Zeitschrift: Candollea : journal international de botanique systématique = international journal of systematic botany

Herausgeber: Conservatoire et Jardin botaniques de la Ville de Genève

Band: 60 (2005)

Heft: 2

Artikel: Discriminant analysis of the spatial distribution of plant species occurrences : I. Theoretical aspects

Autor: Calenge, Clément / Spichiger, Rodolphe / Chessel, Daniel

DOI: https://doi.org/10.5169/seals-879287

Candollea 60(2): 563-575 (2005)

Discriminant analysis of the spatial distribution of plant species occurrences: I. Theoretical aspects

CLÉMENT CALENGE,
RODOLPHE SPICHIGER,
DANIEL CHESSEL
&
CYRILLE CHATELAIN

ABSTRACT

CALENGE, C., R. SPICHIGER, D. CHESSEL & C. CHATELAIN (2005). Discriminant analysis of the spatial distribution of plant species occurrences: I. Theoretical aspects. *Candollea* 60: 563-575. In English, English and French abstracts.

This paper is the first of two articles describing theoretical and practical aspects of spatial discrimination of species distributions in a given area. Three multivariate methods used to show species geographical zonation from lists of species occurrence data from herbarium records are discussed. These methods are (i) *Correspondence Analysis* (CA), (ii) *Canonical Correlation Trend Surface Analysis* (CCTSA), and (iii) *Discriminant Analysis on Eigenvectors of Neighbourhood Operator* (DAENO). Their use is illustrated through the analysis of the spatial distribution of three virtual species. This paper shows that these methods are forms of *Discriminant Analysis* (DA) which use the spatial position of the species occurrences as variables; they just differ in the way they measure this position in space. It is concluded that these three methods are likely to produce the same results from a given data set, providing that the underlying spatial structures are well defined within the study area.

RÉSUMÉ

CALENGE, C., R. SPICHIGER, D. CHESSEL & C. CHATELAIN (2005). L'analyse discriminante de la distribution spatiale d'occurrences d'espèces végétales: I. Aspects théoriques. *Candollea* 60: 563-575. En anglais, résumés anglais et français.

Cet article est le premier d'une série de deux articles décrivant les aspects théoriques et pratiques de la discrimination spatiale de distributions d'espèces dans une zone donnée. Trois méthodes multivariées sont discutées ici et peuvent être utilisées pour mettre en évidence la zonation géographique d'espèces à partir de listes d'occurrences d'espèces issues de données d'herbier. Ces méthodes sont: (i) l'*Analyse Factorielle des Correspondances* (AFC), (ii) l'*Analyse Canonique des Corrélations Appliquées aux Tendances de Surface* (ACCATS), et (iii) l'*Analyse Discriminante sur Vecteurs Propres du Graphe de Voisinage* (ADVPGV). Leur utilisation est illustrée par la distribution spatiale de trois espèces virtuelles. Cet article montre que ces méthodes sont des formes d'*Analyse Discriminante* (AD) qui utilisent comme variables la position spatiale des occurrences d'espèces, et qu'elles ne diffèrent entre elles que par la manière de mesurer dans l'espace la position des espèces. Il est conclu que ces méthodes donnent approximativement le même résultat, pour peu que les structures spatiales sous-jacentes soient suffisamment bien définies sur la zone d'étude.

KEY-WORDS: Plant biogeography – Herbarium records – List of species occurrences – Multivariate statistics – Canonical Correlation Trend Surface Analysis (CCTSA) – Correspondence Analysis (CA) – Discriminant Analysis (DA) – Discriminant Analysis on Eigenvectors of Neighbourhood Operator (DAENO) – Point pattern – Spatial analysis

ISSN: 0373-2967

1. Introduction

The understanding of the geographical zonation of different plant species in a given region is of importance for ecologists because it may help them to develop hypotheses on the ecological history of a region and on the underlying environmental variables driving the observed vegetation patterns (HIRZEL & al., 2003; SPICHIGER & al., 2004). To identify zonation in a given area, many ecological studies focus on the relations between the distributions of several selected species. They try to define a typology grouping together species showing a similar distribution and emphasizing the differences between species showing dissimilar distributions (AUSTIN, 1985; GIMARET-CARPENTIER & al., 2003).

In phytosociology, this issue has been raised frequently (TER BRAAK, 1985; TER BRAAK & LOOMAN, 1986). This field of ecology focuses on processes occurring at a relatively small scale, and the studies of spatial zonation of species is generally based on a systematic sampling of the study area. For example, a virtual grid is superimposed over the study area, and the number of individuals of species sampled in each quadrat is recorded. The statistical methods used to analyse such data are essentially multivariate, as the number of studied species is often large. Ordination methods such as correspondence analysis are generally used in these surveys (HILL, 1974). These methods are optimal when used to show the distribution patterns and vegetation structures in a given area (TER BRAAK & LOOMAN, 1986).

