Journal of Vegetation Science 13: 867-874, 2002

Matching data sets from two different spatial samples

Dray, Stéphane*; Pettorelli, Nathalie & Chessel, Daniel

*UMR CNRS 5558, Laboratoire de Biométrie et Biologie Evolutive, Université Claude Bernard Lyon 1, F-69622 Villeurbanne Cedex, France; *Corresponding author: Fax +33478892719; E-mail dray@biomserv.univ-lyon1.fr*

Abstract. Methods for coupling two data sets (species composition and environmental variables for example) are well known and often used in ecology. All these methods require that variables of the two data sets have been recorded at the same sample stations. But if the two data sets arise from different sample schemes, sample locations can be different. In this case, scientists usually transform one data set to conform with the other one that is chosen as a reference. This inevitably leads to some loss of information. We propose a new ordination method, named spatial-RLQ analysis, for coupling two data sets with different spatial sample techniques. Spatial-RLQ analysis is an extension of co-inertia analysis and is based on neighbourhood graph theory and classical RLQ analysis. This analysis finds linear combinations of variables of the two data sets which maximize the spatial cross-covariance. This provides a co-ordination of the two data sets according to their spatial relationships. A vegetation study concerning the forest of Chizé (western France) is presented to illustrate the method.

Keywords: Co-inertia analysis; Neighbourhood relationship; Ordination; RLQ analysis; Spatial cross-covariance.

Abbreviations: CA = Correspondence analysis; CCA = Canonical Correspondence Analysis; GIS = Geographic Information System; GSVD = Generalized singular value decomposition; PCA = Principal Component Analysis; RDA = Redundancy analysis.

Nomenclature: Rameau et al. (1989).

Introduction

Ecology often deals with the coupling of two data sets. In most cases, the two data sets consist of environmental and species data collected in the same sites. When the sample units are in agreement, a number of ordination methods can be used to link the two tables (ter Braak & Verdonschot 1995). Canonical Correspondence Analysis (CCA; ter Braak 1986, 1987) is probably the most frequently used method for this purpose in finding linear combinations of environmental variables that maximize the separation of species niches (Lebreton et al. 1988). There are several reasons for the success of CCA, and one of these is the dissymmetry of the approach, which uses environmental variables to model species composition by a step of multivariate regression. Redundancy Analysis (RDA; Rao 1964) contains also a multivariate regression step and can be preferable to CCA because the χ^2 metric used by CCA overemphasizes the importance of the rare species in the data set. Note that non-linear relationships can also be modelled using non-linear RDA and CCA (Makarenkov & Legendre 2002). The regression step of CCA or RDA requires that the number of environmental variables must be much lower than the number of samples, like in canonical correlation analysis. In this context, co-inertia analysis (Dolédec & Chessel 1994) is a good alternative because it is more robust than CCA regarding the number of variables compared to the number of individuals (ter Braak & Verdonschot 1995). Furthermore, co-inertia analysis has been extended, under the name of 'RLQ-method', to the case of linking three tables (Dolédec et al. 1996). This method has been named 'RLQ' because it finds linear combination of the variables of table **R** (external information about rows) and linear combinations of the variables of table **Q** (external information about columns) of maximal covariance weighted by data contained in table **L** (link table). For example, RLQ has been used to link species traits to environmental variables by way of a species by sites table (Ribera et al. 2001).

For the methods discussed, measurements of species abundances and environmental variables must have been done in the same locations. In some cases, the two

sample schemes are different and so measurements are not done in the same locations or at the same scale. In biogeographic studies, for example, environmental data are available for meteorological stations whereas species abundances are measured at the quadrat level and are available from museum or atlas data. In vegetation science, people interested in different purposes can sample the same area at different locations and scales. Hitherto, there has been no method that reconciles the two sample schemes, and the simplest way to analyse data is to estimate (e.g. by weighted averaging) the values from one table for the sample points of the other one in order to have the same sample units for the two tables (Mourelle & Ezcurra 1996; Hill 1991).

In this paper we propose a new ordination approach based on the RLQ ordination method and neighbourhood graph theory in order to link two data sets corresponding to different spatial sample plans.

Neighbourhood matrix

The first step of the analysis is to establish a neighbourhood relationship between the sites of the two sample schemes. Let us consider the situation where the first sample involves m_1 sites whereas the second involves m_2 sites. A neighbourhood matrix **G** with m_1 rows and m_2 columns must be constructed where:
$\mathbf{G}_{ij} = 1$ if site i and site j are neighbours
$\mathbf{G}_{ij} = 0$ otherwise.

This kind of matrix is currently used in spatial ordination (Thioulouse et al. 1995). In the case where sites are two-dimensional objects (i.e. quadrats, polygons...) we can easily fill this matrix by considering that:
$\mathbf{G}_{ij} = 1$ if polygon i intersects polygon j
$\mathbf{G}_{ij} = 0$ otherwise.

