Pearson (1901) ou quelques considérations sur les méthodes d'ordination dans R

Emmanuel Castella - Institut F.A. Forel des Sc. de l'Environnement et de l'Eau
DéjeuneRs - 25 avril 2016

LIII. *On Lines and Planes of Closest Fit to Systems of Points in Space.* By KARL PEARSON, *F.R.S., University College, London* *.

(1) IN many physical, statistical, and biological investigations it is desirable to represent a system of points in plane, three, or higher dimensioned space by the "best-fitting" straight line or plane. Analytically this consists in taking

$$y = a_0 + a_1 x, \quad \text{or} \quad z = a_0 + a_1 x + b_1 y,$$
$$\text{or} \quad z = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_3 + \ldots + a_n x_n,$$

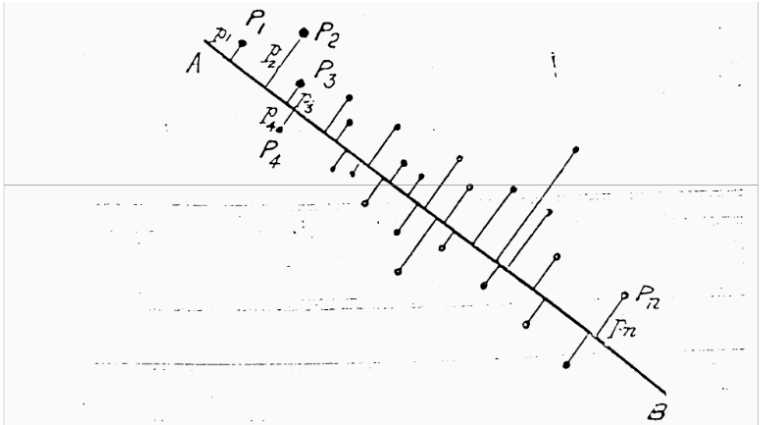
where y, x, z, x_1, x_2, $\ldots x_n$ are variables, and determining the "best" values for the constants a_0, a_1, b_1, a_2, a_1, a_2, a_3, $\ldots a_n$ in relation to the observed corresponding values of the variables. In nearly all the cases dealt with in the text-books of least squares, the variables on the right of our equations are treated as the independent, those on the left as the dependent variables. The result of this treatment is that we get one straight line or plane if we treat some one variable as independent, and a quite different one if we treat another variable as the independent variable. There is no paradox about this; it is, in fact, an easily understood and most important feature of the theory of a system of correlated variables. The most probable value of y for a given value of x, say, is not given by the same relation as the most probable value of x for a given value of y. Or, to take a concrete example, the most probable stature of a man with a given length of leg l being s, the most probable length of leg for a man of stature s will not be l. The "best-fitting" lines and planes for the cases of z up to n variables for a correlated system are given in my memoir on regression †. They depend upon a determination of the means, standard-deviations, and correlation-coefficients of the system. In such cases the values of the independent variables are supposed to be accurately known, and the probable value of the dependent variable is ascertained.

(2) In many cases of physics and biology, however, the "independent" variable is subject to just as much deviation or error as the "dependent" variable. We do not, for example, know x accurately and then proceed to find y, but both x and y are found by experiment or observation. We observe x and y and seek for a unique functional relation between them. Men of given stature may have a variety

* Communicated by the Author.

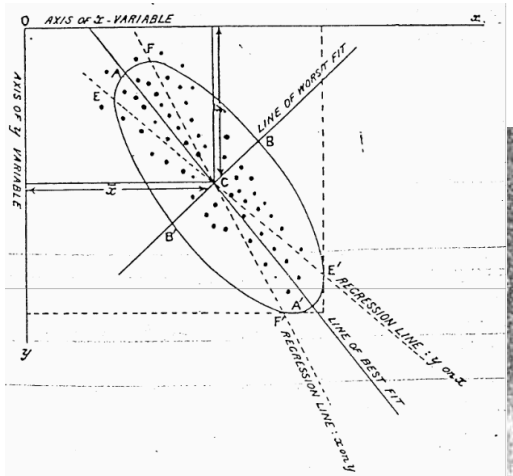
† Phil. Trans. vol. clxxxvii. A, pp. 301 *et seq.*

Pearson, K. 1901. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine* **2**:559-572.



Now clearly $U = S(p^2)$ is the moment of momentum, the second moment of the system of points, supposed equally loaded, about the line A B. But the second moment of a system about a series of parallel lines is always least for the

Pearson, K. 1901. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine* **2**:559-572.
http://pbil.univ-lyon1.fr/R/pearson1901.pdf



Physically the axes of the correlation type-ellipse are the directions of independent or uncorrelated variation. Hence the line of best fit is a direction of uncorrelated variation.

Pearson, K. 1901. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine* **2**:559-572.

Ordination: ordonner des objets les uns par rapport aux autres

-> # classification
-> mesure de distance
-> réduire (résumer) la complexité d'une structure multidimensionnelle
-> technique descriptive / exploratoire

Généralité de la question:

-> espèces * relevés (dénombrement, présence / absence)
-> variables de milieu * relevés (variables quantitatives continues)
-> individus * variables catégorielles (variables qualitatives)

-> pixels * bandes spectrales
-> volcans * propriétés du magma
-> mots * textes
-> populations * fréquences alléliques, …

Catalogue de La Redoute®:

-> pléthore de noms / méthodes dans différents domaines et langues
-> dans l'univers francophone: ACP (Analyse en Composantes Principales) vs AFC (Analyse Factorielle des Correspondances)

1. Une confrontation pédagogique: AFC vs ACP. Quel est le vainqueur ?

2. La grande unification.

3. Un plan factoriel. Et après ?

ade4 dans R

Analyses des Données Ecologiques :
méthodes Exploratoires et Euclidiennes en sciences de l'Environnement

Google ade4 Google ade4

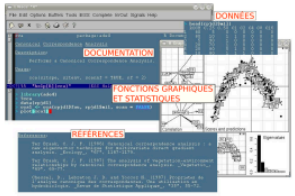
Accueil | adelist | Enseignement | Galerie | Documentation | Telechargements | Rweb | Liens | Historique

Accueil ade4

ade4 est un logiciel développé au laboratoire de Biométrie et Biologie Évolutive (UMR 5558) de l'Université Lyon 1. Il contient des fonctions d'**A**nalyse de **D**onnées destinée d'abord à la manipulation des données **É**cologiques et **E**nvironnementales avec des procédures **E**xploratoires d'essence **E**uclidienne, d'où la dénomination **ade4**.

ade4 se caractérise par :

- l'implémentation de **fonctions statistiques et graphiques**
- la mise à disposition de **données numériques**
- la rédaction d'une **documentation** technique et thématique
- l'inclusion de **références bibliographiques**



fonctions, données, documentation et références dans ade4 (OS : Debian avec Emacs et ESS)

Depuis 2002, **ade4** est un package du logiciel R, disponible sur les systèmes Windows, Mac OS, Linux et Unix.
Les méthodes dédiées à l'analyse d'un tableau sont décrites dans Chessel et al. (2004) (pdf , html).
Les méthodes pour l'analyse de 2 et *K* tableaux sont décrites dans Dray et al. (2007) (pdf , html).
La théorie du schéma de dualité et son implémentation dans **ade4** sont discutés dans Dray et Dufour (2007) (pdf).

D'autres packages liés à **ade4** sont décrits sur la page de téléchargements, en particulier le package **ade4TkGUI** qui propose une interface utilisateur graphique pour **ade4**.

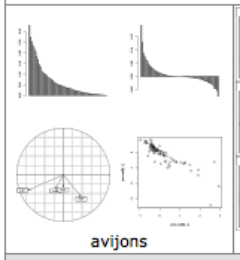
L'ancienne version du logiciel (ADE-4, en majuscules) présenté dans Thioulouse et al. (1997) (pdf , html) est disponible ici mais n'est plus maintenue depuis 2004.

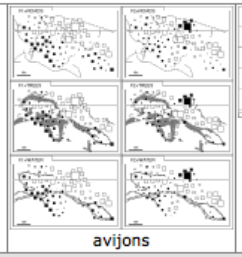
Les documents du stage de formation ade4 de Sept 2008 à Lyon sont disponibles ici

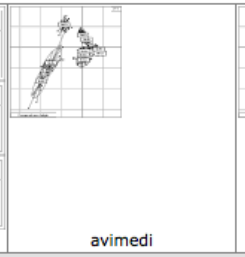
Une liste des articles citant ade4/ADE-4 est disponible ici.