On a larger scale (e.g. continental scale), such designs are difficult to implement, and biologists often look for other sources of information to identify this geographical zonation. In biogeographical studies existing collections of herbarium specimens may be of great use because the location from where each sample was collected is known (GIMARET-CARPENTIER & al., 2003). In this paper, we use the term "occurrence" to describe such data. According to GIMARET-CARPENTIER (1999), an occurrence corresponds to the location of a species specimen in an area that was gathered by an ecologist in an observational study. In the case of herbarium specimens, an occurrence is therefore characterised by three variables: its X and Y coordinates, and the species to which it belongs.

A list of occurrences is fundamentally different from the data collected in a study using systematic sampling. From a statistical point of view, species occurrences have a particular status. The location of an occurrence at a given point ensures that the species was present there, whereas the absence of an occurrence does not guarantee the absence of the species. An absence may also indicate that during the sampling this species was not recorded at the location (GREEN, 1971; HIRZEL & al., 2003). The spatial distribution of the occurrences reflects the distribution of the species as well as the distribution of the sampling intensity; this distribution is not necessarily uniform across the area (WILLIS & al., 2003). The aim here was to discriminate between species over the sampled area and not to describe the actual distribution of a single species. If it is assumed that the probability of collecting an occurrence at a given point does not vary among the different species considered then the lists of occurrences can be used to identify the geographical variation in species composition. They can be used even if this probability varies spatially.

In this paper we focus on three statistical methods that can be used to emphasize the geographical zonation of species in a given region based on lists of occurrences: (i) *Correspondence Analysis* (HILL, 1974), (ii) *Canonical Correlation Trend Surface Analysis* (WARTENBERG, 1985; GIMARET-CARPENTIER & al., 2003), and (iii) *Discriminant Analysis on Eigenvectors of Neighbourhood Operator* (MÉOT & al, 1993; SPICHIGER & al., 2004). Since these three methods rely on the same mathematical and conceptual base our aims were to present mathematical connections between them, and to explore the similarity of results using an example dataset. This virtual example simulates a gradient formed by three virtual species *A*, *B*, and *C* in a study area with 50 occurrences per species (Fig. 1).

2. Correspondence Analysis (CA)

Correspondence Analysis (CA) has been widely used in ecology over the last thirty years. This method has been introduced independently by several authors for the analysis of contingency tables built from the systematic sampling of a given area (ROUX & ROUX, 1967; HATHEWAY, 1971; HILL, 1973). The contingency table **T** (*Q* quadrats × *S* species) contains the number of individuals of each species in each sampled quadrat. CA assigns a set of numerical values to each quadrat and to each species, such that both the discrimination of the species scores between quadrats and the discrimination of the quadrats scores between species is maximum (THIOULOUSE & CHESSEL, 1992). This analysis is optimal to emphasize the spatial organisation of plant species using this kind of sampling scheme (TER BRAAK, 1985).

This method may also be used to analyse the list of species occurrences (Fig. 2A). A grid is superimposed onto the study area, and the occurrences of each species are numbered in each quadrat of the grid. This leads to a contingency table **T**, with *Q* rows (i.e. *Q* quadrats of the grid), and *S* columns (i.e. *S* species). Correspondence analysis of **T** assigns scores to both species and quadrats such that two quadrats with similar floristic composition have similar scores, and two species with a similar distribution in the quadrats of the grid also have similar scores. By maximizing the variance of these scores, this analysis maximizes the spatial floristic variation in the study area. This property allows the main structures and patterns present in the study area to be highlighted.

Correspondence analysis of the contingency table built from our virtual examples assigns positive scores to quadrats located at the northwest corner of the study area, and negative scores to the quadrats located at the southeast corner. The simulated gradient is therefore described well by these scores. The structures highlighted by the analysis may be interpreted through the distribution of the occurrence scores by species. For example, the occurrences of species *A* are restricted to the negative side of the factorial axis, indicating that this species is distributed in the northwest of our study area. Likewise, species *C* is restricted to the southeast corner of the study area, and species *B* is intermediate. The main drawback of this method is that the user has to choose the quadrat size, and that results obtained may depend on the size chosen. It can be further argued that the use of a grid of quadrats might hide local features of the data, and therefore "oversmooth" the data.

3. Canonical Correlation Trend Surface Analysis (CCTSA)

Discriminant Analysis (DA) is another possibility that can be used to emphasize geographical patterns displayed by the studied species. DA is optimal when the objective is to separate *a priori* groups using measurements for individuals from several variables (GREEN, 1971; MANLY, 1994). This method returns several linear combinations of the descriptive variables which maximises the discrimination between groups. For example, if X_1, X_2, and X_3 are the descriptive variables, the analysis returns a new synthetic variable $Y = a_1 \times X_1 + a_2 \times X_2 + a_3 \times X_3$, and the coefficients a_1, a_2, and a_3 are computed such that the ratio (between-group variance of Y) / (total variance of Y) is maximum. This analysis therefore assigns a numerical score to each occurrence so that the percentage of variation of the scores explained by the factor species is as high as possible.