In the case where sites are considered as points, we propose to use tessellation to create neighbourhood relationships (Green & Sibson 1978). We consider two different spatial samples (Fig. 1). Voronoi polygons can be easily constructed for each system of sample with tessellation (Fig. 1). With these two tessellations, sites can be considered as polygons and we can apply the following decision rule:
$\mathbf{G}_{ij} = 1$ if polygon induced by site i intersects polygon induced by site j
$\mathbf{G}_{ij} = 0$ otherwise.

Note that the matrix **G** can also be filled in computing the area of the intersection between polygons with GIS. Taking into account the overlapped area will give more weight to isolated points that produce large Voronoi polygons. Moreover, we propose to define the neighbourhood relations using two tessellations but simplest method can be used. For example, methods based on nearest neighbours criteria or on intervals of Euclidean distances that are used to define neighbourhood in the case of one set of points can be extended in the case of two sets. However, our choice has been guided by the fact that other methods can produce points with no neighbours (methods based on distance) or are difficult to apply in the case of two sets and hence must be devised by the user (number of nearest neighbours or distance must be specified by the user). Our method based on tessellation has the advantage that the whole zone is covered (all the points have neighbours) and that the method is defined without parameters entered by the user. However, one must keep in mind that neighbourhood relation represents the strength of the potential interaction between two points and so the choice of the neighbourhood matrix can greatly influence the results of the analysis.

Measurements of spatial covariance

A major purpose of spatial statistics is to understand the spatial distribution of the values of an attribute sampled over the whole study region (Bailey & Gatrell 1995). Is the value observed at a particular location correlated to those observed at neighbouring points? To answer this question, spatial covariance and covariograms are well known tools (see Cressie 1991 for example). We consider a single variable **x** measured at n locations $(x_1, \ldots, x_n)$ defining a n by n neighbourhood matrix $\mathbf{G}^*$. The matrix $\mathbf{G}^*$ is symmetric and allows to define a diagonal matrix of neighbouring weights

$$\mathbf{D}^* = diag(p^*_{1+}, \ldots, p^*_{i+}, \ldots, p^*_{n+}) \quad (1)$$

where

$$p^*_{i+} = \sum_{j=1}^{n} \mathbf{G}^*_{ij} / 2m \quad (1a)$$

is the neighbouring weight for the point i and

$$m = \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbf{G}^*_{ij} / 2 \quad (1b)$$

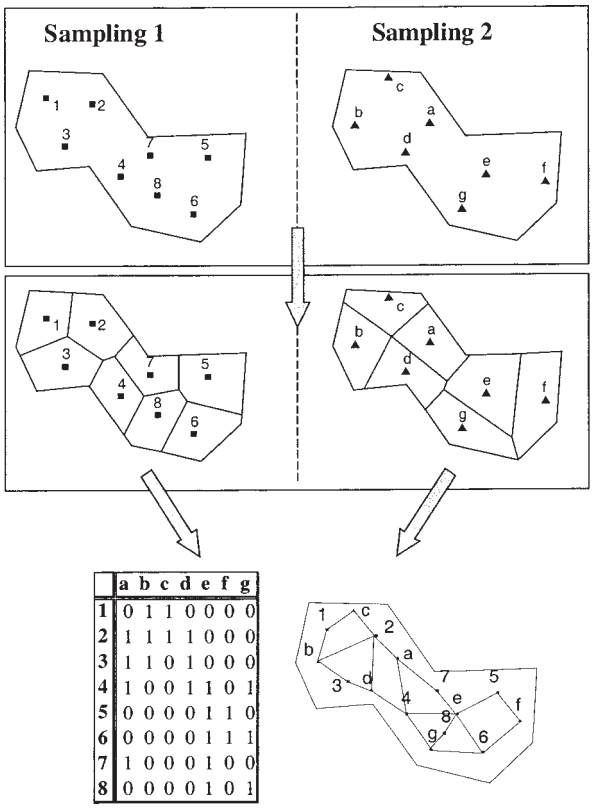
which is the number of pairs of neighbours. For a single variable **x**, the spatial covariance is (Thioulouse et al. 1995):

$$Cov_{spat}(\mathbf{x}) = \frac{1}{2m} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbf{G}^*_{ij} (x_i - \bar{x}_{\mathbf{D}^*})(x_j - \bar{x}_{\mathbf{D}^*}) \quad (2)$$

where

$$\bar{x}_{\mathbf{D}^*} = \sum_{i=1}^{n} p^*_{i+} x_i \quad (2a)$$

which is the mean of the variables **x** given the weights $\mathbf{D}^*$. The word covariance is rather ambiguous here because in the spatial context, it concerns the same variable



	a	b	c	d	e	f	g
1	0	1	1	0	0	0	0
2	1	1	1	1	0	0	0
3	1	1	0	1	0	0	0
4	1	0	0	1	1	0	1
5	0	0	0	0	1	1	0
6	0	0	0	0	1	1	1
7	1	0	0	0	1	0	0
8	0	0	0	0	1	0	1

