Cartographie et analyse de données multivariées

Stéphane DRAY
Laboratoire de Biométrie et Biologie Evolutive
Université Claude Bernard Lyon I, Bat G. Mendel
69622 Villeurbanne Cedex, FRANCE
email : dray@biomserv.univ-lyon1.fr

Résumé : L'identification de structures spatiales induisant plusieurs variables est une problématique récurrente dans de nombreux domaines d'application. Cet article présente quelques aspects permettant la prise en compte de l'espace lors de l'analyse de données multivariées. Une illustration basée sur des données socio-économiques est proposée.

Mots clés : analyse en composantes principales, analyse de données, structure spatiale multivariée

1. Introduction

Selon Poidevin [21], près de 85 % des bases de données contiennent un composant géographique associé à un lieu précis. Cette masse « astronomique » de données spatialisées a entraîné un fort développement de l'utilisation des systèmes d'information géographique (SIG) dans de nombreux domaines d'activités tels que l'environnement, l'agriculture, le marketing, les transports ou la gestion du territoire. L'efficacité des SIG pour stocker, gérer, manipuler, analyser, modéliser, représenter des données à références spatiales n'est plus à démontrer.

Parmi les fonctionnalités de base d'un SIG, la cartographie thématique est un outil d'analyse, d'aide à la décision et de communication largement utilisée pour représenter quelques variables : *une bonne carte vaut mieux qu'un long discours*. Lorsque l'on s'intéresse à des phénomènes impliquant plus de deux variables, il est difficile de représenter l'ensemble de l'information. La rédaction d'une seule carte conduit à des objets surchargés dont la lisibilité est réduite. La rédaction de plusieurs cartes rend difficile l'identification de corrélations et de structures communes (figure 1). Le besoin d'outils permettant d'obtenir des variables synthétiques résumant l'information disponible se fait alors ressentir. C'est précisément l'objectif de *l'analyse de données* et des méthodes *d'analyse multivariée*.