The geographical zonation of the species in the study area is shown in Figure 1. The distributions of tree species can be worked out using the variables measuring the spatial distribution of their occurrences (Fig. 2B). Polynomial functions of geographical coordinates of occurrences are often used to achieve this, with f.i. variables x, y, x^2, y^2, $x \times y$, x^3, etc., where x and y are the occurrence coordinates (GIMARET-CARPENTIER & al., 2003). Discriminant analysis of species by such polynomial functions has often been used in ecology and is called *Canonical Correlation Trend Surface Analysis* (CCTSA) (GITTINS, 1968; WARTENBERG, 1985; GIMARET-CARPENTIER & al., 2003).

The analysis maximizes the distinctions between species according to the spatial distribution of their occurrences. As for many statistical techniques, the discriminant analysis requires a number of occurrences much larger than the number of variables. However, there is no strict rule to choose the number of polynomial functions to be used in this analysis (MANLY, 1994). With sample sizes commonly encountered in the literature, this number rarely exceeds 10 (e.g. GIMARET-CARPENTIER & al., 2003).

CCTSA is mathematically equivalent to the commonly used *Canonical Correspondence Analysis* (CCA) (TER BRAAK, 1986). The difference is that CCA requires a grid of quadrats (in this case the sampling unit is the quadrat, as in CA), whereas this is not necessary in DA (where the sampling unit is the occurrence). CCA is therefore a special form of DA (TER BRAAK & VERDONSCHOT, 1995).

We present an application of CCTSA using a virtual example. As with CA, CCTSA assigns very positive scores to occurrences located at the northwest corner of the study area, (Fig. 1) and very negative scores to the occurrences located at the southeast corner of the area (Fig. 2B). The analysis clearly emphasizes the gradient simulated in the study area. Smoothing methods, such as lowess regression (CLEVELAND & DEVLIN, 1988), may be used to smooth the occurrences scores in the study area; this may give an even clearer picture of the structure. As in CA the distribution of the scores by species reflects the position of the species on the gradient. The histograms reveal that species *A* is located at the northwestern end of the gradient, species *C* at the southwestern end, and species *B* falls between species *A* and *C*.

Thus, CCTSA gives the same conclusions as CA without any need of excessive discretization. For this reason, CCTSA has been recommended by many authors (GITTINS, 1968; WARTENBERG, 1985; GIMARET-CARPENTIER & al., 2003). The main drawback of this method is that the polynomial functions of occurrence coordinates are often strongly correlated (e.g. the x-coordinate is strongly dependent on its square x^2). This is a problem because DA requires that the discriminating variables are not too strongly correlated (MANLY, 1994). In addition, this correlation implies that a large number of polynomial functions have to be included in the analyses to account for a large part of the spatial variation. To circumvent this drawback some authors have proposed the use of orthogonal polynomials in this analysis instead of the classical polynomials (BORCARD & LEGENDRE, 1994).

4. Discriminant Analysis on Eigenvectors of Neighbourhood Operator (DAENO)

An interesting alternative to CCTSA is to replace the set of polynomial functions of geographical coordinates by a set of coordinates which position the occurrences relative to the other occurrences of the pattern. This type of method constitutes the core of *Discriminant Analysis on Eigenvectors of Neighbourhood Operator* (DAENO).

Firstly, a network of neighbouring relationships relating the occurrences of the pattern needs to be computed. Several algorithms are available to compute such networks (e.g. GABRIEL & SOKAL, 1969; PACE & ZOU, 2000). For simplicity here we use Delaunay triangulation (UPTON & FINGLETON, 1985). This network is built from a Voronoi tessellation of the occurrence pattern (UPTON & FINGLETON, 1985). This tessellation is special kind of decomposition of the study area into a set of polygons, determined by distances to the occurrence pattern (Fig. 3A). Each polygon is associated with an occurrence and includes all the places closer to this occurrence than to any other. The Delaunay Triangulation is derived from this tessellation, and connects two occurrences by a line if their Voronoi polygons share a common edge. In the following analyses two occurrences connected by a line on this graph are considered as neighbours.

It is possible to build a matrix describing the network of neighbouring relationships from this graph. Let **V** be a matrix with *I* rows and *I* columns (where *I* is the total number of occurrences, all species being pooled). At the intersection of the row *i* and of the column *j*, the matrix contains

1 if the i^{th} occurrence is a neighbour of the j^{th} occurrence, and 0 otherwise. The matrix **V** is called neighbourhood operator and it can be used to derive a set of scores which describe the position of occurrences.