Fig. 1. Definition of a neighbourhood matrix from two different spatial samplings. From the two data sets, two tessellations are performed. Two voronoi polygons are neighbours if they intersect. Then, a neighbourhood matrix is constructed with 1 if the two individuals are neighbours and 0 otherwise.

and not two variables as in the general statistical context. But we can extend the idea of spatial covariance for a single variable (**x**) to the cross-covariance between two variables (**x**, **y**):

$$Cov_{spat}(\mathbf{x},\mathbf{y}) = \frac{1}{2m}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{G}^{*}_{ij}(x_i - \bar{x}_{\mathbf{D}^*})(y_j - \bar{y}_{\mathbf{D}^*}) \quad (3)$$

The use of cross-covariance has been introduced in kriging methods for spatial interpolation (see Bailey & Gatrell 1995). Suppose that **x** has been recorded at n sites, and additional information on possible covariate **y** is recorded at $n+a$ sites. Then, we can apply co-kriging (based on cross-covariance) by using covariate information to improve the prediction of **x** at a general point in the whole study region. While co-kriging requires that **x** and **y** be measured at the same n locations, the notion of cross-covariance is still legitimate when **x** and **y** have been sampled at different locations.

Spatial-RLQ ordination

Let **R** be an $m_1 \times p$ matrix containing the measurements of p variables at m_1 sites. Let **Q** be an $m_2 \times q$ matrix containing the measurements of q variables at m_2 sites. Spatial-RLQ is based exactly on the same principles as classical RLQ ordination. The only difference concerns the central table **L**. The table **L** derives from a species by sites abundance table in classical RLQ whereas it derives from an $m_1 \times m_2$ neighbourhood matrix in spatial-RLQ analysis. The first step of the method consists of separate ordinations of **R**, **L** and **Q**. The second step is the study of the common structure of **R** and **Q** through **L** with spatial-RLQ analysis.

Correspondence analysis of the central table

Let us consider the $m_1 \times m_2$ matrix **G** where $\mathbf{G}_{ij} = 1$ if sites i and j are neighbours and $\mathbf{G}_{ij} = 0$ otherwise. The table **P** of neighbourhood relative frequencies has m_1 rows and m_2 columns and contains the relative frequencies $\mathbf{P}_{ij} = \mathbf{G}_{ij}/g_{++}$. Moreover, let

$$g_{i+} = \sum_{j=1}^{m_2}\mathbf{G}_{ij}\,,\ g_{+j} = \sum_{i=1}^{m_1}\mathbf{G}_{ij} \text{ and } g_{++} = \sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\mathbf{G}_{ij} \quad (4)$$

be the row totals, the column totals and the grand total of neighbours, respectively. We define $\mathbf{D}_{m_1}$, the diagonal matrix of row neighbouring weights:

$$\mathbf{D}_{m_1} = diag(p_{1+},...,p_{i+},...,p_{m_1+}) \text{ where } p_{i+} = g_{i+} / g_{++}. \quad (5a)$$

In the same way, the diagonal matrix of column neighbouring weights is

$$\mathbf{D}_{m_2} = diag(p_{+1},...,p_{+j},...,p_{+m_2}) \quad (5b)$$

where $p_{+j} = g_{+j} / g_{++}$. Correspondence Analysis (CA) of neighbourhood table **G** is the generalized singular value decomposition (GSVD) of the statistical triplet ($\mathbf{L}$, $\mathbf{D}_{m_2}$, $\mathbf{D}_{m_1}$) with

$$\mathbf{L}_{ij} = \frac{\mathbf{G}_{ij}}{g_{i+}g_{+j}} - 1. \quad (6)$$

This GSVD finds a $\mathbf{D}_{m_2}$-normed vector **c** and a $\mathbf{D}_{m_1}$-normed vector **d** maximizing the quantity (Dolédec et al. 1996)

$$\mathbf{d}^t\mathbf{D}_{m_1}\mathbf{L}\mathbf{D}_{m_2}\mathbf{c} = \frac{1}{g_{++}}\sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\mathbf{G}_{ij}(d_i - \bar{d}_{\mathbf{D}_{m_1}})(c_j - \bar{c}_{\mathbf{D}_{m_2}}) = Cov_{spat}(\mathbf{c},\mathbf{d}) \quad (7)$$

Since **c** and **d** are normed vectors, the above quantity is simply the spatial cross-correlation:

$$Cov_{spat}(\mathbf{c},\mathbf{d}) = Corr_{spat}(\mathbf{c},\mathbf{d})\sqrt{Var(\mathbf{c})}\sqrt{Var(\mathbf{d})} = Corr_{spat}(\mathbf{c},\mathbf{d}) \quad (8)$$

Lebart (1984) has applied correspondence analysis to find row and column scores with maximum spatial correlation for the symmetric matrix **G** of a simple graph. In the case of an asymmetric neighbourhood matrix, correspondence analysis finds row and column scores maximizing the spatial cross-correlation between neighbours. Eigenvectors corresponding to highest eigenvalues describe global structures whereas eigenvectors corresponding to the lowest eigenvalues describe local structures. These eigenvectors can be used as a good alternative to polynomial functions in trend surface analyses to describe spatial patterns (Thioulouse et al. 1995).