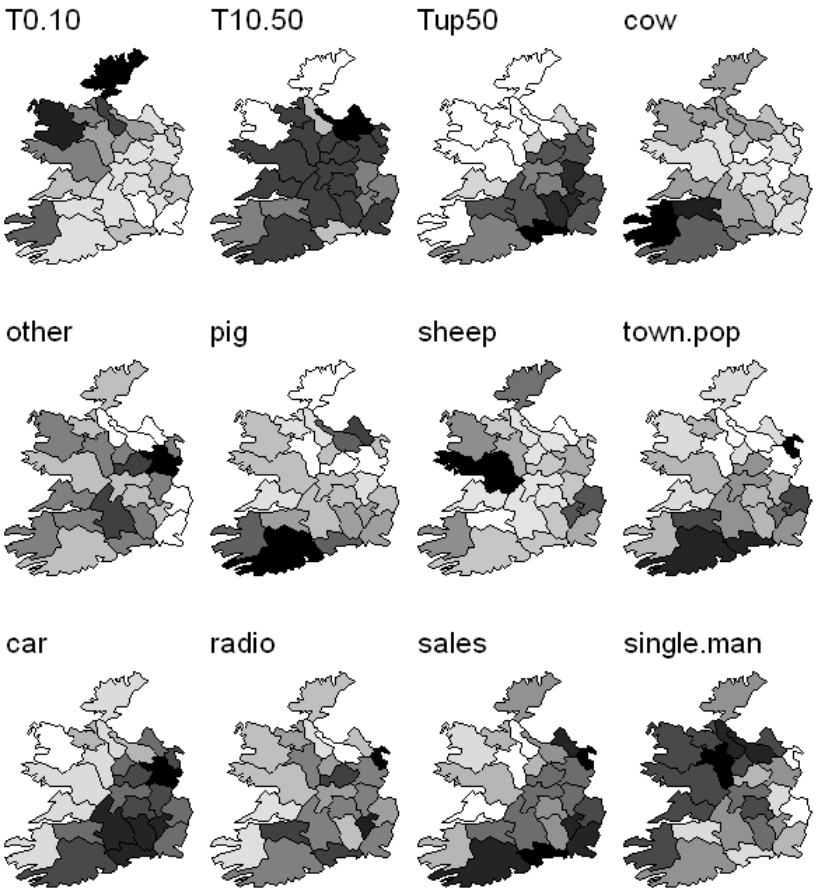

Figure 1. Tableau de 12 variables mesurées sur les 24 districts d'Irlande (le district de Dublin est extrait du jeu de données). Données reprises dans [5, p. 53]. Code des variables : 1-2-3 répartition (en 1 pour 1000) des propriétés agricoles en 3 groupes d'imposition (**T0.10** < 10 £, **T10.50** 10-50 £, **Tup50** > 50 £). 4-5-6-7 Nombres moyens d'animaux pour 1000 acres de prairies et cultures respectivement 4- **cow** vaches laitières, 5- **other** autres bestiaux, 6- **pig** cochons, 7- **sheep** moutons. 8- **town.pop** Pourcentage de population urbanisée (villes et villages) en 1 pour 1000 9- **car** Nombre de voitures pour 1000 habitants 10- **radio** Nombre de licences de radio pour 1000 habitants 11- **sales** Ventes de détail moyenne par habitant en £ 12- **single.man** Pourcentage de célibataires parmi les hommes de 30-34 ans en 1 pour 1000. Données normalisées.

2. Analyse de données multivariées

1. Objectif

« L'analyse de données » recouvre un ensemble de techniques ayant pour objectif la description statistique des grands tableaux. Ces méthodes permettent de rechercher les structures cachées dans les données et *« d'obtenir une description de nature statistique pour un certain phénomène qui a donné lieu au recueil de mesures ou observations trop nombreuses et dépendantes les unes des autres pour être interprétables en première lecture»* [18]. De par la quantité d'information analysée, ces méthodes nécessitent un usage quasi-impératif de l'informatique. Elles se caractérisent par leur objectif exploratoire et par l'abandon d'hypothèses probabilistes (contrairement à la statistique inférentielle) au profit de la géométrie euclidienne.
En général, un tableau contient les mesures de p variables (colonnes) sur n individus (lignes). Dans le cas qui nous intéresse, ces individus correspondent à des unités géographiques. L'analyse exploratoire d'un tel tableau en fournit un résumé en

insistant sur les représentations graphiques des individus qui sont considérés au même titre que les variables. Les domaines d'utilisation de l'analyse de données sont nombreux et diversifiés : biométrie, psychométrie, économétrie…
Il existe une multitude de méthodes d'analyses multivariées permettant de traiter différentes structures de données : l'analyse factorielle des correspondances [15] pour les tables de contingence, l'analyse factorielle multiple [24] pour les variables qualitatives, l'analyse discriminante [11] pour la prise en compte d'une partition des individus en groupe, les méthodes de couplage [22, 16, 8, 25] pour l'analyse d'une paire de tableaux…
Dans cet article, nous présenterons uniquement l'analyse en composantes principales (ACP) qui permet de traiter un tableau de variables quantitatives.

2. Analyse en composantes principales

Un tableau **X** contient les mesures de p variables quantitatives (colonnes) mesurées pour n localisations géographiques (lignes). On associe deux métriques **Q** et **D**, respectivement, aux colonnes et des lignes du tableau **X**. La pondération uniforme est la plus utilisée et conduit aux matrices diagonales $\mathbf{Q} = diag(1,...,1)$ et $\mathbf{D} = diag(\frac{1}{n},...,\frac{1}{n})$. Il est alors possible de représenter les n individus comme des points dans un espace à p dimensions où chaque axe représente une variable. Un deuxième point de vue consiste à représenter les p variables comme des points dans un espace à n dimensions où chaque axe représente un individu. L'objectif de l'ACP est d'obtenir des vecteurs (axes) définissant un sous-espace de dimension réduite permettant une représentation des individus et des variables.