The diagonal matrix $\mathbf{D}_n$ is computed as following:

$$\mathbf{D}_n = \mathrm{Diag}(\mathbf{V}\ \mathbf{1}_n)$$

where $\mathbf{1}_n$ is the n-vector of 1.

At the intersection of the row i and of the column i, $\mathbf{D}_n$ contains the number of neighbours of the occurrence i. Then, the matrix **S** is computed by:

$$\mathbf{S} = \frac{1}{m}\mathbf{D}_n - \frac{1}{m}\mathbf{V}$$

where m is equal to the number of pairs of neighbours (the sum of all values in **V**).

The diagonalization of **S** returns a set of eigenvectors that are orthogonal (i.e. uncorrelated). The first vector assigns a numerical score to each occurrence so that the spatial autocorrelation of the scores is as high as possible (Fig. 3B). In other words, two neighbouring occurrences will have similar scores whereas two very distant occurrences will have very different scores. The second vector maximizes the spatial autocorrelation under the constraint of orthogonality with the first vector, and so on. Theoretical justifications for the above formulas can be found in MÉOT & al. (1993).

Eigenvectors from the diagonalization of **S** can be used in spatial analysis in place of the polynomials of geographical coordinates, such as f.i. in DAENO. This latter analysis has many advantages over CCTSA. Firstly, whereas the Cartesian coordinate system (i.e. x, y) gives the position of a point in space relative to an arbitrary point of reference (i.e. the origin of space, with coordinates x = 0 and y = 0), the eigenvectors of neighbourhood operators give the position of an occurrence relative to the whole occurrence pattern (i.e. the studied system). Secondly, two occurrences may be close in geographic space, i.e. with similar Cartesian coordinates, but separated by an impassable boundary (e.g. a mountain). The polynomial functions of geographical coordinates do not take into account this possibility whereas the eigenvectors of the neighbourhood operator do, provided that the neighbouring relationships between occurrences on either side of the boundary are deleted from the matrix **V**. Finally, the eigenvectors of the neighbourhood operator are orthogonal, i.e. uncorrelated. This property important when these variables are used as descriptive variables in a discriminant analysis (MANLY, 1994). MORELLET (1998) noted that the eigenvectors of neighbouring operator explains a larger quantity of spatial variability than polynomial functions of coordinates. For all these reasons, the eigenvectors of the neighbourhood operator have been recommended in spatial analyses by THIOULOUSE & al. (1995).

Practical application of DAENO using virtual species distributions

Using Delaunay triangulation we computed a network of neighbouring relationships. Figure 3A shows the resulting neighbouring relationships between the occurrences. From this graph we derived the neighbourhood operator. The neighbourhood operator was then diagonalized and the 12 first eigenvectors of this matrix were kept for DAENO, although there is no strict rule to decide how many eigenvectors should be kept for further analyses (Fig. 3B). As previously mentioned, the analysis shows the simulated gradient. Very positive scores are seen for occurrences in the northwest corner and very negative scores for the occurrences in the southeast corner (Fig. 2C). Note that the difference between the northwest and the southeast appears even more clearly with this analysis than with CCTSA. Likewise, species *A* and *C* are more strongly discriminated with DAENO.

5. Discussion

The three methods presented above may be used to discriminate species according to their spatial distribution. CCTSA and DAENO are forms of DA. CA is also a form of DA (THIOULOUSE & CHESSEL, 1992) achieved through the superimposing of a grid onto the occurrence pattern. A matrix **W** can be derived, with *I* rows (the *I* occurrences) and *Q* columns (the *Q* quadrats). At the intersection of the row *i* and of the column *j*, this matrix contains 1 if the i^{th} occurrence belongs to the quadrat *j*, and 0 otherwise. Discriminant analysis of the species by the table **W** returns exactly the same results as CA of the contingency table **T** described above.

These three methods are all discriminant analyses of species using variables measuring the position of occurrences. The only difference is the measure of the spatial position of occurrences; this measure involves, for each method respectively:

- quadrats in CA;
- polynomials of the geographical coordinates in CCTSA;
- eigenvectors of neighbourhood operator in DAENO.

Despite the fact that these variables differ from one method to another the results are similar whatever the method used. All three analyses are likely to identify major patterns in the data. However, theoretical arguments lead us to prefer DAENO, since the eigenvectors of neighbourhood operator have a higher ecological meaning (THIOULOUSE & al., 1995). We used Delaunay triangulation to generate neighbouring relationships but we stress that numerous other methods are available in the literature to build such a network (for a review, see OLLIER, 2005).