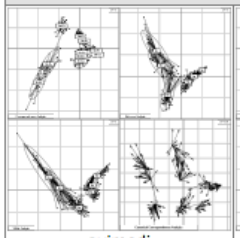
http://pbil.univ-lyon1.fr/ade4/

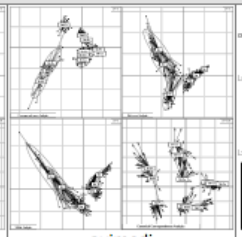
Images per page: 25

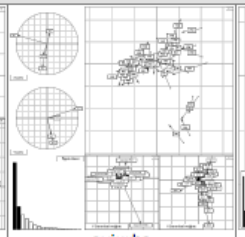

avijons

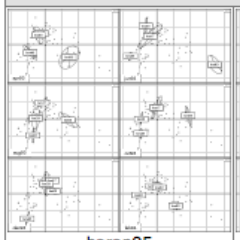

avijons

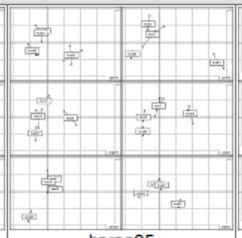

avimedi

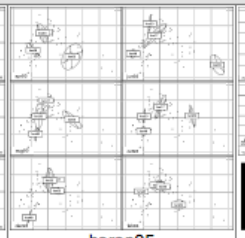

avimedi


avimedi

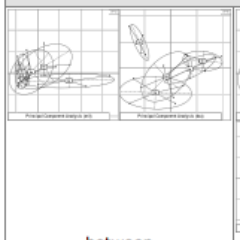

aviurba

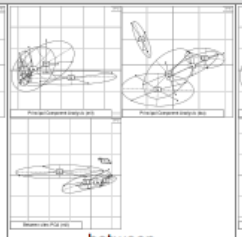

baran95

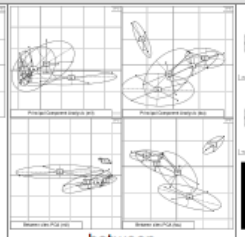

baran95


baran95


be


between


between

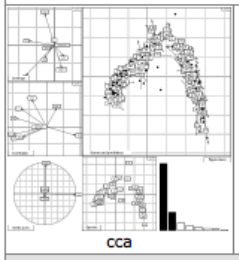

between

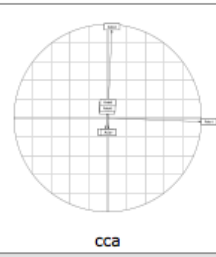

bf88

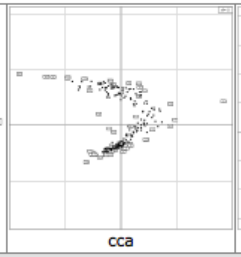

bf88

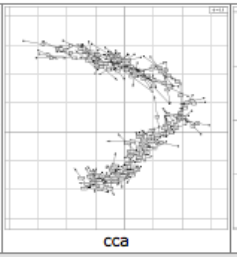

bf88

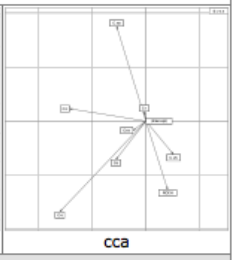
Page 31: «Prev - 1 2 [3] 4 5 6 7 8 9 10 11 12 13 14 15 16 17

