Couplage de données issues de deux plans d'échantillonnage distincts d'un même espace

Stéphane Dray

Laboratoire Biométrie et Biologie Evolutive, UMR CNRS 5558
Université Claude Bernard Lyon I
F-69622 VILLEURBANNE Cedex

dray@biomserv.univ-lyon1.fr

RÉSUMÉ. *Les méthodes permettant d'étudier la co-structure d'un couple de tableaux exigent que les variables des deux jeux de données soient mesurés pour les mêmes sites d'échantillonnage. Nous présentons une nouvelle méthode, l'analyse RLQ spatialisée, permettant de coupler les données provenant de deux plans d'échantillonnage distincts d'un même espace. Cette méthode, extension de l'analyse de co-inertie, est basée sur la théorie des graphes et l'analyse RLQ classique. Elle cherche des combinaisons linéaires des variables de chaque jeu de données maximisant la covariance spatiale croisée et fournit une co-ordination des deux tableaux en tenant compte de la relation spatiale. Une application concernant l'analyse de données d'atlas écologiques est proposée.*

ABSTRACT. *Methods for studying the common structure of a pair of tables require that variables of the two data sets have been recorded at the same sample stations. We present a new method, spatial RLQ analysis, to link data arising from different sample schemes of the same area. This method 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 maximise the spatial cross-covariance. This provides a co-ordination of the two data sets according to their spatial relationships. An application concerning the analysis of ecological atlas data is proposed.*

MOTS-CLÉS : *analyse de co-inertie ; analyse RLQ ; covariance spatiale croisée ; ordination ; relation de voisinage.*

KEYWORDS: *co-inertia analysis, neighbourhood relationship, ordination, RLQ analysis, spatial cross-covariance*

1. Introduction

Les systèmes d'information géographique (SIG) permettent de stocker, gérer, manipuler, analyser des grands jeux de données à référence spatiale. L'analyse de ces données est souvent difficile car les découpages surfaciques peuvent varier

suivant les jeux de données. Ce « décalage spatial » peut apparaître en cas de confidentialité des données, de données provenant de différentes institutions ou de variations des zones d'étude au cours du temps. Ce problème est bien connu et la plupart des SIG possèdent des méthodes d'interpolation surfacique (Goodchild & Lam, 1980) permettant de transférer les données collectées pour un système de surfaces (source) à un autre système de surfaces (cible).

Les méthodes d'analyses multivariées permettent d'obtenir un résumé graphique de grands jeux de données et elles ont été fréquemment utilisées pour identifier des structures spatiales multivariées. Il y a près de cinquante ans, Goodall (1954) cartographiait les scores d'une ACP à l'aide de courbes de niveaux afin de représenter le caractère spatial des structures identifiées. Depuis ces premiers travaux, différentes approches ont été proposées afin de tenir compte de l'espace dans les méthodes d'analyses multivariées. Ainsi, les mesures d'autocorrélation spatiale, comme l'indice de Moran (Cliff & Ord, 1973) ont des extensions multivariées et il est alors possible d'identifier les structures spatiales locales et globales des jeux de données (Thioulouse *et al.*, 1995).

L'utilisation croissante des SIG a introduit de nouvelles problématiques thématiques requérrant des développements méthodologiques. Par exemple, il peut être intéressant de coupler deux jeux de données provenant de systèmes d'échantillonnage distincts d'un même espace. Une réponse méthodologique à ce problème est apportée par l'analyse RLQ spatialisée (Dray *et al.*, 2002). Nous présentons les principales caractéristiques de cette méthode. Une illustration, basée sur l'analyse de données provenant de deux atlas, est proposée concernant l'influence de facteurs environnementaux (mesurés pour des régions éco-géographiques) sur la distribution d'insectes carabiques (mesurée pour des quadrats) dans la région Rhône-Alpes (figure 1).

