienden a interactuar negativamente.

Si g_{pqr} es negativo, son posibles otras cuatro combinaciones de la terna $(a_{ip}^{X}, b_{jq}^{Y}, c_{kr})$, formada por los signos de sus pesos:

$[(-)P,(+)Q,(+)R], [(+)P,(-)Q,(+)R], [(+)P,(+)Q,(-)R]$ y $[(-)P,(-)Q,(-)R]$

$[(-)P,(+)Q,(+)R]$: significa que combinaciones en los niveles de los modos i y j con bajos valores en a_{ip}^{X} y altos valores en b_{jq}^{Y}, tienden a interactuar positivamente en cada nivel del modo k con altos valores en c_{kr}.

$[(+)P,(-)Q,(+)R]$: significa que combinaciones en los niveles de los modos i y j con altos valores en a_{ip}^{X} y bajos valores en b_{jq}^{Y}, tienden a interactuar positivamente en cada nivel del modo k con altos valores en c_{kr}.

$[(+)P,(+)Q,(-)R]$: significa que combinaciones en los niveles de los modos i y j con altos valores en a_{ip}^{X} y b_{jq}^{Y}, tienden a interactuar positivamente en cada nivel del modo k con bajos valores en c_{kr}.

$[(-)P,(-)Q,(-)R]$: significa que combinaciones en los niveles de los modos i y j con bajos valores en a_{ip}^{X} y b_{jq}^{Y}, tienden a interactuar positivamente en cada nivel del modo k con bajos valores en c_{kr}.

Las combinaciones de signos para la terna indicada que no correspondan a ninguna de las cuatro combinaciones anteriores, tienden a interactuar negativamente.

Al igual que en el caso de dos vías, este análisis se realiza solamente con las combinaciones de los modos i, j y k que tienen los mayores valores absolutos en las respectivas matrices de marcadores.

En la figura 2, síntesis de lo anterior, se muestra el esquema general de la interpretación de los signos.