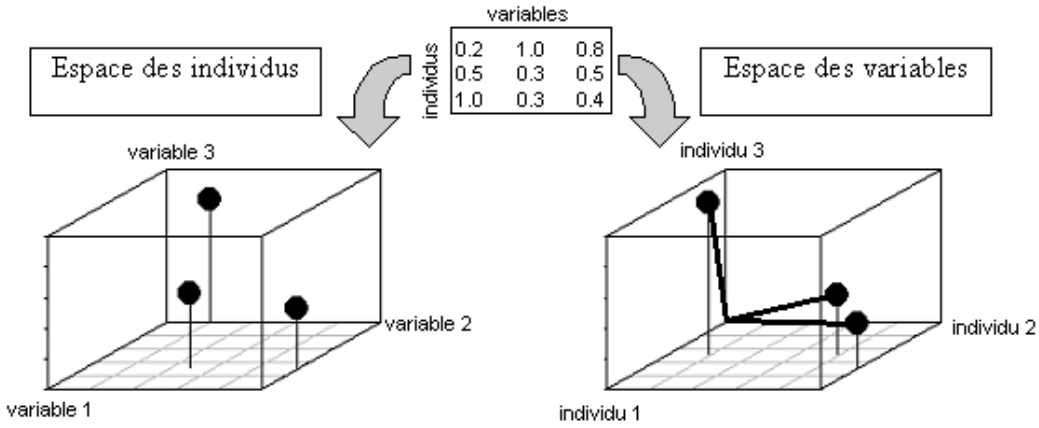

Figure 2. Représentation géométrique d'un tableau comme un nuage de points dans l'espace des individus ou dans l'espace des variables.

L'ACP est l'analyse du triplet (**X**, **Q**, **D**) et consiste à déterminer les valeurs propres et vecteurs propres associés aux différentes possibilités de produits matriciels $\mathbf{X}^t\mathbf{DXQ}$, $\mathbf{DXQX}^t$, $\mathbf{XQX}^t\mathbf{D}$, et $\mathbf{QX}^t\mathbf{DX}$. Dans la pratique, on ne réalisera qu'une seule diagonalisation (dans la dimension la plus petite) et l'on déduira les trois autres systèmes de vecteurs propres. Par exemple, si $p < n$ alors on diagonalisera la matrice $\mathbf{X}^t\mathbf{DXQ}$ afin d'obtenir les vecteurs propres contenus dans la matrice $\mathbf{A}_r$ associés aux r valeurs propres non nulles et distinctes (avec $r \leq p$):

$$\mathbf{X}^t\mathbf{DXQA}_r = \mathbf{A}_r\Lambda_r \text{ et } \mathbf{A}_r^t\mathbf{QA}_r = \mathbf{I}_r$$

La matrice $\Lambda_r = diag(\lambda_1, \lambda_2, ..., \lambda_r)$ contient les r valeurs propres non nulles et la matrice $\mathbf{A}_r$ (p x r) contient les r vecteurs propres **Q**-normés associés. Les matrices $\mathbf{QX}^t\mathbf{DX}$, $\mathbf{DXQX}^t$ et $\mathbf{XQX}^t\mathbf{D}$ ont les mêmes valeurs propres. On déduit les trois autres systèmes de vecteurs propres $\mathbf{B}_r$, $\mathbf{A}_r^*$, $\mathbf{B}_r^*$ associés, respectivement, aux produits matriciels $\mathbf{XQX}^t\mathbf{D}$, $\mathbf{QX}^t\mathbf{DX}$, $\mathbf{DXQX}^t$ grâce aux formules suivantes :

$$\mathbf{A}_r^* = \mathbf{QA}_r \qquad \mathbf{B}_r = \mathbf{XA}_r^*\Lambda_r^{-\frac{1}{2}} \qquad \mathbf{B}_r^* = \mathbf{DB}_r \qquad \mathbf{A}_r = \mathbf{X}^t\mathbf{B}_r^*\Lambda_r^{-\frac{1}{2}}$$