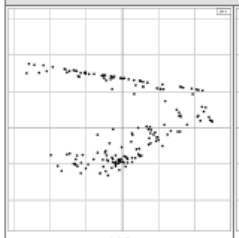

cca

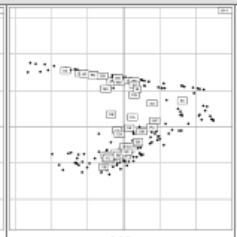

cca

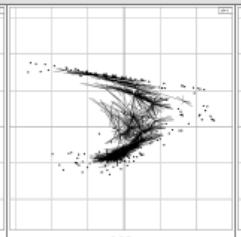

cca

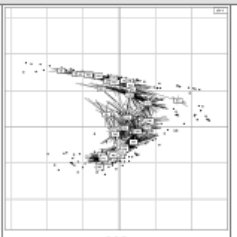

cca

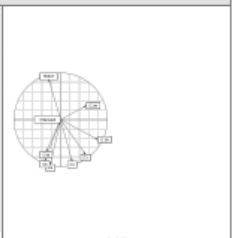

cca

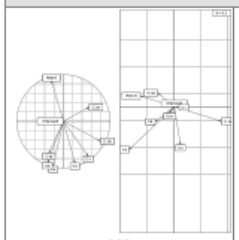

cca

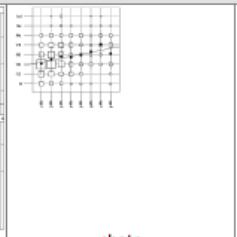

cca

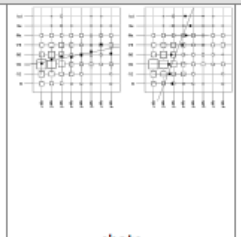

cca

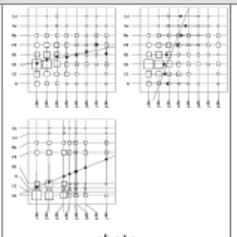

cca

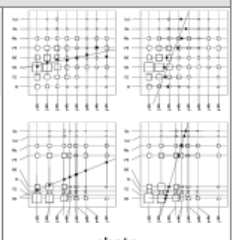

cca

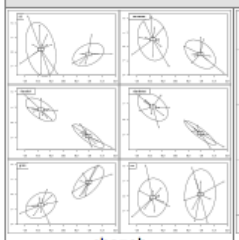

cca

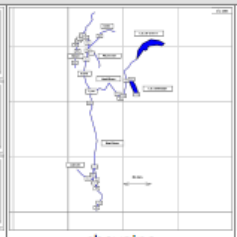

chats

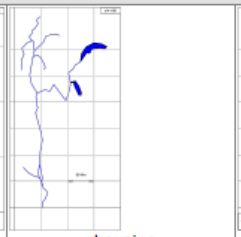

chats

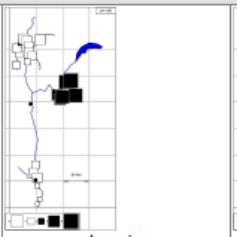

chats

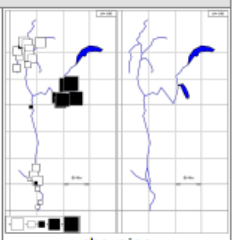

chats


chazeb


chevaine


chevaine


chevaine


chevaine

chevaine

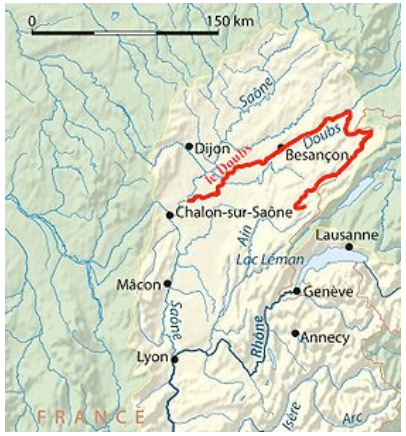
chevaine

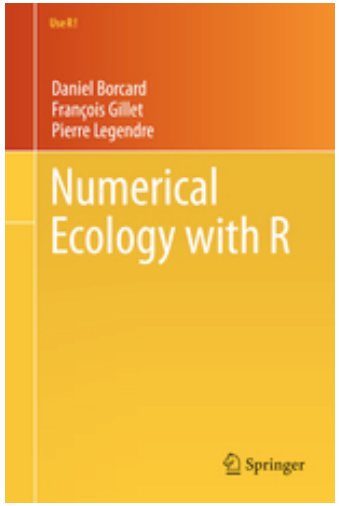
chevaine

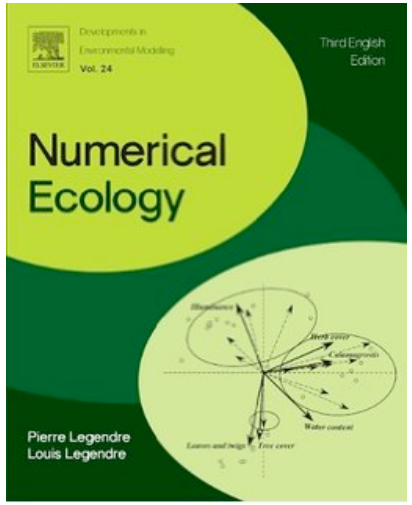
chevaine

clementines

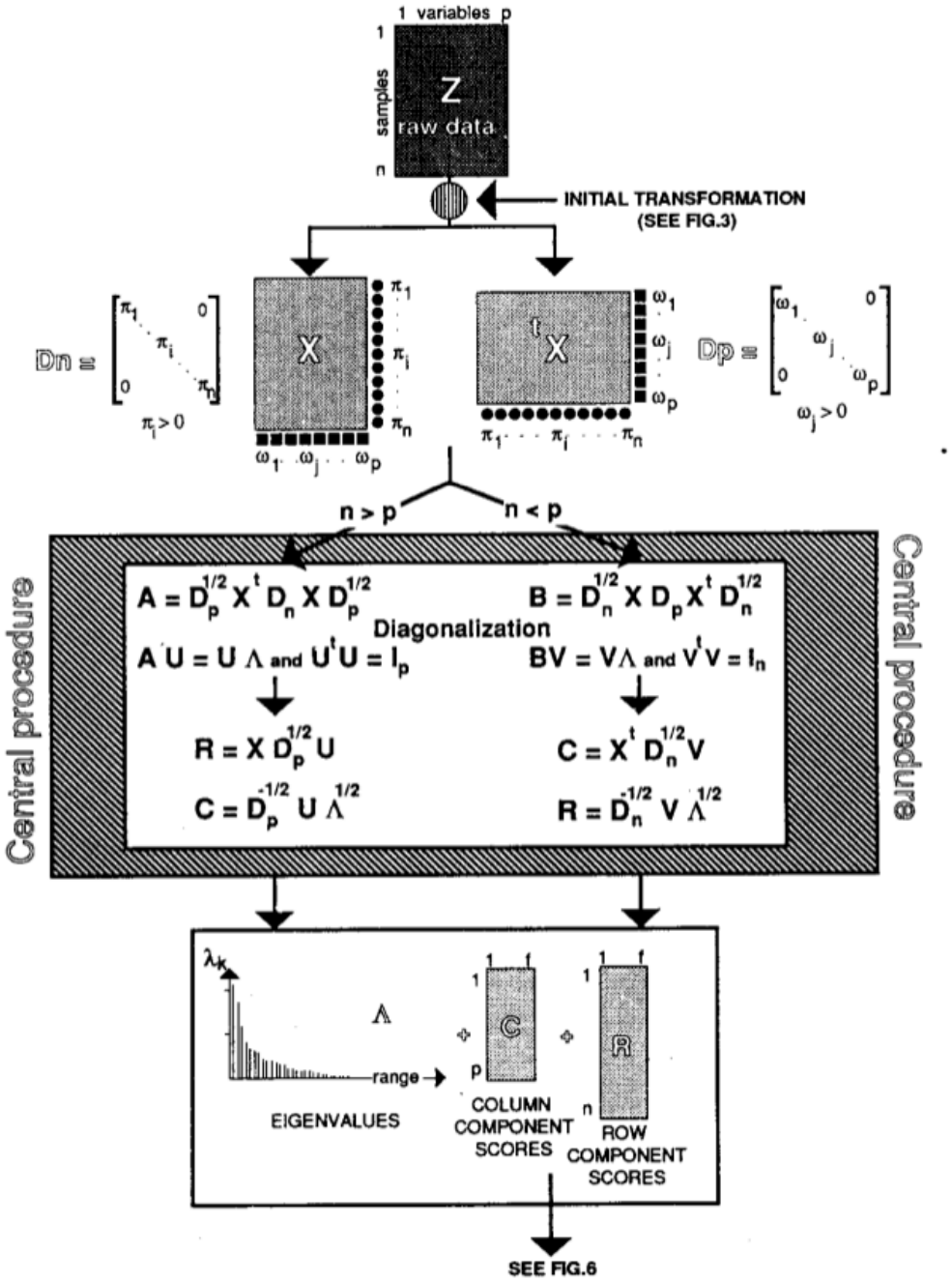
Page 31: «Prev - 1 2 3 4 [5] 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 ... Next»





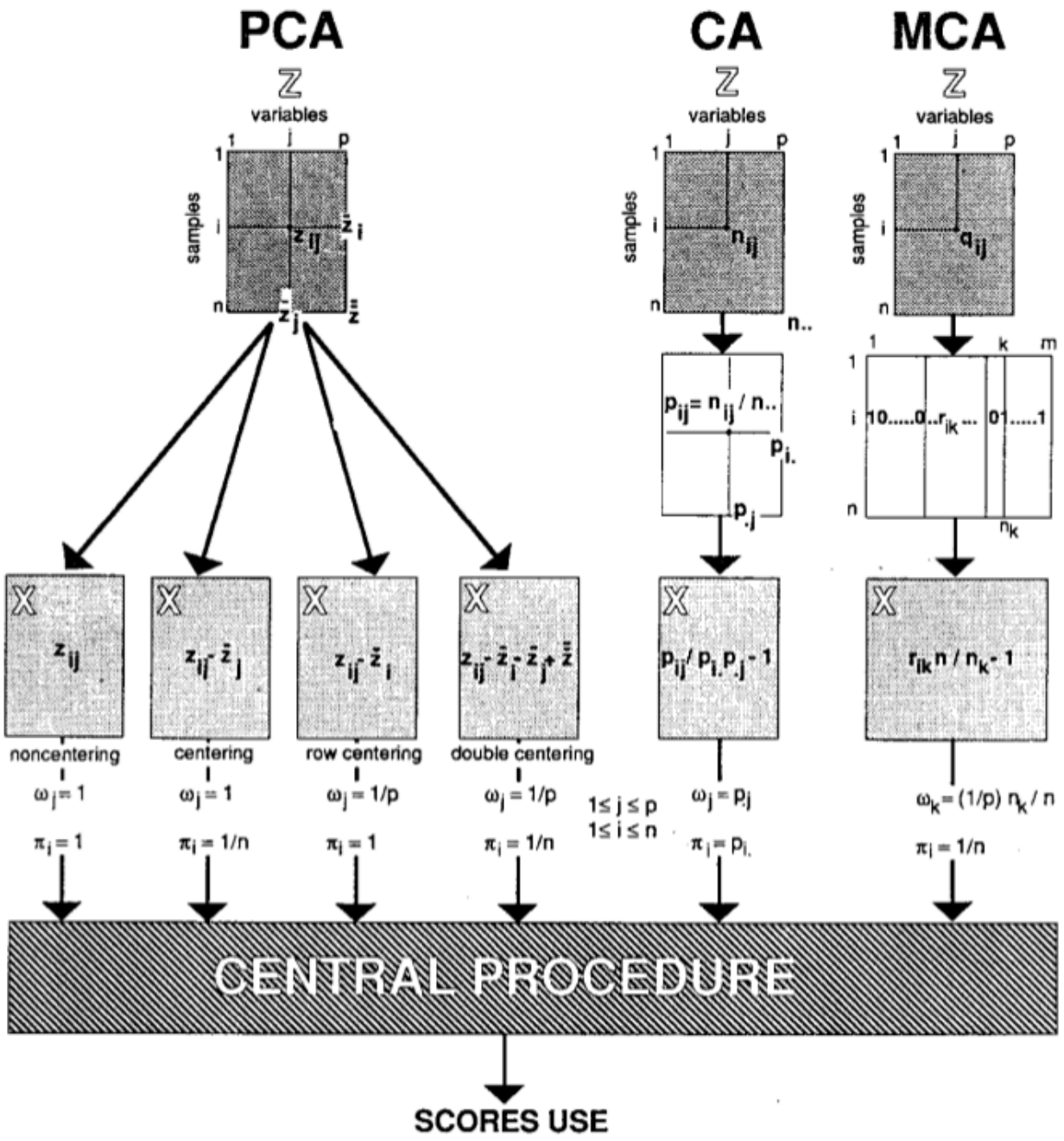


Verneaux J., 1973. Cours d'eau de Franche-Comté (Massif du Jura). Recherches écologiques sur le réseau hydrographique du Doubs. Essai de biotypologie. Thèse de doctorat, Université de Besançon, 257 p.



Dolédec, S., and D. Chessel. 1991. Recent developments in linear ordination methods for environmental sciences. Advances in Ecology, India 1:133-155.