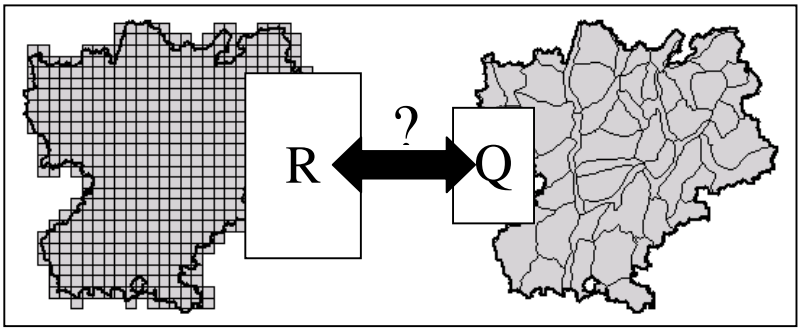

Figure 1. *L'analyse RLQ spatialisée permet d'étudier la co-structure entre deux ensembles de variables provenant de système d'échantillonnage distincts d'une même zone. Dans notre exemple le tableau* ***R*** *contient des mesures d'abondance d'insectes carabiques dans des quadrats et le tableau* ***Q*** *contient des mesures de variables environnementales pour des régions éco-géographiques.*

2. Mesures d'une relation spatiale multivariée

2.1. *Graphe de voisinage*

Considérons la partition d'une région par deux systèmes de m_1 et m_2 zones respectivement. Il est facile, à l'aide d'un SIG, de calculer une matrice de voisinage $\mathbf{G}$ (m_1, m_2) exprimant le lien entre ces deux systèmes d'échantillonnage ($\mathbf{G}_{ij} = 1$ si la zone i du premier système coupe la zone j du second système, $\mathbf{G}_{ij} = 0$ sinon). Deux matrices contenant les pondérations de voisinage pour les lignes et les colonnes de $\mathbf{G}$ sont définies :

$$\mathbf{D}_{m_1} = diag\left(\sum_{j=1}^{m_2}\mathbf{G}_{ij} \Big/ \sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\mathbf{G}_{ij}\right) \text{ et } \mathbf{D}_{m_2} = diag\left(\sum_{i=1}^{m_1}\mathbf{G}_{ij} \Big/ \sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\mathbf{G}_{ij}\right)$$

2.2. *Covariance spatiale*

La covariance spatiale, ou auto-covariance, est bien connue des géostatisticiens. Elle exprime, pour une variable, le lien entre la valeur prise en un point et les valeurs prises aux points voisins. Cette mesure peut être étendue au cas de deux variables $\mathbf{x}$ et $\mathbf{y}$ mesurées respectivement pour m_1 et m_2 sites. La covariance spatiale croisée entre $\mathbf{x}$ et $\mathbf{y}$ s'écrit alors :

$$Cov_{spat} = \sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\mathbf{G}_{ij}\left(x_i - \overline{x}_{\mathbf{D}_{m_1}}\right)\left(y_j - \overline{y}_{\mathbf{D}_{m_2}}\right) \Big/ \sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\mathbf{G}_{ij}$$

Il est remarquable que, dans cette formule, les moyennes des variables $\mathbf{x}$ et $\mathbf{y}$ sont calculées à l'aide des pondérations de voisinage.

2.3. *Analyse factorielle des correspondances d'un graphe de voisinage*

L'analyse factorielle des correspondances (AFC) d'un graphe de voisinage symétrique (Lebart, 1984) peut être étendue au cas d'un graphe non symétrique. Cela correspond à l'analyse du triplet ($\mathbf{L}$, $\mathbf{D}_{m_2}$, $\mathbf{D}_{m_1}$) qui recherche un vecteur $\mathbf{c}$ $\mathbf{D}_{m_2}$-normé et un vecteur $\mathbf{d}$ $\mathbf{D}_{m_1}$-normé maximisant :

$$\mathbf{d}^t\mathbf{D}_{m_1}\mathbf{L}\mathbf{D}_{m_2}\mathbf{c} = \sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\mathbf{G}_{ij}\left(d_i - \overline{d}_{\mathbf{D}_{m_1}}\right)\left(c_j - \overline{c}_{\mathbf{D}_{m_2}}\right) \Big/ \sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\mathbf{G}_{ij} = Cov_{spat}(\mathbf{c},\mathbf{d})$$

$$\text{avec } \mathbf{L}_{ij} = \left(\mathbf{G}_{ij} \Big/ \sum_{i=1}^{m_1}\mathbf{G}_{ij}\sum_{j=1}^{m_2}\mathbf{G}_{ij}\right) - 1$$