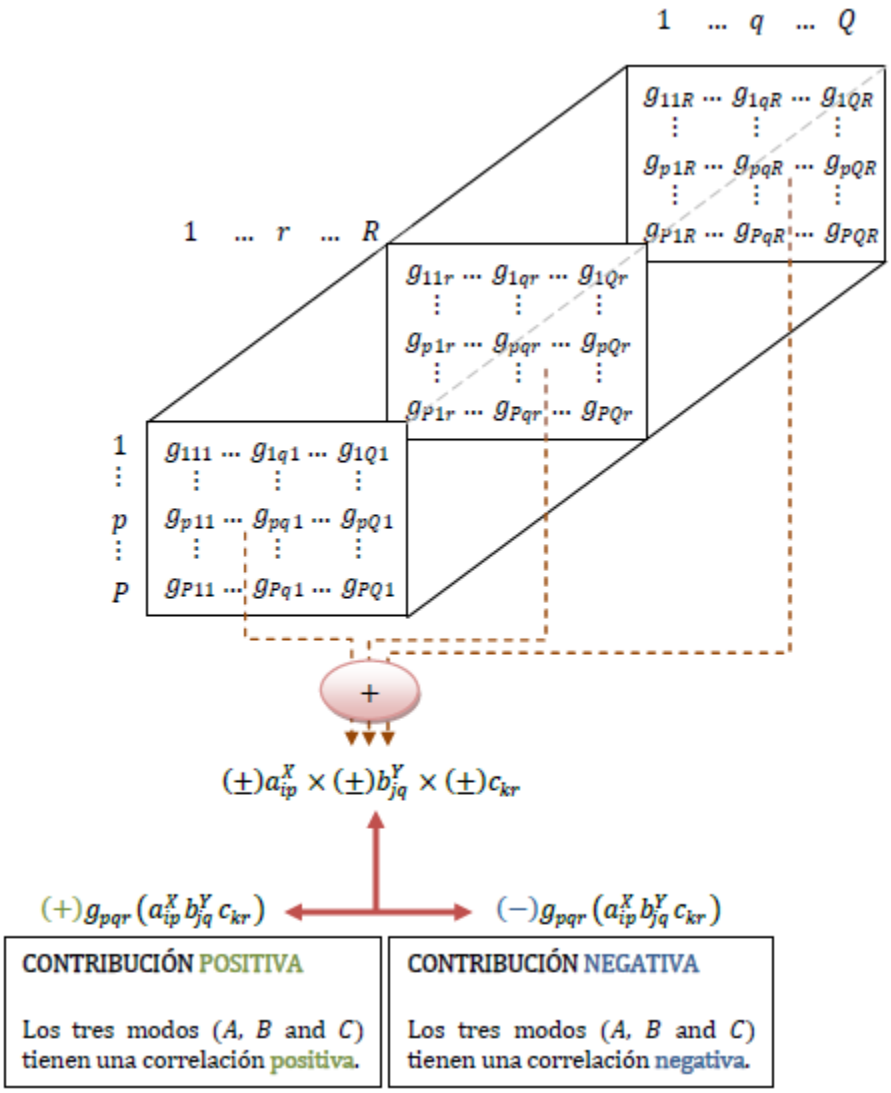


Figura 2. Esquema general para la interpretación de los signos.

Los procedimientos de análisis inherente al modelo CO-TUCKER se proyectaron siguiendo los siguientes pasos (Fig. 3):

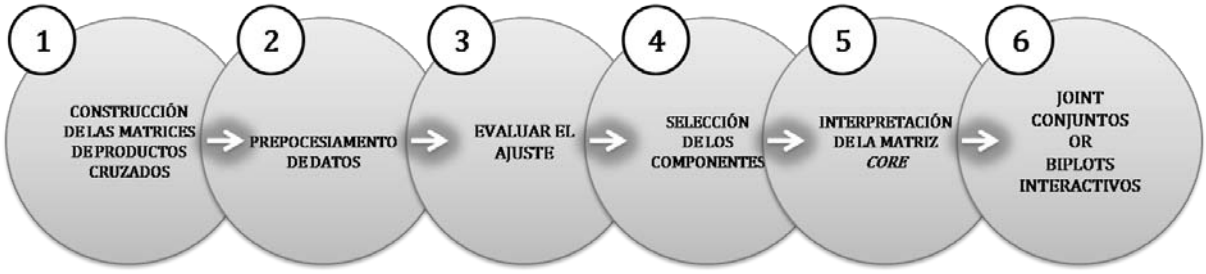


Figura 3. Etapas del modelo CO-TUCKER

En el **CAPÍTULO CINCO** de la tesis se muestra la aplicación del CO-TUCKER al caso real que motivó el desarrollo de los aportes presentados en la memoria que se presenta.

Cabe señalar el hecho de que, debido a la aproximación de mínimos cuadrados utilizada en el modelo Tucker3 (para estimar los parámetros), es posible conocer la bondad del ajuste obtenida para el modelo ajustado, sujetos, variables y condiciones. Esto puede ser una ventaja sobre el STATICO, ya éste que se basa en la maximización de la similitud de toda la serie de co-estructuras, lo que no permite conocer si se ajusta adecuadamente a las partes individuales de datos.

Por otra parte, esta investigación puede mostrar que ciertos niveles (i, j, k) de uno modo $(A^X(I \times P), B^Y(J \times Q), C(K \times R))$ están dominando la solución más de lo que es aceptable (teniendo en cuenta su "papel" en la investigación).