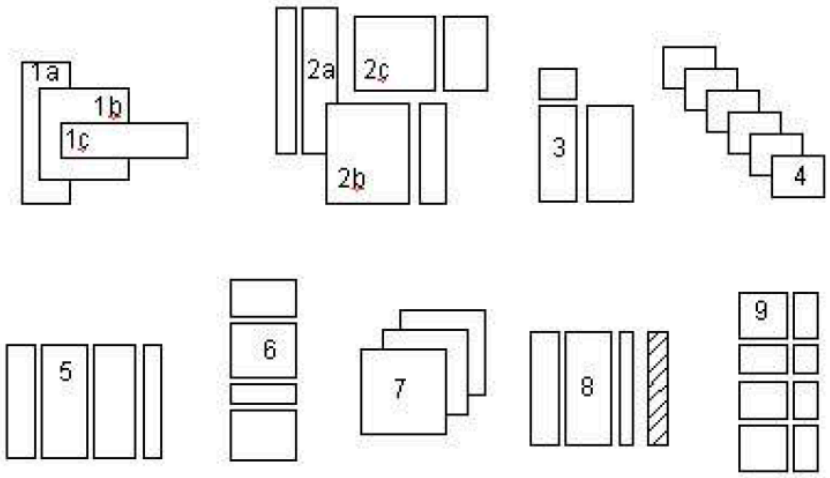
Dolédec, S., and D. Chessel. 1991. Recent developments in linear ordination methods for environmental sciences. Advances in Ecology, India 1:133-155.



Des données originales

Ce qui retiendra le statisticien en écologie des communautés, c'est l'extraordinaire diversité de structures de données, de conditions numériques et de questions plus ou moins clairement formulées susceptibles d'orienter les analyses. L'acquisition d'observations basées sur la notion de *relevés* conduit naturellement à des questions portant sur :

1 - les analyses à un tableau relevés-espèces et les méthodes d'ordination
2 - les couples de tableaux dans des conditions d'analyse canonique (a), de variables instrumentales (b), de co-inertie (c)
3 - les analyses à trois tableaux (biologique-écologique-environnemental)
4 - les cubes de données (stations-espèces-dates)
5 - les K-tableaux par blocs de variables (groupes faunistiques)
6 - les K-tableaux par blocs de stations (expertises régionales)
7 - les K-matrices de distances (spatiale, génétique, écologique, environnementale)
8 - les K+1-tableaux (groupes taxonomiques et variables de milieu)
9 - les K-couples de tableaux (faune-milieu par dates, régions, …)



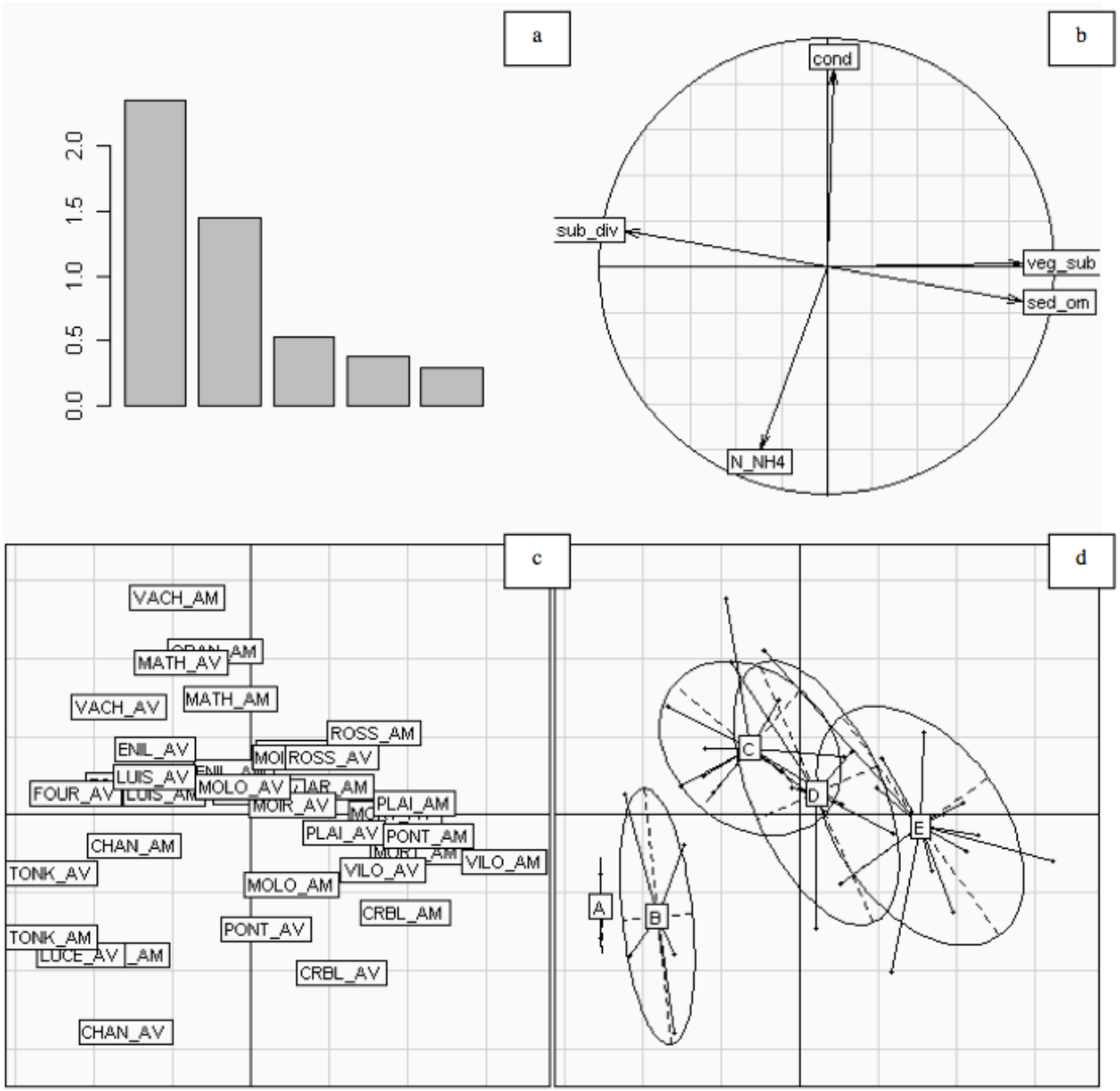
Structures de données fréquentes en écologie des communautés.

D. Chessel, 2000

36 sites alluviaux

5 variables de milieu « traduisant » la connectivité au fleuve

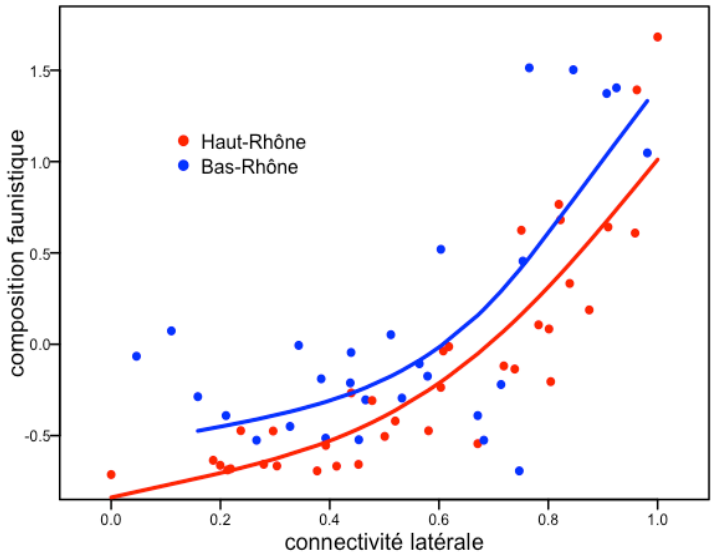
Le 1er axe de l'ACP comme « variable synthétique »