Les vecteurs $\mathbf{c}$ et $\mathbf{d}$ sont normés et la quantité précédente est simplement la corrélation spatiale croisée entre $\mathbf{c}$ et $\mathbf{d}$:

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

3. Analyse RLQ spatialisée

Les données collectés pour chaque système d'échantillonnage induisent deux analyses séparées associées aux triplets statistiques ($\mathbf{R}$, $\mathbf{D}_p$, $\mathbf{D}_{m_1}$) et ($\mathbf{Q}$, $\mathbf{D}_q$, $\mathbf{D}_{m_2}$). Dans ces deux triplets, les pondérations des lignes (sites) proviennent de l'AFC du graphe de voisinage. L'analyse RLQ spatialisée permet d'étudier la structure commune des tableaux **R** et **Q** au travers du lien exprimé par la relation de voisinage spatiale. Elle se définit par l'analyse du triplet $\left(\mathbf{R}^t\mathbf{D}_{m1}\mathbf{L}\mathbf{D}_{m2}\mathbf{Q}, \mathbf{D}_q, \mathbf{D}_p\right)$ qui recherche **u**, axe principal $\mathbf{D}_p$-normé et **v,** composante principale $\mathbf{D}_q$-normé maximisant la quantité :

$$\mathbf{u}^t\mathbf{D}_p\mathbf{R}^t\mathbf{D}_{m_1}\mathbf{L}\mathbf{D}_{m_2}\mathbf{Q}\mathbf{D}_q\mathbf{v}$$

Une autre formulation de l'équation précédente est obtenue en posant $\mathbf{a} = \mathbf{R}\mathbf{D}_p\mathbf{u}$ et $\mathbf{b} = \mathbf{Q}\mathbf{D}_q\mathbf{v}$:

$$\mathbf{u}^t\mathbf{D}_p\mathbf{R}^t\mathbf{D}_{m_1}\mathbf{L}\mathbf{D}_{m_2}\mathbf{Q}\mathbf{D}_q\mathbf{v} = \mathbf{a}^t\mathbf{D}_{m_1}\mathbf{L}\mathbf{D}_{m_2}\mathbf{b} = Cov_{spat}(\mathbf{a},\mathbf{b})$$

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

L'analyse RLQ spatialisée recherche donc un vecteur **a**, combinaison linéaire des variables de **R**, et un vecteur **b**, combinaison linéaire des variables de **Q**, de covariance spatiale croisée maximale. Cette analyse est un compromis entre les trois analyses séparées des tableaux **L**, **R** et **Q**. La première partie ($Corr_{spat}(\mathbf{a},\mathbf{b})$) correspond à l'AFC de **L**, la seconde ($\sqrt{Var(\mathbf{a})}$) à l'analyse de **R** et la troisième ($\sqrt{Var(\mathbf{b})}$) à l'analyse de **Q**.

Remerciements au Muséum d'Histoire naturelle de Lyon, au R.E.R.A et à la Région Rhône-Alpes pour la mise à disposition des données carabiques.

A. D. Cliff , J. K. Ord, *Spatial autocorrelation*, London, Pion, 1973

S. Dray, N. Pettorelli , D. Chessel, "Matching data sets from two different spatial samplings", *Journal of Vegetation Science*, vol. 13, 2002, p. 867-874.

D. W. Goodall, "Objective methods for the classification of vegetation III. An essay on the use of factor analysis", *Australian Journal of Botany*, vol. 2, 1954, p. 304-324.

M. Goodchild , N. S. Lam, "Areal interpolation: a variant of the traditional spatial problem", *Geoprocessing*, vol. 1, 1980, p. 297-312.

L. Lebart, "Correspondence analysis of a graph structure", *Bulletin Technique du CESIA*, vol. 2, 1984, p. 5-19.

J. Thioulouse, D. Chessel , S. Champely, "Multivariate analysis of spatial patterns: a unified approach to local and global structures", *Environmental and Ecological Statistics*, vol. 2, 1995, p. 1-14.