Les vecteurs propres de $\mathbf{X}^t\mathbf{DXQ}$ contenus dans la matrice $\mathbf{A}_r$ sont les r axes principaux. Les vecteurs propres de $\mathbf{XQX}^t\mathbf{D}$ contenus dans la matrice $\mathbf{B}_r$ sont les r composantes principales. Les vecteurs propres de $\mathbf{QX}^t\mathbf{DX}$ contenus dans la matrice $\mathbf{A}_r^*$ sont les r facteurs principaux. Les vecteurs propres de $\mathbf{DXQX}^t$ contenus dans la matrice $\mathbf{B}_r^*$ sont les r cofacteurs principaux.
Les représentations graphiques des individus et des variables sont alors obtenues par simple projection de **X** sur les composantes principales et axes principaux. Les coordonnées des lignes ($\mathbf{L}_r$) et les coordonnées des colonnes ($\mathbf{K}_r$) sont calculées comme suit :

$$\mathbf{K}_r = \mathbf{X}^t\mathbf{DB}_r = \mathbf{A}_r\Lambda_r^{\frac{1}{2}}$$
$$\mathbf{L}_r = \mathbf{XQA}_r$$

Plusieurs critères sont optimisés par l'analyse d'un triplet $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$. Les vecteurs $\mathbf{a}_1, ..., \mathbf{a}_j, ..., \mathbf{a}_r$, contenus dans la matrice $\mathbf{A}_r$, maximisent successivement sous contrainte d'orthonormalité au sens de **Q** la forme quadratique $\|\mathbf{XQa}_i\|_{\mathbf{D}}^2$. Le maximum atteint est λ_i. Les vecteurs $\mathbf{b}_1, ..., \mathbf{b}_j, ..., \mathbf{b}_r$, contenus dans la matrice $\mathbf{B}_r$, maximisent successivement sous contrainte d'orthonormalité au sens de **D** la forme quadratique $\|\mathbf{X}^t\mathbf{Db}_i\|_{\mathbf{Q}}^2$. Le maximum atteint est λ_i. Les couples de vecteurs $(\mathbf{a}_1, \mathbf{b}_1), ..., (\mathbf{a}_j, \mathbf{b}_j), ..., (\mathbf{a}_r, \mathbf{b}_r)$ maximisent successivement sous contrainte d'orthonormalité les produits scalaires $\langle \mathbf{XQa}_i | \mathbf{b}_i \rangle_{\mathbf{D}} = \langle \mathbf{X}^t\mathbf{Db}_i | \mathbf{a}_i \rangle_{\mathbf{Q}}$. Le maximum atteint est $\sqrt{\lambda_i}$.
Dans le cas où le tableau **X** contient des variables mesurées dans différentes unités, il est indispensable de centrer (moyenne nulle) et de normaliser (variance unitaire) les variables avant la recherche des éléments propres (ACP centrée-normée). Cette étape permet d'éviter que l'influence des variables ayant une forte variabilité (due à l'unité de mesure) ne soit sur-évaluée. Pour l'ACP centrée-normée, la maximisation de $\|\mathbf{X}^t\mathbf{Db}_i\|_{\mathbf{Q}}^2$ conduit à rechercher un axe $\mathbf{b}_i$ maximisant la somme des carrés de corrélations avec les variables de **X** ($\sum_{j=1}^{p} corr^2(\mathbf{b}, \mathbf{X}_j)$) et la maximisation de $\|\mathbf{XQa}_i\|_{\mathbf{D}}^2$ recherche une combinaison linéaire des variables de **X** de variance maximale ($\mathrm{var}(\mathbf{Xa}_i)$)

3. Application

Les données concernant les districts d'Irlande ont été traitées à l'aide d'une ACP centrée-normée. Le graphique des valeurs propres (figure 3) montre que la structure sur le premier axe contient beaucoup d'information et qu'il faut également s'intéresser aux trois axes suivants.