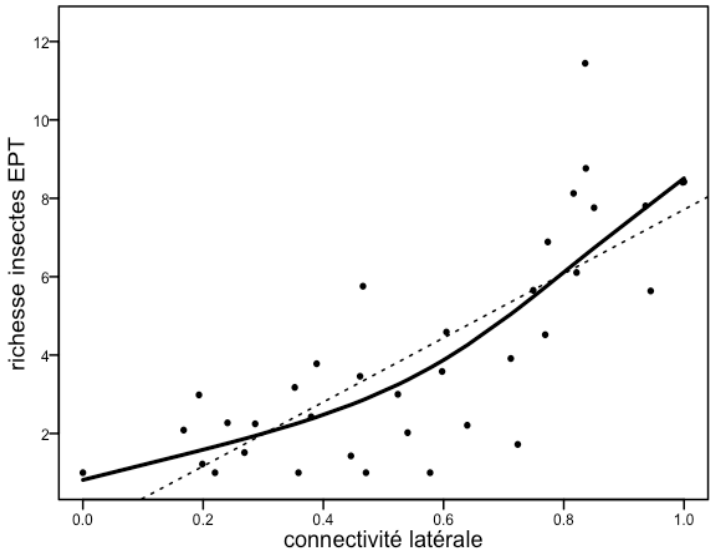


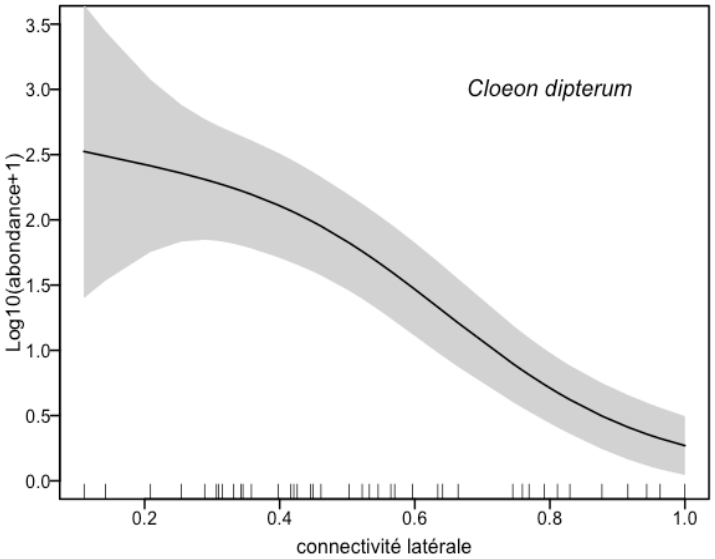
36 sites alluviaux

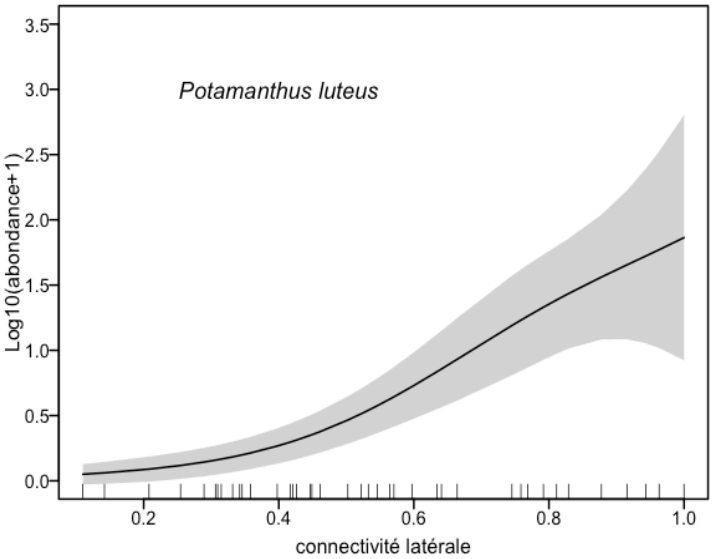
5 variables de milieu « traduisant » la connectivité au fleuve

Le 1er axe de l'ACP comme « variable synthétique »









Connexion avec le fleuve

Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Theoretical Biology 228 (2004) 523–537
Journal of Theoretical Biology
www.elsevier.com/locate/jtbi

From dissimilarities among species to dissimilarities among communities: a double principal coordinate analysis

Sandrine Pavoine*,*, Anne-Béatrice Dufour*, Daniel Chessel*

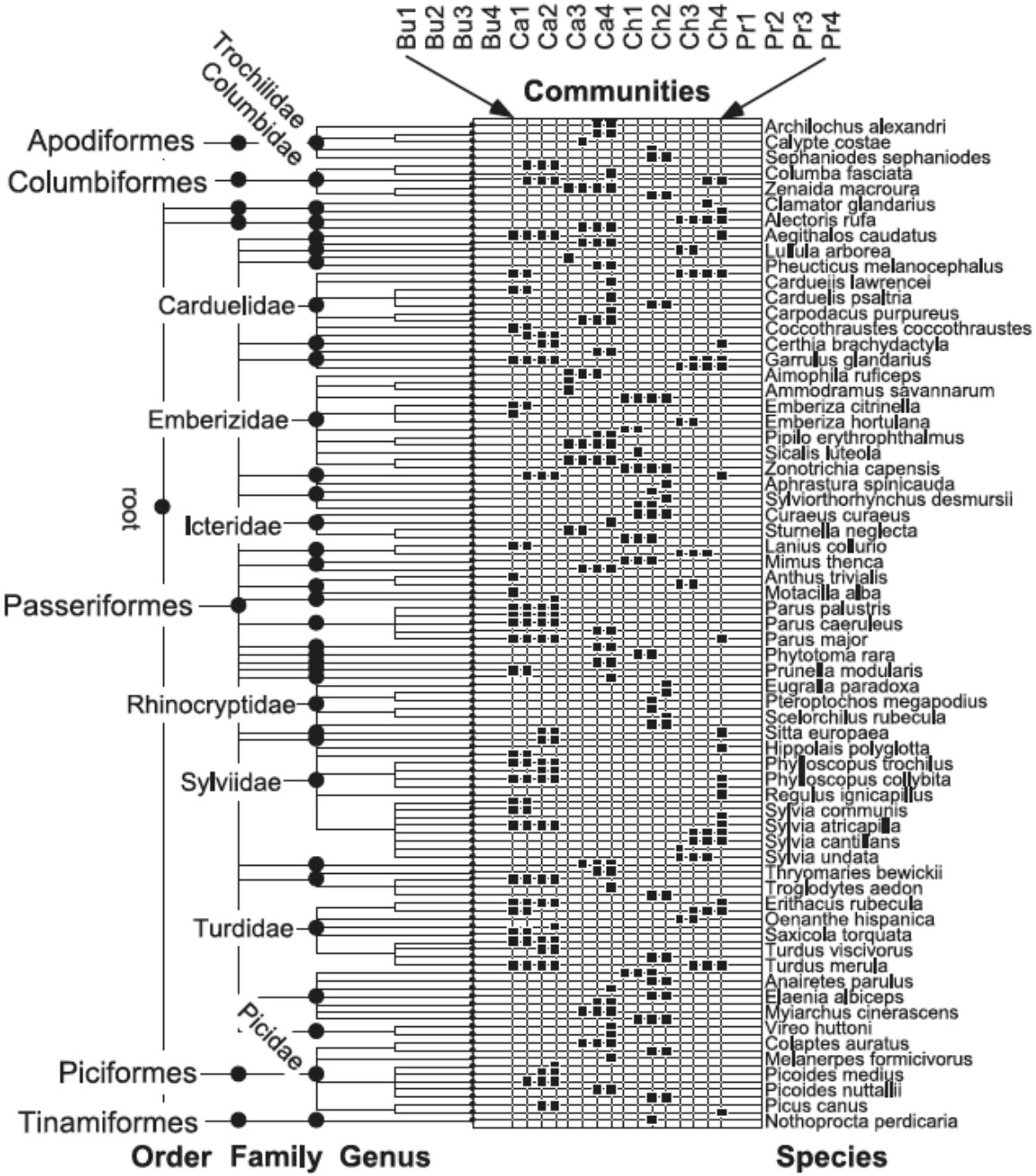
* *Laboratoire de Biométrie et Biologie Evolutive, UMR CNRS 5558, Université Claude Bernard LYON I, 43, boulevard du 11 novembre 1918, 69622 Villeurbanne Cedex, France*

Received 18 June 2003; received in revised form 28 November 2003; accepted 9 February 2004

Abstract

This paper presents a new ordination method to compare several communities containing species that differ according to their taxonomic, morphological or biological features. The objective is first to find dissimilarities among communities from the knowledge about differences among their species, and second to describe these dissimilarities with regard to the feature diversity within communities. In 1986, Rao initiated a general framework for analysing the extent of the diversity. He defined a diversity coefficient called quadratic entropy and a dissimilarity coefficient and proposed a decomposition of this diversity coefficient in a way similar to ANOVA. Furthermore, Gower and Legendre (1986) built a weighted principal coordinate analysis. Using the previous context, we propose a new method called the double principal coordinate analysis (DPCoA) to analyse the relation between two kinds of data. The first contains differences among species (dissimilarity matrix); the second the species distribution among communities (abundance or presence/absence matrix). A multidimensional space assembling the species points and the community points is built. The species points define the original differences between species and the community points define the deduced differences between communities. Furthermore, this multidimensional space is linked with the diversity decomposition into between-community and within-community diversities. One looks for axes that provide a graphical ordination of the communities and project the species onto them. An illustration is proposed comparing bird communities which live in different areas under mediterranean bioclimates. Compared to some existing methods, the double principal coordinate analysis can provide a typology of communities taking account of an abundance matrix and can include dissimilarities among species. Finally, we show that such an approach generalizes some of these methods and allows us to develop new analyses.