As stressed by PALMER (1993), "it is one thing to entrust the validity of equations in the abstract, and yet another to entrust our data to them". A biological application of DAENO is presented in this issue (see SPICHIGER & al., 2005) showing how DAENO may be used to analyse the geographical zonation of several tree species in Paraguay using herbarium specimens. Point patterns are very common in ecology and the need for methods dealing with such data is still considerable (UPTON & FINGLETON, 1985). CA, CCTSA and DAENO may have a role in this context. Despite our theoretical preference for DAENO the other methods are also valid tools which may be preferred because of their wider availability in commercial statistical software. These three methods are freely available (see ade4 package for R software, downloadable at http://cran.r-project.org).

ACKNOWLEDGEMENTS

We would like to thank Michelle Price for correcting the English.

REFERENCES

AUSTIN, M. P. (1985). Continuum concept, ordination methods, and niche theory. *Ann. Rev. Ecol. Syst.* 16: 39-61.

BORCARD, D. & P. LEGENDRE (1994). Environmental control and spatial structure in ecological communities: an example using oribatid mites (Acari, Oribatei). *Environm. Ecol. Statist.* 1: 37-61.

CLEVELAND, W. S. & S. J. DEVLIN (1988). Locally weighted regression: an approach to regression analysis by local fitting. *J. Amer. Statist. Assoc.* 83: 596-610.

GABRIEL, K. R. & R. R. SOKAL (1969). A new statistical approach to geographic variation analysis. *Syst. Biol.* 18: 259-278.

GIMARET-CARPENTIER, C. (1999). *Analyse de la biodiversité à partier d'une liste d'occurrences d'espèces: nouvelles methods d'ordination appliqués à l'étude de l'endémisme dans les Ghâts occidentaux.* Thèse de doctorat. Université de Lyon.

GIMARET-CARPENTIER, C., S. DRAY & J. P. PASCAL (2003). Large-scale biodiversity pattern analyses of the endemic tree flora of the western Ghats (India) using canonical correlation analysis of point data. *Ecography* 26: 429-444.

GITTINS, R. (1968). Trend-surface analysis of ecological data. *J. Ecol.* 56: 845-869.

GREEN, R. H. (1971). A multivariate statistical approach to the Hutchinsonian niche: bivalve molluscs of central Canada. *Ecology* 52: 543-556.

HATHEWAY, W. H. (1971). Contingency-table analysis of rain forest vegetation. *In*: PATIL, G. P., E. C. PIELOU & W. E. WATERS (ed.), *Statistical Ecology. Vol.3: Many species populations, ecosystems and systems analysis*: 271-314 Pennsylvania State University Press.

HILL, M. O. (1973). Reciprocal averaging: an eigenvector method of ordination. *J. Ecol.* 61: 237-249.

HILL, M. O. (1974). Correspondence analysis: A neglected multivariate method. *Appl. Statist.* 23: 340-354.

HIRZEL, A. H., J. HAUSSER, D. CHESSEL & N. PERRIN (2003). Ecological-Niche Factor analysis: How to compute habitat-suitability maps without absence data? *Ecology* 83: 2027-2036.

MANLY, B. F. J. (1994). *Multivariate Statistical Methods. A primer, Second edition*, Chapman & Hall.

MÉOT, A., D. CHESSEL & R. SABATIER. (1993). Opérateurs de voisinage et analyse des données spatio-temporelles. *In*: LEBRETON, J. D. & B. ASSELAIN *Biométrie et Environnement*, 45-71. Masson.

MORELLET, N. (1998). *Des outils biométriques appliqués aux suivis des populations animales: l'exemple des cervidés.* Thèse de doctorat. Université de Lyon.

OLLIER, S. (2005). *Des outils pour l'intégration des contraintes spatiales, temporelles et évolutives en analyse de données écologiques.* Thèse de doctorat. Université de Lyon.

PACE, R. K. & D. ZOU (2000). Closed-form maximum likelihood estimates of nearest neighbour spatial dependence. *Geogr. Analysis* 32: 154-172.

PALMER, M. W. (1993) Putting things in even better order: the advantages of canonical correspondence analysis. *Ecology* 74: 2215-2230.

ROUX, G. & M. ROUX (1967). A propos de quelques méthodes de classification en phytosociologie. *Rev. Statist. Appl.* 15: 59-72.

SPICHIGER, R., C. CALENGE & B. BISE (2004). The geographical zonation in the Neotropics of tree species characteristic of the Paraguay-Paraná Basin. *J. Biogeogr.* 31: 1489-1501.

SPICHIGER, R., C. CALENGE & B. BISE (2005). The discriminant analysis of the spatial distribution of plant species occurrences: II. Distribution of major tree communities in Paraguay. *Candollea* 60: 577-593.

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 gradient analysis. *Ecology* 67: 1167-1179.

TER BRAAK, C. J. F. & C. W. N. LOOMAN (1986) Weighted averaging, logistic regression and the Gaussian response model. *Vegetatio* 65: 3-11.

TER BRAAK, C. J. F. & P. F. M. VERDONSCHOT (1995). Canonical correspondence analysis and related multivariate methods in aquatic ecology. *Aquatic Sci.* 57: 255-289.