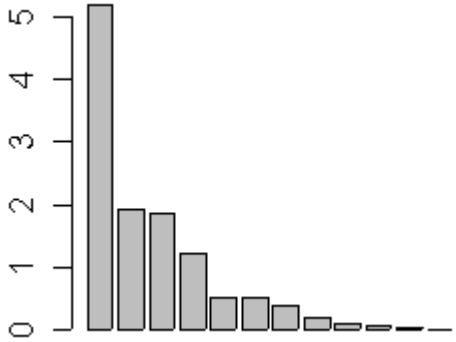

Figure 3. Valeurs propres de l'ACP centrée-normée

Les districts et les variables ont été représentés pour les deux premiers plans factoriels formés par les axes 1, 2 et 3, 4 de l'ACP centrée normée (figure 4). Les flèches quantifient la corrélation des variables avec les axes. La carte des variables peut être directement mise en relation avec la représentation des individus. Il est d'ailleurs possible de superposer variables et individus sur le même graphique (biplot [12]).

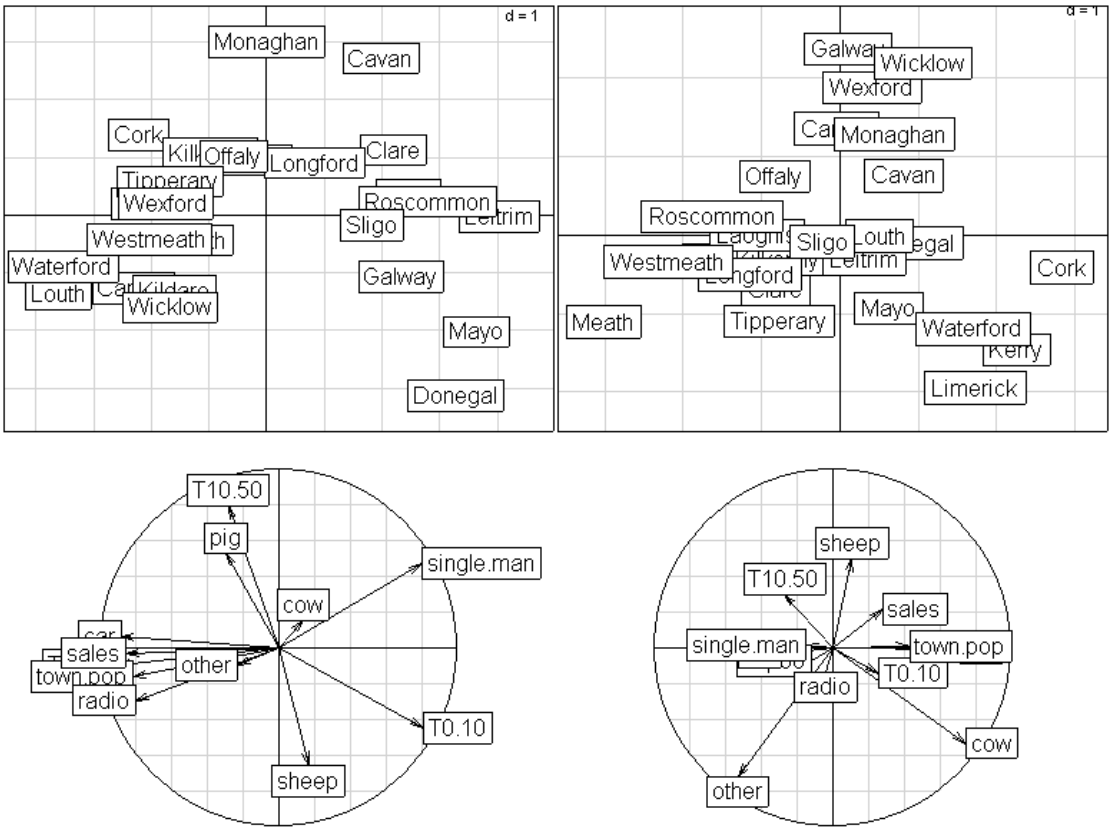

Figure 4. Scores des individus et des variables pour les quatre premiers axes de l'ACP centrée-normée