Keywords: Dissimilarity; Diversity; Quadratic entropy; PCoA


Bu1
Bu2
Bu3
Bu4
Ca1
Ca2
Ca3
Ca4
Ch1
Ch2
Ch3
Ch4
Pr1
Pr2
Pr3
Pr4
Communities
Trochilidae
Columbidae
Apodiformes
Columbiformes
Carduelidae
Emberizidae
root
Icteridae
Passeriformes
Rhinocryptidae
Sylviidae
Turdidae
Picidae
Piciformes
Tinamiformes
Archilochus alexandri
Calypte costae
Sephaniodes sephaniodes
Columba fasciata
Zenaida macroura
Clamator glandarius
Alectoris rufa
Aegithalos caudatus
Lullula arborea
Pheucticus melanocephalus
Carduelis lawrencei
Carduelis psaltria
Carpodacus purpureus
Coccothraustes coccothraustes
Certhia brachydactyla
Garrulus glandarius
Aimophila ruficeps
Ammodramus savannarum
Emberiza citrinella
Emberiza hortulana
Pipilo erythrophthalmus
Sicalis luteola
Zonotrichia capensis
Aphrastura spinicauda
Sylviorthorhynchus desmursii
Curaeus curaeus
Sturnella neglecta
Lanius collurio
Mimus thenca
Anthus trivialis
Motacilla alba
Parus palustris
Parus caeruleus
Parus major
Phytotoma rara
Prunella modularis
Eugralla paradoxa
Pteroptochos megapodius
Scelorchilus rubecula
Sitta europaea
Hippolais polyglotta
Phylloscopus trochilus
Phylloscopus collybita
Regulus ignicapillus
Sylvia communis
Sylvia atricapilla
Sylvia cantillans
Sylvia undata
Thryomanes bewickii
Troglodytes aedon
Erithacus rubecula
Oenanthe hispanica
Saxicola torquata
Turdus viscivorus
Turdus merula
Anairetes parulus
Elaenia albiceps
Myiarchus cinerascens
Vireo huttoni
Colaptes auratus
Melanerpes formicivorus
Picoides medius
Picoides nuttallii
Picus canus
Nothoprocta perdicaria
Order
Family
Genus
Species

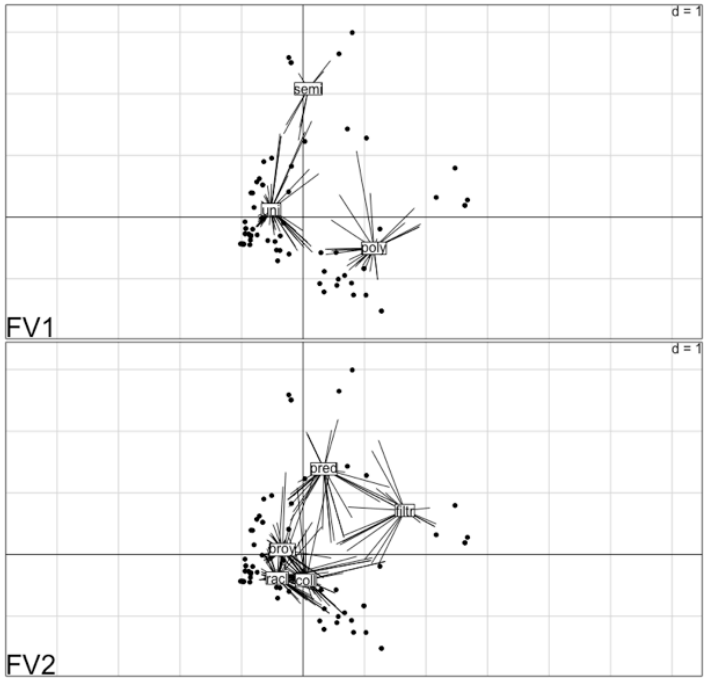
```
> traits2
    semi uni poly racl broy coll filtr pred
g1     0   3    0  9.0  0.0  1.0   0.0  0.0
g2     0   3    0 10.0  0.0  0.5   0.0  0.0
g3     0   3    0  3.0  0.0  5.0   2.0  0.0
g4     0   1    3  2.0  1.0  6.0   0.5  0.0
g5     0   0    3  3.0  3.0  4.0   0.0  0.0
g6     1   3    0  4.0  3.0  2.0   0.0  0.0
g7     0   3    0  3.0  3.0  3.0   0.0  0.0
g9     0   3    1  4.0  2.0  3.0   0.0  0.0
g10    0   3    0 10.0  0.0  0.0   0.0  0.0
g11    0   3    0  5.0  3.0  1.0   0.0  0.0
g12    0   3    0  4.5  1.5  2.5   0.0  0.0
g13    0   3    0  5.0  2.0  2.0   0.0  0.0
g14    0   3    0  5.0  3.0  1.0   0.0  0.0
g15    0   3    0  4.0  4.0  0.5   0.0  0.0
g16    0   3    0  6.0  1.0  3.0   0.0  0.0
g18    0   3    0  6.0  2.0  1.0   0.0  0.0
g19    0   3    0  5.0  2.0  3.0   0.0  0.0
g20    0   3    0  6.0  2.0  1.0   0.0  0.0
g21    0   3    1  5.0  2.0  2.0   0.0  0.0
g23    0   3    0  3.0  0.0  7.0   0.5  0.0
g24    0   3    0  0.5  0.0 10.0   0.5  0.0
g25    1   3    0  5.5  0.0  2.0   2.5  0.0
c1     0   0    3  1.0  7.0  2.0   0.0  0.5
c2     0   0    3  1.0  7.0  2.0   0.0  0.5
c3     0   0    3  1.0  2.0  2.0   0.0  3.0
c4     0   1    3  1.0  7.0  2.0   0.0  0.5
etc…
```

Trait 1 (3 modalités) : voltinisme

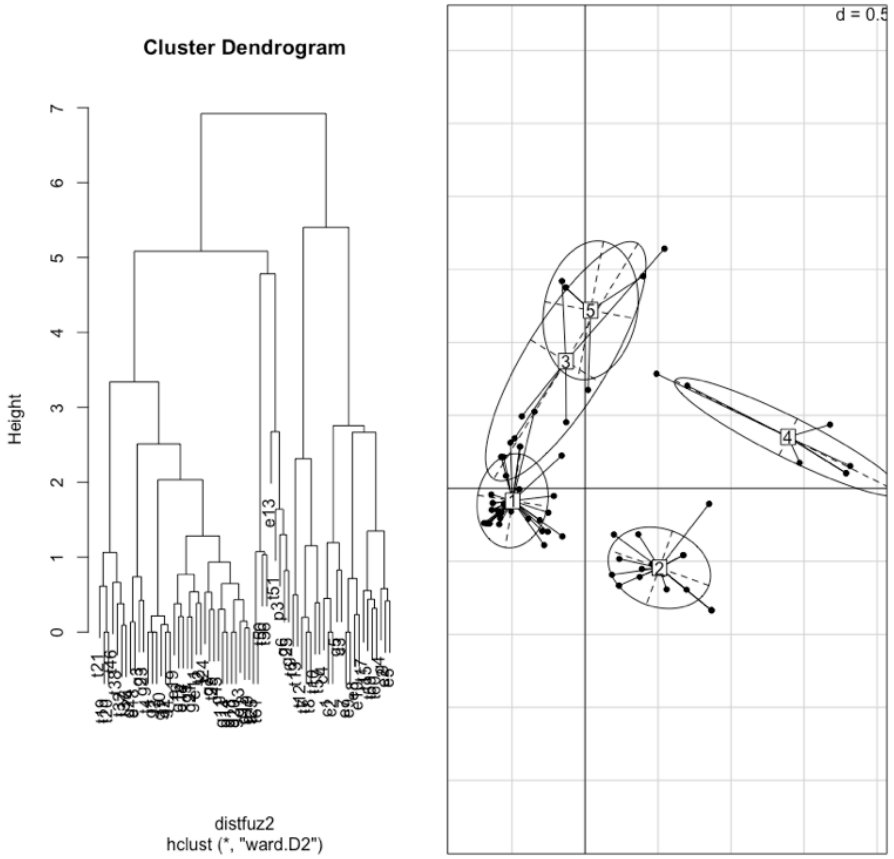
- semi : semivoltin – la génération se déroule sur plus d'un an
- uni : univoltin – la génération dure un an
- poly : polyvoltin – il y a plus d'une génération par an