THIOULOUSE, J., & D. CHESSEL. (1992) A method of reciprocal scaling of species tolerance and sample diversity. *Ecology* 73: 670-680.

THIOULOUSE, J., D. CHESSEL & S. CHAMPELY (1995). Multivariate analysis of spatial patterns: a unified approach to local and global structures. *Environm. Ecol. Statist.* 2: 1-14.

UPTON, G. & B. FINGLETON (1985). *Spatial data analysis by example. Vol 1: Point pattern and quantitative data.* John Wiley and Sons.

WARTENBERG, D. (1985) Canonical trend surface analysis: a method for describing geographic pattern. *Syst. Zool.* 34: 259-279.

WILLIS, F., J. MOAT & A. PATON (2003). Defining a role for herbarium data in Red List assessments: a case study of Plectranthus from eastern and southern tropical. *Biodivers. & Conservation* 12: 1537-1552.

Submitted on June 27, 2005
Accepted on July 12, 2005

Addresses of the authors: CCa, DC: Université Claude Bernard Lyon, UMR CNRS 5558, Laboratoire de Biométrie et Biologie Evolutive, 43, boulevard du 11 novembre 1918, 69622 Villeurbanne Cedex, France. Email (CCa): calenge@biomserv.univ-lyon1.fr

RS, CCh: Conservatoire et Jardin botaniques de la Ville de Genève, ch. de l'Impératrice 1, CP 60, CH-1292 Chambésy, Switzerland. Email (RS): rodolphe.spichiger@cjb.ville-ge.ch

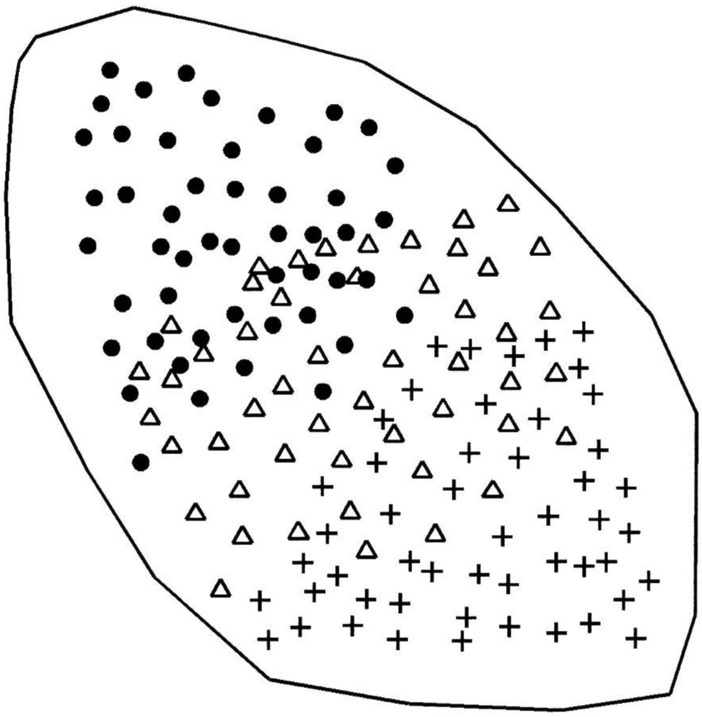

Fig. 1. – Distribution of the three virtual species studied in this paper: 50 occurrences were simulated for three species A (•), B (Δ) and C (+). The three species are distributed along a gradient from the northwest of the area toward the southwest. The polygon defines the study area.