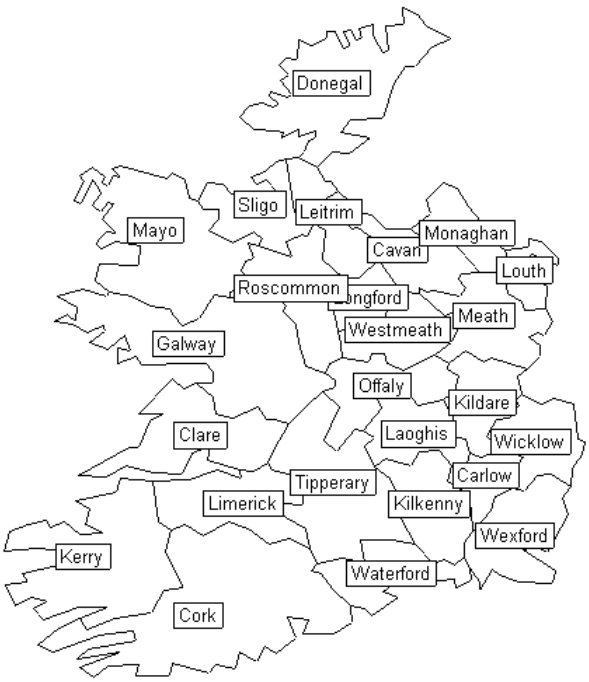

Figure 5. Noms des districts d'Irlande

Le premier axe sépare, à droite, les districts ayant un grand pourcentage de célibataires et un grand nombre d'exploitation faiblement imposé (e.g. Leitrim, Mayo, Donegal) aux districts fortement imposés et urbanisés faisant beaucoup de vente et possédant un grand nombre de voitures et de radios (e.g. Waterford, Louth) situés à gauche de l'axe. L'interprétation peut être répétée pour les autres axes et il est ainsi possible d'identifier des districts ayant des similarités d'un point de vue socio-économique et de connaître la nature de ces similarités.

4. Cartographie

Les représentations graphiques obtenues à l'aide de l'ACP permettent de visualiser les principales structures mais il est difficile d'appréhender le caractère spatial des structures identifiées car il est nécessaire de faire le lien entre la carte géographique (figure 5) et les résultats de l'ACP (figure 4). Une solution bien plus efficace a été utilisée dès les premières applications de l'ACP [14]. Elle consiste à cartographier les scores de l'analyse dans l'espace géographique (figure 6). Cette représentation fait clairement apparaître que la structure identifiée sur le premier axe de l'analyse oppose les districts ruraux et pauvres du Nord-Ouest aux districts plus riches et urbanisés du Sud-Est. Le troisième axe est aussi structuré spatialement, le quatrième beaucoup moins, le second à peine.

Les mesures d'autocorrélation [5] peuvent être appliquées aux scores des districts, obtenus avec l'ACP, afin de tester le caractère spatial des structures identifiées. Les tests confirment les impressions visuelles (tableau 1).

	Axe 1	Axe 2	Axe 3	Axe 4
I (Moran)	0.567	0.1754	0.4677	0.2716
Probabilité	0.0002	0.0704	0.0002	0.0186

Table 1. Test d'autocorrélation spatiale (Moran) pour les quatre premiers axes de l'ACP centrée-normée.