Trait 2 (5 modalités) : mode trophique

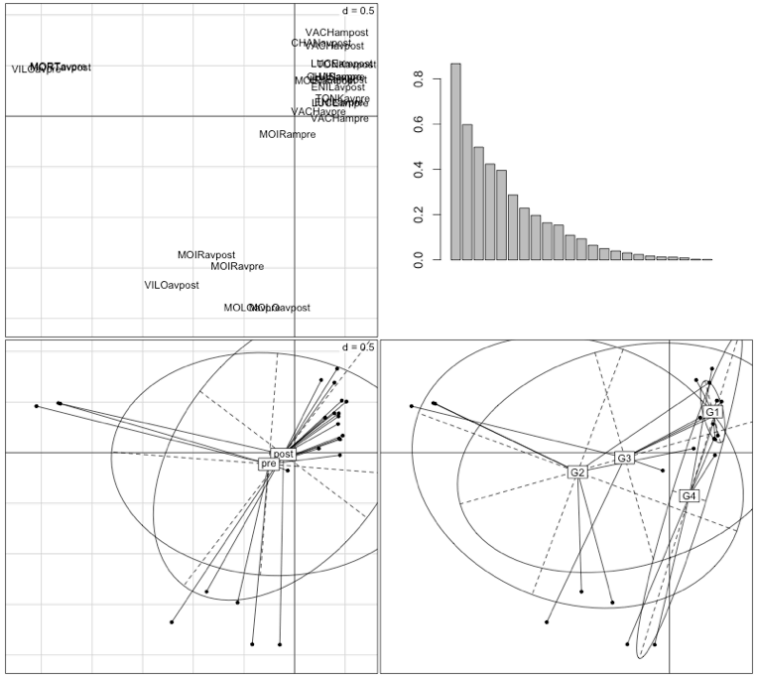
- racl : racleur – consommation de périlithon/-phyton par raclage du substrat
- broy : broyeur – fragmentation de matière organique grossière
- coll : collecteur – collection de matière organique fine associée au substrat minéral
- filtr : filtreur – collection de matière organique fine en suspension dans l'eau
- pred : prédateur.



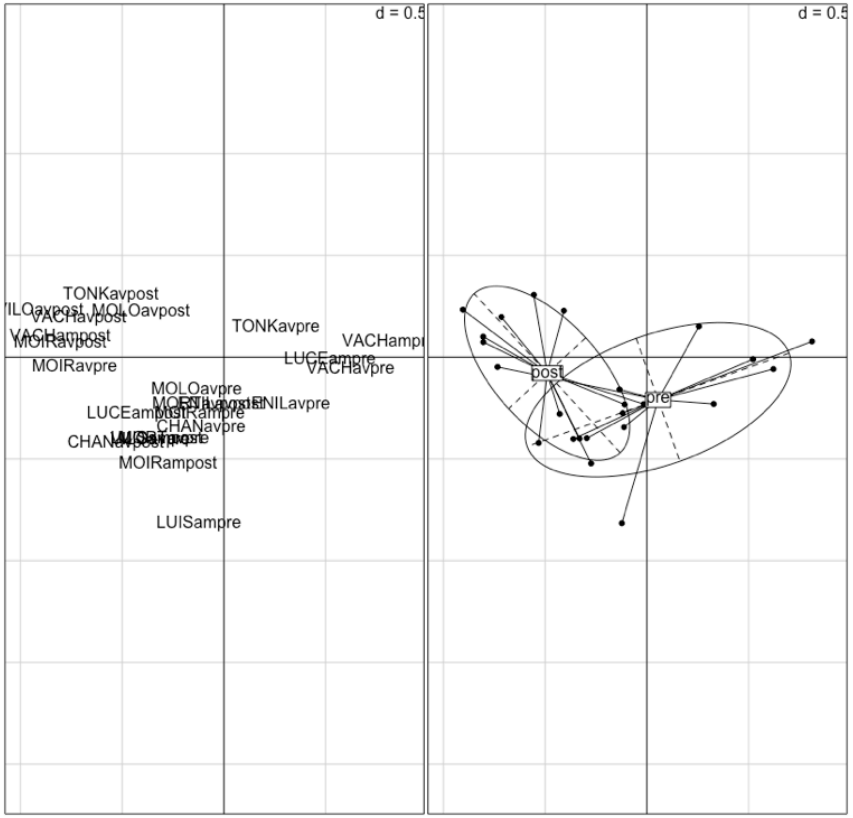
Ordination par une Analyse des Correspondances Floue (fca) du tableau contenant 67 taxons et 2 traits biologiques (voltisnisme – 3 catégories, modes trophiques – 5 catégories). Ordination des 67 taxons par les 2 premiers axes de la fca. En haut : catégories du trait « voltinisme », en bas : catégories du trait « mode trophique »



Classification par la méthode de Ward des 67 taxons à partir de leur distance dans une fca de deux traits biologiques (voltinisme et modes trophiques). A droite, les 5 groupes retenus sont projetés sur le premier plan factoriel des taxons dans la fca



*Analyse Factorielle des Correspondances (AFC) du tableau contenant l'abondance de 75 taxons dans 24 sites * périodes.*
*En haut à gauche : Ordination des sites * périodes le long de deux premiers axes factoriels (f1 – horizontal, 20% de l'information ; f2 – vertical, 14% de l'information).*
E haut à droite : Valeurs propres associées aux axes factorielles de l'AFC.
En bas à gauche : Les 24 sites regroupés graphiquement par période (pre / post).
En bas à droite : Les 24 sites regroupés graphiquement par groupe (G1 à G4).



*Ordination par une DPCoa du tableau contenant 67 taxons et 24 stations*périodes. Une distance inter-taxons est intégrée, basée sur deux traits biologiques (voltinisme et modes trophiques). A gauche : ordination des 24 stations*périodes par les 2 premiers axes de la DPCoa, à droite : les mêmes stations*périodes regroupées par période.*

FEMS MICROBIOLOGY ECOLOGY

MINIREVIEW

A guide to statistical analysis in microbial ecology: a community-focused, living review of multivariate data analyses

Pier Luigi Buttigieg[1,2,3] & Alban Ramette[1,4]

[1]HGF-MPG Group for Deep Sea Ecology and Technology, Bremerhaven, Germany; [2]Alfred Wegener Institute Helmholtz Centre for Polar and Marine Research, Bremerhaven, Germany; [3]MARUM Center for Marine Sciences, Bremen, Germany; and [4]Max Planck Institute for Marine Microbiology, Bremen, Germany

Correspondence: Pier Luigi Buttigieg, HGF-MPG Group for Deep Sea Ecology and Technology, Max Planck Institute for Marine Microbiology, Celsiusstrasse 1, 28359 Bremen, Germany. Tel.: +49 421 2028 984; fax: +49 421 2028 690; e-mail: pbuttigi@mpi-bremen.de

Present address: Alban Ramette, Institute of Social and Preventive Medicine (ISPM), University of Bern, Finkenhubelweg 11, 3012 Bern, Switzerland

Received 16 June 2014; revised 30 September 2014; accepted 6 October 2014. Final version published online 5 November 2014.

DOI: 10.1111/1574-6941.12437

Editor: Gerard Muyzer

Keywords
multivariate statistics; online resource; interactive guide; complex data.

Abstract

The application of multivariate statistical analyses has become a consistent feature in microbial ecology. However, many microbial ecologists are still in the process of developing a deep understanding of these methods and appreciating their limitations. As a consequence, staying abreast of progress and debate in this arena poses an additional challenge to many microbial ecologists. To address these issues, we present the GUide to STatistical Analysis in Microbial Ecology (GUSTA ME): a dynamic, web-based resource providing accessible descriptions of numerous multivariate techniques relevant to microbial ecologists. A combination of interactive elements allows users to discover and navigate between methods relevant to their needs and examine how they have been used by others in the field. We have designed GUSTA ME to become a community-led and -curated service, which we hope will provide a common reference and forum to discuss and disseminate analytical techniques relevant to the microbial ecology community.