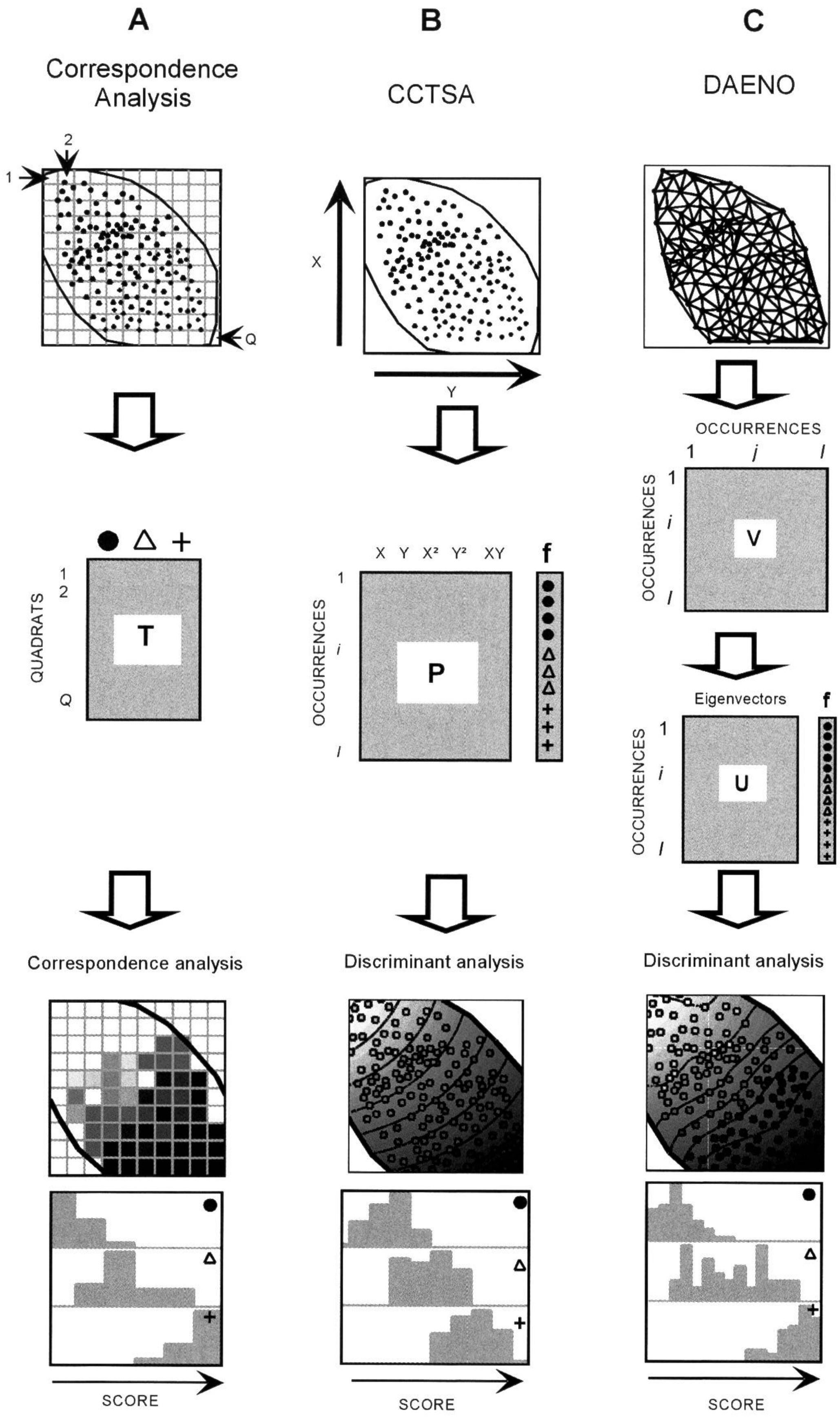
A
Correspondence Analysis
B
CCTSA
C
DAENO
2
1
Q
X
Y
OCCURRENCES
1 j l
OCCURRENCES
1
i
l
V
QUADRATS
1
2
Q
T
X Y X² Y² XY
f
OCCURRENCES
1
i
l
P
Eigenvectors
f
OCCURRENCES
1
i
l
U
Correspondence analysis
Discriminant analysis
Discriminant analysis
SCORE
SCORE
SCORE

Fig. 2. – Basic principle of the three statistical methods compared. Each method is used to analyse the geographical zonation of the three virtual species A (•), B (Δ) and C (+). The spatial distribution of the occurrences of these species is displayed on Fig. 1.

A. *Correspondence analysis* (CA). A virtual grid is superposed to the study area, and the occurrences of each species are numbered in each of the Q quadrats. Correspondence analysis is then performed on the resulting table **T**. The quadrat scores are then mapped, and the map is interpreted using the histograms of the occurrence scores for each species.

B. *Canonical correlation trend surface analysis* (CCTSA). Polynomial functions of the cartesian coordinates of the species occurrences are in the table **P**, and the species associated with each occurrence is stored in the vector **f**. Discriminant analysis of **P** by the factor **f** assigns scores to each occurrence so that the discrimination between species is maximised according to the spatial distribution of their occurrences. The occurrences scores are then smoothed over the study area using a lowess regression (smoothing is displayed in grey levels). Histograms of the occurrence scores for each species allow the interpretation of the results.

C. *Discriminant analysis on eigenvectors of neighbourhood operator* (DAENO). The neighbourhood operator **V** is a square matrix with a 1 if the occurrence in row and the occurrence in column are neighbours and 0 otherwise. The diagonalization of **V** produces a set of eigenvector, stored in a table **U**, for which the spatial autocorrelation is maximised. A discriminant analysis of **U** by the factor species **f** assigns scores to each occurrence so that the discrimination between species is maximised according to the spatial distribution of their occurrences. The occurrences scores are then smoothed over the study area using a lowess regression (smoothing is displayed in grey levels). Histograms of the occurrence scores for each species allow the interpretation of the results.