Analyses of R and Q

$\mathbf{D}_q$ and $\mathbf{D}_p$ are, respectively, a $q \times q$ and a $p \times p$ diagonal matrix of column weights associated to tables **Q** (m_2 by q) and **R** (m_1 by p). Different ordination methods can be used to analyse the tables **R** and **Q**. Different table transformations imply different analyses (see Dolédec et al. 2000). For example, if the variables in **R** are centred then the GSVD of ($\mathbf{R}$, $\mathbf{I}_p$, $\mathbf{D}_{m_1}$), ($\mathbf{I}_p$ being the $p \times p$ identity matrix) is a centred PCA whereas if the variables in **R** are centred and scaled to unit variance then the same GSVD is a normed PCA. Note that in the GSVD of ($\mathbf{R}$, $\mathbf{I}_p$, $\mathbf{D}_{m_1}$), row weights (p_{i+}) derive from the correspondence analysis of table **L**. Finally, GSVD of ($\mathbf{R}$, $\mathbf{D}_p$, $\mathbf{D}_{m_1}$) and ($\mathbf{Q}$, $\mathbf{D}_q$, $\mathbf{D}_{m_2}$) can result in different types of analyses (normed PCA, centred PCA, CA, or multiple CA…).

Spatial-RLQ analysis

The purpose of spatial-RLQ analysis is to study the common structure of tables **R** and **Q** through the neighbourhood relationship instead of analysing tables **R** and **Q** separately and trying to find a common spatial structure. Spatial-RLQ is defined as the GSVD of

$$\left(\mathbf{R}^t\mathbf{D}_{m_1}\mathbf{L}\mathbf{D}_{m_2}\mathbf{Q}, \mathbf{D}_q, \mathbf{D}_p\right).$$

This analysis consists in finding a $\mathbf{D}_p$-normed axis and a $\mathbf{D}_q$-normed component $\mathbf{v}_1$ so that the quantity: $\mathbf{u}_1{}^t\mathbf{D}_p\mathbf{R}^t\mathbf{D}_{m_1}\mathbf{L}\mathbf{D}_{m_2}\mathbf{Q}\mathbf{D}_q\mathbf{v}_1$ is maximized.

We can rewrite the previous equation with $\mathbf{a} = \mathbf{R}\mathbf{D}_p\mathbf{u}_1$ and $\mathbf{b} = \mathbf{Q}\mathbf{D}_q\mathbf{v}_1$:

$$\mathbf{u}_1{}^t\mathbf{D}_p\mathbf{R}^t\mathbf{D}_{m_1}\mathbf{L}\mathbf{D}_{m_2}\mathbf{Q}\mathbf{D}_q\mathbf{v}_1 = \mathbf{a}^t\mathbf{D}_{m_1}\mathbf{L}\mathbf{D}_{m_2}\mathbf{b}. \quad (9)$$

It can be easily demonstrated that:

$$\mathbf{a}^t\mathbf{D}_{m_1}\mathbf{L}\mathbf{D}_{m_2}\mathbf{b} = Cov_{spat}(\mathbf{a},\mathbf{b}) \quad (10)$$

So, spatial-RLQ finds linear combinations of variables of **R** ($\mathbf{a} = \mathbf{R}\mathbf{D}_p\mathbf{u}_1$) and linear combinations of variables of **Q** ($\mathbf{b} = \mathbf{Q}\mathbf{D}_q\mathbf{v}_1$) that have maximum spatial cross-covariance. The spatial cross-covariance can be decomposed into three terms, like classical covariance:

$$Cov_{spat}(\mathbf{a},\mathbf{b}) = Corr_{spat}(\mathbf{a},\mathbf{b})\sqrt{Var(\mathbf{a})}\sqrt{Var(\mathbf{b})} \quad (11)$$

This decomposition shows that spatial-RLQ is a compromise between the three separate analyses. The first part ($Corr_{spat}$ ($\mathbf{a}$,$\mathbf{b}$)) corresponds to the correspondence analysis of **L**, the second ($\sqrt{Var(\mathbf{a})}$) to the analysis of **R** and the third ($\sqrt{Var(\mathbf{b})}$) to the analysis of **Q**. In RLQ, we maximize a compromise that finds a score induced by the variables of **R** and a score induced by the variables of **Q** which have a maximum spatial cross-correlation. Maximum values are obtained from separate analyses and so we can compare ordination obtained from spatial-RLQ analysis to those obtained from separate analyses. Furthermore, a Monte-Carlo test is available by permuting rows of tables **R** and **Q** in order to test the legitimacy of the spatial-RLQ analysis.

Application

Data were collected from a 2614-ha managed forest located at Chizé (western France, Fig. 2a). The first data set was collected at the subplot scale, which corresponds to the level of forestry management (4 ha on average). These data were available for the two main vegetation strata, the timber stands and the coppices. This data set was collected by foresters of the Office National des Forêts and is used essentially for forest management. For each subplot, foresters determine the three dominant species for the coppices and the four dominant species for the timber stand, and their cover (in %). Subspecies and rare species were not determined and data were pooled resulting in ten categories for timber stand data and five for coppice data (Table 1).