Introduction

Multivariate statistical analyses are typically used to summarise high-dimensional data, test hypotheses involving multiple response variables, and examine relationships between large sets of variables (Legendre & Legendre, 1998; Härdle & Simar, 2007). The use of multivariate analyses is supplanting 'simple' descriptive analyses across ecology (see James & McCulloch, 1990 and Økland, 2007 for comment) and has become common in microbial ecology, where complex, multidimensional data sets abound (e.g. Ramette, 2007; Bertics & Ziebis, 2009; Frossard *et al.*, 2012; Thioulouse *et al.*, 2012; Hartmann *et al.*, 2013; Rivers *et al.*, 2013). Indeed, numerous software tools used by microbial ecologists implement multivariate analysis techniques and have been recommended as standard components of, for example, microbiome analysis (Kuczynski *et al.*, 2012) and environmental studies (e.g. Zinger *et al.*, 2012). Notable examples include the MOTHUR software (Schloss *et al.*, 2009), the Quantitative Insights Into Microbial Ecology (QIIME) platform (Caporaso *et al.*, 2010), the PHYLOSEQ package (McMurdie & Holmes, 2013) and the Biodiversity Virtual e-Laboratory (BIOVEL; http://www.biovel.eu/) project. While such developments may lead one to conclude that standard statistical recipes and 'workflows' now exist for microbial ecology data, it is vital to recognise that gauging the appropriateness of a given technique to the data and phenomena under investigation is not necessarily a 'cut and dried' affair.

Firstly, it is essential to recognise that the application of statistical techniques to ecological data is the focus of a living field of study: numerical ecologists and statisticians routinely re-evaluate the properties and limitations of even well-known techniques in relation to ecological

MINIREVIEW

Multivariate analyses in microbial ecology

Alban Ramette

Microbial habitat group, Max Planck Institute for Marine Microbiology, Bremen, Germany

Correspondence: Alban Ramette, Microbial habitat group, Max Planck Institute for Marine Microbiology, Celsiusstrasse 1, 28359 Bremen, Germany. Tel.: +49 421 2028 863; fax: +49 421 2028 690; e-mail: aramette@mpi-bremen.de

Received 17 January 2007; revised 18 July 2007; accepted 20 July 2007.
First published online 25 September 2007.

DOI:10.1111/j.1574-6941.2007.00375.x

Editor: Ian Head

Keywords
ordination; multivariate; modeling; statistics; gradient.

Abstract

Environmental microbiology is undergoing a dramatic revolution due to the increasing accumulation of biological information and contextual environmental parameters. This will not only enable a better identification of diversity patterns, but will also shed more light on the associated environmental conditions, spatial locations, and seasonal fluctuations, which could explain such patterns. Complex ecological questions may now be addressed using multivariate statistical analyses, which represent a vast potential of techniques that are still underexploited. Here, well-established exploratory and hypothesis-driven approaches are reviewed, so as to foster their addition to the microbial ecologist toolbox. Because such tools aim at reducing data set complexity, at identifying major patterns and putative causal factors, they will certainly find many applications in microbial ecology.

Introduction

Microbial 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 (e.g. Hughes Martiny *et al.*, 2006; Ramette & Tiedje, 2007a, b). Because DNA sequences are being accumulated at an unprecedented rate due to high-throughput technologies such as pyrosequencing (Edwards *et al.*, 2006a, b), single-cell genome sequencing (Zhang *et al.*, 2006), or metagenomics (Venter *et al.*, 2004; Field *et al.*, 2006; Gill *et al.*, 2006), future challenges will very likely consist of interpreting the observed diversity patterns as a function of contextual environmental parameters. This would help answer fundamental questions in microbial ecology such as whether microbial diversity responds qualitatively and quantitatively to the same factors as macroorganism diversity (Horner-Devine *et al.*, 2004; van der Gast *et al.*, 2005; Green & Bohannan, 2006; Hughes Martiny *et al.*, 2006).

Most obstacles encountered by microbial ecologists when they try to summarize and further explore large data sets concern the choice of the adequate numerical tools to further evaluate the data statistically and visually. Such tools, which have been developed by community ecologists to work on distribution and diversity patterns of plants and animals, could be readily applied in microbial ecology. Although multivariate analyses of community diversity patterns are well described in the literature, microbial ecologists have used multivariate analyses either rarely or mostly for exploratory purposes. A brief survey of the literature confirms this trend (Table 1; Fig. 1). Table 1 indicates that bacterial studies rank third after plant and fish studies for their use of multivariate analyses. Complex data sets are mostly explored via principal component analysis, or cluster analysis, and hypothesis-driven techniques such as redundancy analysis, canonical correspondence analysis (CCA), or Mantel tests are more rarely used (Fig. 1). Axis 1 (horizontal) clearly differentiates microscopic (bacteria, microorganisms, fungi) from macroscopic (fish, bird, plant, insect) life, and this may be related to the use of more exploratory methods (e.g. cluster analysis, PCA) in the first group. It is important to state that the figures presented in Table 1 and Fig. 1 have to be taken with caution because many articles do not include a description of statistical approaches in their titles or abstracts, and so

MINIREVIEW

Multivariate analyses in microbial ecology

Alban Ramette

Microbial habitat group, Max Planck Institute for Marine Microbiology, Bremen, Germany

FEMS Microbiol Ecol **62** (2007) 142–160

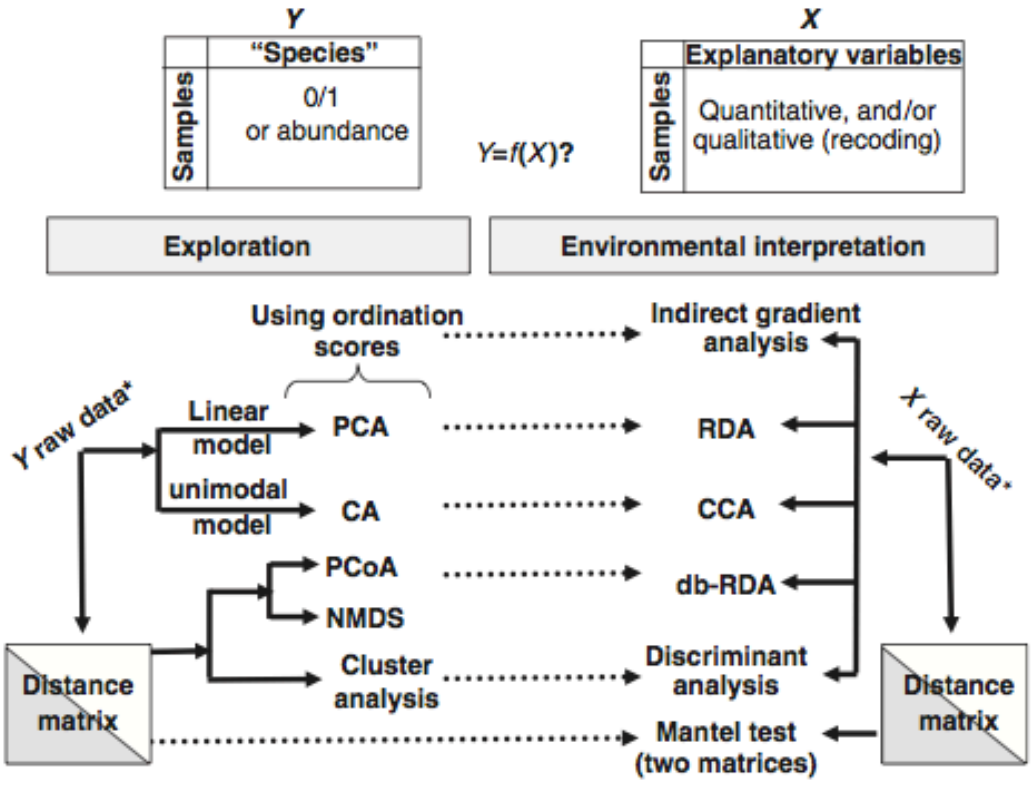


Fig. 4. Relationships between numerical methods. Exploratory tools such as PCA, CA, PCoA, NMDS, or cluster analysis can be applied to a sample-by-species table to extract the main patterns of variation, to identify groups or clusters of samples, or specific species interactions. Sample scores on the main axes of variation can be related to variation in environmental variables using indirect gradient analyses. When a constrained analysis is desired (i.e. direct gradient analysis), RDA, db-RDA, CCA, or linear discriminant analysis can be used as extensions of the unconstrained methods. Mantel tests are appropriate to test the significance of the correlation between two distance matrices (e.g. one based on species data and the other on environmental variables). Raw data may be transformed, normalised or standardised as appropriate before analysis.

Famille Vanin, Miribel 1904