Por otro lado, se debe prestar especial atención a la organización los arreglos de una serie de k-tablas. Para una matriz de tres vías, hay tres formas de organizar el cubo de datos en una serie de tablas. Sin embargo, sólo dos son realmente interesantes: la opción de poner una tabla por una variable no es interesante y tampoco coherente con el objetivo de la mayoría de las investigaciones. Por lo tanto, la elección entre poner una tabla para cada condición k o una tabla para cada objeto i juega un papel importante. Es obvio que esta elección depende del objetivo en cuestión. Sin embargo, esta posibilidad en CO-TUCKER debe utilizarse con cautela, ya que esta flexibilidad puede ser obtenida a costa de perder algunas de las estructuras en el conjunto de datos. En efecto, si la

organización es tener una tabla para cada condición k (es decir, I filas × J variables × K condiciones) luego, con la ejecución del modelo, la información sobre los objetos se "perderá". Análogo a un tipo de organización I filas × J variables × K objetos, donde la información que se "pierde" está relacionada con las condiciones. Esto es, de hecho, un inconveniente de enfoque en el CO-TUCKER.

Básicamente, la mayoría de los métodos multi-tablas buscan las semejanzas entre los datos de las tablas (es decir, matrices) y no las diferencias. Así, si hay diferencias entre los datos de las tablas, estos métodos señalarán solamente que no existe una estructura común y por tanto habrá que explorar los datos con otras técnicas que si nos permitan explicar las referidas diferencias.

Además, como los dos primeros ejes del análisis de la interestructura obtenida con estas técnicas solamente mostrarán el consenso entre los datos de las tablas, el poder observar las dimensiones de órdenes superiores o analizar los residuos puede proporcionar importante información sobre las diferencias entre las diferentes tablas.

Por lo tanto (y probablemente), este es uno de las principales ventajas del CO-TUCKER ya que permite la selección de un número distinto de componentes para cada modo y debido a la aproximación mediante el método de los mínimos cuadrados utilizado para estimar los parámetros, es posible indicar en qué cuantía la variabilidad de cada sujeto/objeto, variable, y condición está representado/ajustado por el modelo de tres vías.

En conclusión, la disponibilidad de métodos capaces del análisis de conjuntos de datos con una organización compleja, como pares de cubos de datos, es muy importante porque permite tener en cuenta esta organización y permite analizar los conjuntos de datos a nivel general. Es obvio que la adición de la complejidad en la exploración de las relaciones de los cubos de datos podría aumentar la dificultad en la interpretación. Sin embargo, este es un paso indispensable para comprender la estructura de los datos y su significado. Por lo tanto, el método que aquí se presenta, el CO-TUCKER, puede ser visto bien como un complemento bien como una alterativa a otros métodos existentes en la literatura, para el completo conocimiento de las estructuras subyacentes en complejos conjuntos de datos.

PUNTOS ABIERTOS

En el trabajo aquí presentado se indican una serie de áreas que aún necesitan ser consideradas en más detalle. La técnica propuesta en esta tesis, denominada el CO-TUCKER, es potencialmente aplicable también en muchas otras áreas, como lo demuestran los numerosos trabajos de áreas diferentes a la Ecología, en los que se aplican técnicas ya existentes, focalizadas inicialmente hacia el análisis de datos ecológicos.

En esta tesis, algunos de los resultados más fructíferos surgieron de la colaboración con los campos fuera de la Estadística (en concreto, de la Ecología). Tal labor interdisciplinar a menudo produce información útil, que de otra manera no tendría la posibilidad de florecer. La colaboración interdisciplinar puede ayudar a proporcionar los modelos y

algoritmos más rápidos y dirigidos específicamente a hacer frente a problemas particulares.

Esta tesis tiene la intención de ser una contribución a la maduración del análisis de tres vías, pero sobre todo a la maduración del campo de análisis multivariante ligado a los temas relacionados con la Ecología.

Una tesis doctoral no debe convertirse en el final de un camino, por tanto esta memoria, resultado del esfuerzo y la investigación de años, se configura simplemente como el inicio de un trabajo futuro con el que esperamos dar continuidad y solución a la multitud de problemas abiertos que en ella se han planteado y tratado de resolver haciendo frente a dicho futuro con la eficacia necesaria en la articulación del análisis de datos de tablas múlti-vía.