The second data set concerns the same area but it contains information about vegetation accessible to roe deer (height < 1.20 m, Duncan et al. 1998). This data set was collected at the scale of 1-m^2 sample plots. This sample technique is part of a population dynamic study aiming to understand relationships between roe deer population and their available food. For this second data set, taxonomic information is recorded at the genera level. For each quadrat, the presence (or absence) of genera was recorded. In total, 613 subplots (data set 1) and 578 points (data set 2) were recorded (Fig. 2). The first data set results in table **R** with 613 rows and 15 columns and the second one in table **Q** with 578 rows and 58 columns. The purpose of our study is an examination of the relationships between canopy (timber stand and coppice) and understorey (vegetation < 1.20 m height) when the two data sets are measured on different spatial scales.

The data have been geo-referenced and introduced in a Geographic Information System (GIS). Then, a tessellation on data points has been carried out and we have used the above decision rules to construct a neighbourhood matrix **L** with 613 rows and 578 columns. Separates analyses have been performed: PCAs for tables **R** and **Q** and CA for central table **L**. We performed a randomization test to check for the statistical significance of the relationship between **R** and **Q**. This test is based on permutation of rows of tables **R** and **Q** and for each permutation the total inertia of the analysis is computed. The total inertia increases with the intensity of the link between **R** and **Q** through **L**. We used a Monte-Carlo version of the test with 1000 permutations, demonstrating a significant relationship ($p < 0.001$: all permutations have values smaller than observed total inertia) validating the use of spatial-RLQ analysis. The first axis of the spatial-RLQ analysis takes into account 94% (935/990, Table 2) of the total co-structure and we focus on results for this axis only. As seen before, spatial-RLQ analysis maximize the spatial covariance between linear combinations of the variables of **R** and linear combinations of the variables of **Q**. This covariance can be decomposed as the product of two standard deviations by their spatial correlation. Hence, it is possible to measure the proportion of variance attributed to each table and this can be compared to those obtained by separate analyses (Table 2). For table **R**, the first axis of spatial-RLQ analysis takes into account 97% (2988/3075) of the maximal potential inertia obtained by separate analysis and 89% (0.7419/0.8281) for table **Q**. Regarding spatial cross-correlation, the maximum is achieved by CA of table **L** and is equal to the square root of the first eigenvalue of CA ($\sqrt{0.9944} = 0.9972$). The spatial cross-correlation resulting from RLQ analysis shows a decrease (0.6495 in comparison to 0.9972). CA aims to find a couple of normed vectors of maximal cross-correlation whereas RLQ is based on the maximisation of the cross-covariance between linear combinations of R and Q. The observed decrease results from the tables R and Q which do not allow to reconstruct the complete original

Table 1. Taxonomic names and codes.

Name	Code	Name	Code
Data set 1		*Ficaria ranunculoides*	Ficra
Timber stand		*Fragaria* spec.	Fra
Quercus spec.	QueT	*Fraxinus excelsior*	Fraex
Acer spec.	AceT	*Galium* spec.	Gal
Pinus spec.	PinT	*Geum* spec.	Geu
Other deciduous	DecT	*Glechoma* spec.	Gle
Cedrus spec.	CedT	*Hedera helix*	Hedhe
Carpinus betulus	CarT	*Hieracium* spec.	Hie
Prunus avium	PruT	*Hyacinthoides* spec.	Hya
Fagus sylvatica	FagT	*Hypericum* spec.	Hyp
Abies douglasi	AbiT	*Ilex aquifolium*	Ileaq
Other coniferous	ConT	*Lathyrus* spec.	Lat
		Ligustrum vulgare	Ligvu
Coppices		*Lithospermum* spec.	Lit
Quercus spec.	QueC	*Lonicera periclymenum*	Lonpe
Acer spec.	AceC	*Mellitis* spec.	Mel
Other deciduous	DecC	other *Lamiaceae*	Othla
Carpinus betulus	CarC	other *Prunus*	Othpr
Fagus sylvatica	FagC	*Ornithogalum* spec.	Orn
		Poaceae	Poa
Data set 2		*Pinus* spec.	Pin
Vegetation lower than 1.20 m		*Potentilla sterilis*	Potst
Acer spec.	Ace	*Prunus spinosa*	Prusp
Ajuga reptans	Ajure	*Pulmonaria* spec.	Pul
Allium sativum	Allsa	*Quercus* spec.	Que
Anemone nemorosa	Anene	*Ranunculus* spec.	Ran
Arum spec.	Aru	*Rhamnus* spec.	Rha
Calamintha spec.	Cal	*Rosa* spec.	Ros
Carex spec.	Car	*Rubia peregrina*	Rubpe
Carpinus betulus	Carbe	*Rubus* spec.	Rub
Clematis vitalba	Clevi	*Ruscus aculeatus*	Rusac
Convolvulus spec.	Con	*Senecio* spec.	Sen
Cornus spec.	Cor	*Sorbus domestica*	Sordo
Corylus avellana	Corav	*Sorbus torminalis*	Sorto
Crataegus spec.	Cra	*Stachys* spec.	Sta
Epilobium spec.	Epi	*Ulmus* spec.	Ulm
Eupatorium cannabinum	Eupca	*Veronica* spec.	Ver
Euphorbia spec.	Eup	*Viburnum lantana*	Vibla
Euonymus europaeus	Euoeu	*Vicia* spec.	Vic
Fagus sylvatica	Fagsy	*Viola* spec.	Vio

data, i.e. ordination of subplots and points of the separate analysis of L.