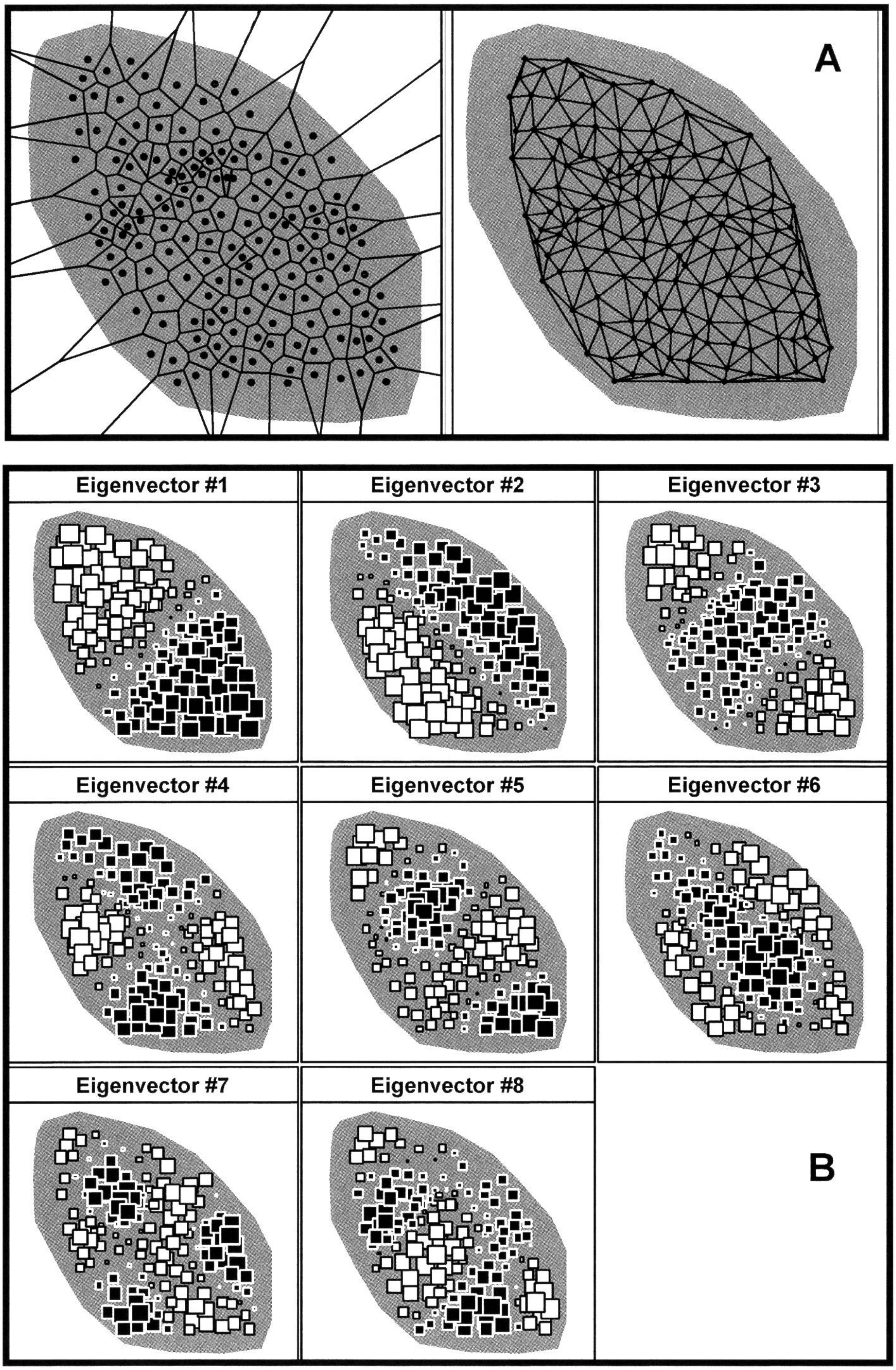
A
Eigenvector #1
Eigenvector #2
Eigenvector #3
Eigenvector #4
Eigenvector #5
Eigenvector #6
Eigenvector #7
Eigenvector #8
B

Fig. 3. – Computation of discriminant analysis on eigenvectors of neighbourhood operator.

A. Computation of the network of neighbouring relationships of the occurrences pattern using Delaunay Triangulation (right graph), the dual structure of Voronoï tessellation of the plane derived from the point pattern (left graph). Each line connects two neighbouring occurrences. The neighbourhood operator is computed from this network.

B. Map of the scores given by the first eight eigenvectors of the neighbourhood operator. Each square represents a species occurrence. For a given occurrence, the size of the square is proportional to the absolute value of the score. Black squares correspond to positive values and white squares correspond to negative values. The spatial autocorrelation of the scores is maximized on the first eigenvector. This autocorrelation is also maximised on the second vector, under the constraint of orthogonality with the first one, and so on.