CONCLUSIONES

La exhaustiva revisión de la bibliografía especializada en el campo de la Ecología, nos ha permitido:

1. Conocer el impacto de los métodos multivariantes descriptivos para el análisis de datos organizados en una matriz de dos vías, en un cubo de datos, en dos matrices y/o en dos cubos, llegando a la conclusión de que el Análisis de Correspondencias y el Análisis Canónico de Correspondencias son las técnicas que se usan en un altísimo porcentaje de estos artículos. El uso de los métodos de la familia STATIS es mucho menor, aunque se ha detectado un crecimiento a lo largo del período de estudio. Los modelos de la escuela

holandesa, a pesar de capturar información que puede ser mucho más relevante, prácticamente son desconocidos por los ecólogos.

2. No existe en la literatura especializada ningún procedimiento que permita describir no sólo la parte estable de las relaciones entre dos conjuntos de datos observados en diferentes tiempos o condiciones sino también la parte cambiante en las estructuras.

Hemos propuesto un marco general para el análisis de datos multivariantes de tres vías procedentes de estudios ecológicos, en el cual hemos puesto de manifiesto que:

3. En muchos conjuntos de datos de tres vías (objetos × variables × condiciones) la información relevante (o no-trivial) está más allá de las dos primeras componentes principales.
4. El uso de los métodos multi-tablas que investigan las configuraciones consenso (como en el caso de PTA y STATICO) proporcionan, en muchos casos, "sólo" la información trivial que se manifiesta a través de las dos primeras componentes principales.
5. El uso de dichos métodos multi-tablas sólo ponen de manifiesto la estructura común, y cuando esta estructura común es inexistente deben realizarse algunas exploraciones.
6. El uso de modelos de tres vías (para conjuntos de datos de tres vías) obtiene una ventaja, ya que permite elegir un número diferente de componentes de cada modo.
7. Hemos desarrollado un nuevo método de análisis multivariante para analizar dos conjuntos de tablas de tres vías al que hemos denominado CO-TUCKER

siguiendo la terminología de Thiolouse. Asimismo, el nombre CO-TUCKER significa "CO-inertia y TUCKER", dado que en primer lugar utiliza *k* Análisis de Co-Inercia (para calcular la secuencia de las *k* tablas de covarianzas cruzada), y luego el modelo Tucker3 para analizar esta nueva *k*-tabla.

8. El modelo resuelve los problemas cuando el objetivo es poner de manifiesto las diferencias y no las configuraciones comunes dentro de los datos. A diferencia de otros modelos relacionados, este modelo demuestra tener algunas propiedades intrínsecas y singulares debido a su relación con el STATICO y el modelo Tucker3.
9. El hecho de que el nuevo modelo propuesto requiera de la utilización de algoritmos de cálculo, implica que modificaciones en el algoritmo pueden proporcionar modificaciones del modelo.
10. Con la nueva propuesta se ha logrado:

    Resultados estructurales de alta calidad y más interpretables debido a la parsimonia y la correspondencia entre la naturaleza de los datos y el modelo. Además, a través de la matriz *core* y de los biplots conjuntos, dichos resultados captan mejores y/o más reales predicciones. Asimismo, destaca información sobre las diferencias entre las tablas de datos y proporciona una mejor comprensión de los patrones de la variabilidad asociada a los cambios temporales y espaciales de los complejos conjuntos de datos.
11. La contribución del grupo de expertos en Ecología Marina que nos ha facilitado la interpretación de los resultados del análisis de los datos reales utilizados, ha puesto de manifiesto que la interacción entre estadísticos y ecólogos enriquece no sólo a los unos y a los otros sino también a la Ciencia.

REFERENCIAS BIBLIOGRÁFICAS

La revisión bibliográfica realizada, es un ejemplo claro del auge de la aplicación de estos métodos en diversos campos de estudio. La revisión exhaustiva de la literatura, conjuntamente con los resultados de diversos ejemplos donde estas técnicas son aplicadas, evidencia la potencialidad e importancia del tema.