Taxonomic information has been projected onto the first axis of spatial-RLQ (Fig. 3). On this axis, genera are plotted according to the link of their spatial distribution with global spatial patterns. It is apparent that stands with *Fagus sylvatica* in the canopy (upper side of the first axis) tend to have *Rubus* spec. and *Ruscus aculeatus* in the understorey and are mostly distributed in the south of the forest. The north of the forest is mainly occupied by stands with oak in the canopy and *Carpinus betulus* and *Ornithogalum* spec. (lower side of the first axis). So, it seems that the first axis indicates a species turnover from the north to the south of the forest involving different species communities in these two parts of the forest. Representation of scores of subplots and sample points by a smoothing by two-dimensional weighted local regression (Cleveland & Devlin 1988) confirms these trends (Fig. 4a, b). Sample scores, which are defined by species composition, are structured from the north to the south. However, there are some differences between these two maps especially in the southwest of the forest where the two scores are quite different. This lack of correspondence between the two scores is probably due to the fact that in this area trees are young and these subplots do not contain mature timber stands.

Table 2. Inertia decomposition for spatial-RLQ analysis (three tables ordination). Inertia: maximal projected variability; Var: variance of the sets of factorial scores computed for the first axis; Cov: covariance between the two sets of factorial scores projected on the first spatial-RLQ axis; Cor: correlation between the two sets of factorial scores projected on the first spatial-RLQ axis.

	Spatial-RLQ analysis	Maximal potential values (obtained by separate analysis)
Total inertia	990	
Inertia projected on F1	935	
Cov (F1-**R**, F1-**Q**)	30.58	
Cor (F1-**R**, F1-**Q**)	0.6495	0.9972 (CA of **L**)
Var (F1-**R**)	2988	3075 (PCA of **R**)
Var (F1-**Q**)	0.7419	0.8281 (PCA of **Q**)

Conclusion

Spatial-RLQ analysis is a new methodological perspective for coupling two data sets. This method is close to co-inertia analysis, which is used for linking two data sets with the same samples. In our method, the two data sets are considered symmetrically but it could be interesting to introduce an asymmetric part in order to explain one data set by the other. As CCA can be considered as an asymmetric co-inertia analysis, double-constrained CCA (Böckenholt & Böckenholt 1990; Lavorel et al. 1998, 1999) could be a good starting point for an

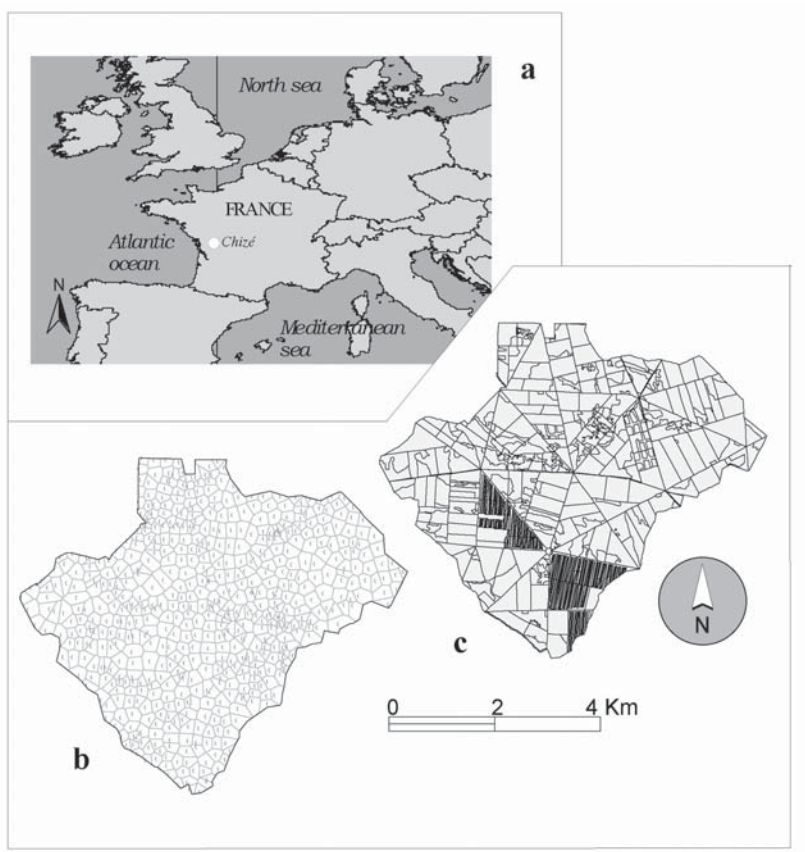


Fig. 2. Spatial location (**a**) and sampling scheme (**b**, **c**) of the Chizé forest (Western France). Information has been collected for points (**b**) and for sub-plots (**c**). From the point data, a tessellation is applied in order to construct a neighbourhood matrix.

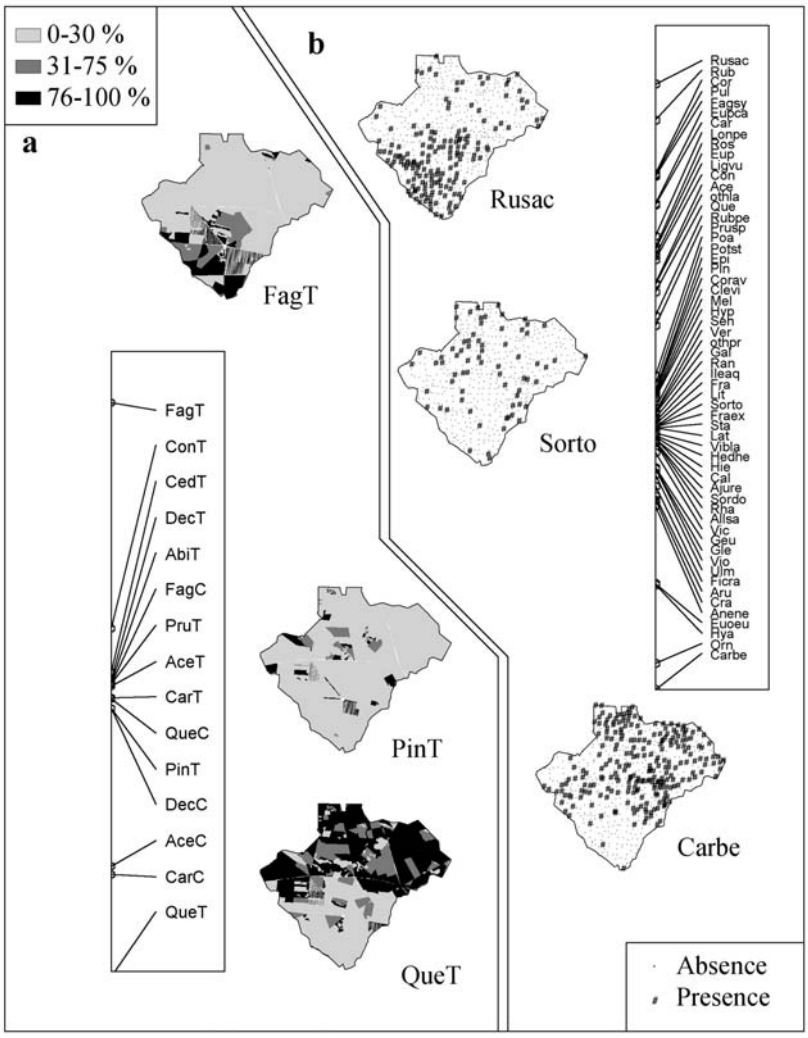


Fig. 3. Spatial-RLQ analysis of the Chizé data. Axis 1 of the species ordination for timber stand and coppice data (**a**) and for understorey vegetation (**b**). Spatial distributions of some species are represented.

asymmetric form of spatial-RLQ analysis. Moreover, we define in this paper a way to construct the neighbourhood matrix. This matrix represents the strength of the potential interactions between locations. The use of GIS permits the definition of a number of procedures to construct spatial neighbourhood matrices (Anselin & Getis 1992). Obviously, our choice is very simple but spatial-RLQ is flexible and can admit different kinds of neighbourhood matrices. The only constraint is due to the CA of the central table implying that all elements of this table must be non-negative. For example, taking into account the area of one sample crossing another one would probably make the analysis more realistic. Since the analytical results may be sensitive to the specification

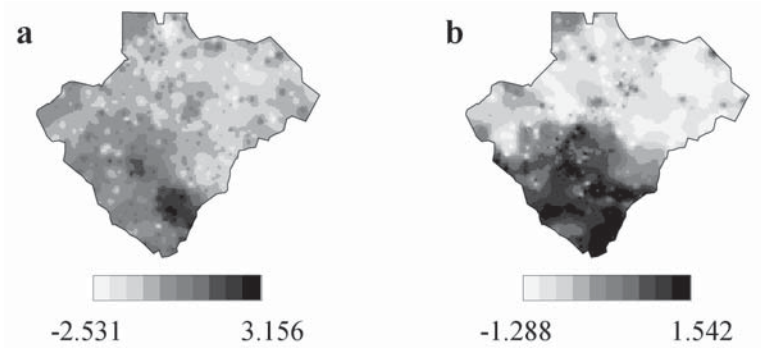


Fig. 4. Spatial-RLQ analysis of Chizé data. Smoothing by two-dimensional weighted local regression of (**a**) point scores of understory vegetation, (**b**) sub-plot scores of timber stand and coppices.

of the neighbourhood matrix, different spatial neighbourhood matrices may be needed for different purposes of studies. There is no universal type of neighbourhood matrix that can be used in spatial analysis. The choice of neighbourhood matrix and multiple possibilities of analyses of marginal tables **R** and **Q** make spatial-RLQ analysis appears as a flexible and a general method for spatial co-ordination of data.

Acknowledgements. We are grateful to the Office National des Forêts and to all field assistants and volunteers that spent time collecting data on the study site. We also wish to thank J. Oksanen, E. van der Maarel and P. Legendre and an anonymous reviewer, whose suggestions and comments have allowed us to improve the first version of this text.

References

Anselin, L. & Getis, A. 1992. Spatial statistical analysis and geographic information systems. *Ann. Reg. Sci.* 26: 19-33.

Bailey, T.C. & Gatrell, A.C. 1995. *Interactive spatial data analysis*. Longman, Harlow, UK.

Böckenholt, U. & Böckenholt, I. 1990. Canonical analysis of a contingency tables with linear constraints. *Psychometrika* 55: 633-639.

Cleveland, W.S. & Devlin, S.J. 1988. Locally weighted regression: an approach to regression analysis by local fitting. *J. Am. Stat. Ass.* 83: 596-610.

Cressie, N. 1991. *Statistics for spatial data*. John Wiley, New-York, NY.

Dolédec, S. & Chessel, D. 1994. Co-inertia analysis: an alternative method for studying species-environment relationships. *Freshwater Biol.* 31: 277-294.

Dolédec, S., Chessel, D., ter Braak, C.J.F. & Champely, S. 1996. Matching species traits to environmental variables: a new three-table ordination method. *Environ. Ecol. Stat.* 3: 143-166.

Dolédec, S., Chessel, D. & Gimaret-Carpentier, C. 2000. Niche separation in community analysis: a new method. *Ecology* 81: 2914-2927.

Duncan, P., Tixier, H., Hofman, R.R. & Lechner-Doll, M. 1998. Feeding strategies and the physiology of digestion in roe deer. In: Andersen R., Duncan, P. & Linnell, J.D.C. (eds.) *The European roe deer: the biology of success*, pp. 91-116. Scandinavian University Press, Oslo, NO.

Green, P. & Sibson, R. 1978. Computing Dirichlet tessellations in the plane. *Comput. J.* 21: 168-173.

Greenacre, M.J. 1984. *Theory and applications of Correspondence Analysis*. Academic Press, London, UK.

Hill, M.O. 1991. Patterns of species distribution in Britain elucidated by canonical correspondence analysis. *J. Biogeogr.* 18: 247-255.

Lavorel, S., Touzard, B., Lebreton, J.D. & Clément, B. 1998. Identifying functional groups for response to disturbance in an abandoned pasture. *Acta Oecol.* 19: 227-240.

Lavorel, S., Rochette, C. & Lebreton, J.D. 1999. Functional groups for response to disturbance in Mediterranean old fields. *Oikos* 84: 480-498.

Lebart, L.1984. Correspondence analysis of a graph structure. *Bull. Tech. CESIA*. 2: 5-19.

Lebreton, J.D., Chessel, D., Prodon, R. & Yoccoz, N. 1988. L'analyse des relations espèces-milieu par l'analyse canonique des correspondances. I. Variables de milieu quantitatives. *Acta Oecol. - Oecol. Gener.* 9: 53-67.

Makarenkov, V. & Legendre P. 2002. Nonlinear redundancy analysis and canonical correspondence analysis based on polynomial regression. *Ecology* 83: 1146-1161.

Mourelle, C. & Ezcurra, E. 1996. Species richness of Argentine cacti: A test of biogeographic hypotheses. *J. Veg. Sci.* 7: 667-680.

Rameau, C., Mansion, D. & Dumé, G. 1989. *Flore forestière française, Plaine et Colline*. Institut pour le développement forestier, Paris, FR.

Rao, C.R. 1964. The use and interpretation of principal component analysis in applied research. *Sankhya A*. 26: 329-359.

Ribera, I., Dolédec, S., Downie, I.S. & Foster, G.N. 2001. Effect of land disturbance and stress on species traits: a three table analysis of ground beetle assemblage. *Ecology* 82: 1112-1129.

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. & Verdonschot, P.F.M. 1995. Canonical correspondence analysis and related multivariate methods in aquatic ecology. *Aquat. Sci.* 57: 255-289.

Thioulouse, J., Chessel, D. & Champely, S. 1995. Multivariate analysis of spatial patterns: a unified approach to local and global structures. *Environ. Ecol. Stat.* 2: 1-14.

Received 25 October 2001;
Revision received 3 July 2002;
Accepted 7 September 2002.
Coordinating Editor: J. Oksanen & E. van der Maarel.