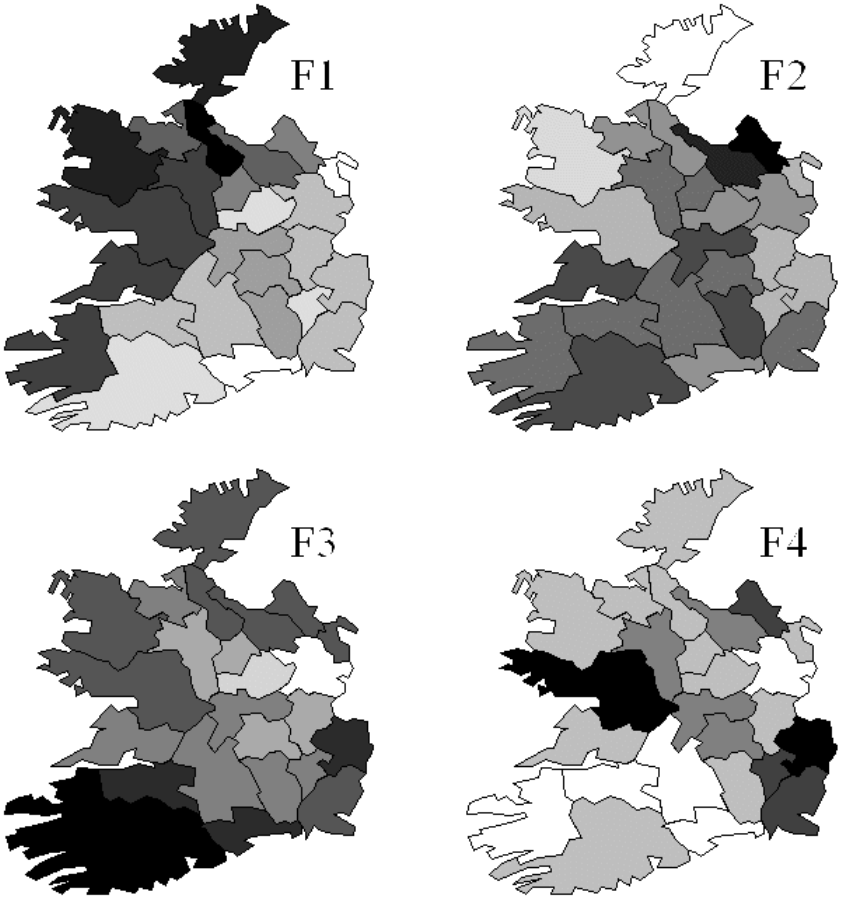

Figure 6. Cartographie des scores des 24 districts pour les quatre premiers axes de l'ACP centrée-normée

3. Vers une prise en compte de l'information spatiale

Les données d'origine dans l'article fondateur de Geary sont curieusement multivariées. Si on sait faire l'analyse du tableau, ici une analyse en composantes principales normée, la question est de reproduire cette analyse en l'optimisant du point de vue de l'intégration de l'espace sous-jacent au découpage administratif. L'exemple précédent montre bien que les structures identifiées par l'ACP peuvent être spatialisées (axes 1 et 3) ou non (axes 2 et 4). De nouvelles méthodes d'analyse de données intégrant une prise en compte de l'espace ont donc été développées afin d'identifier des structures multivariées et spatialisées.

Ces nouvelles méthodes sont basées sur des principes de géostatistique et d'analyse de données. Le couplage de polynômes des coordonnées spatiales des sites et des variables quantitatives par une analyse canonique des corrélations (*canonical trend surface analysis*, [13, 30]) ou par une analyse canonique des correspondances [3, 2,

19, 26] permet d'identifier une structure spatiale multivariée. Cette approche est assez efficace et simple à mettre en oeuvre. Cependant, elle requiert que la zone d'étude soit globalement homogène et que les sites d'échantillonnage soient régulièrement distribués dans l'espace. Dans des situations plus complexes (barrières géographiques, réseau …), l'utilisation de graphes de voisinage pour définir l'espace est plus approprié. De nombreuses méthodes intégrant la structure de voisinage ont donc été développées. Certaines de ces méthodes sont basées sur la variabilité locale [17, 1, 4], d'autres approches sont basées sur l'indice de Moran [31] ou sur la covariance spatiale [10]. L'approche la plus élégante, d'un point de vue mathématique, réconcilie les deux points de vue de Moran et Geary et permet d'identifier les structures spatiales locales et globales [27]. L'analyse multi-échelles de données multivariées a également entraîné le développement de nombreuses méthodes [6, 7, 20, 23, 29, 28].
L'utilisation de ces méthodes est assez marginale et les applications demeurent peu nombreuses. De nombreuses approches ont été proposées mais aucune méthode ne s'est réellement imposée. La problématique méthodologique reste donc toujours ouverte et le développement d'une approche fondée sur des concepts mathématiques clairement définis et efficace d'un point de vue pratique reste à faire.
En tout état de cause, il est évident que la mise en place d'interfaces entre analyses multivariées, systèmes d'information géographique et géostatistiques permet d'améliorer et de faciliter l'analyse de données multivariées et spatialisées [9]. Cette combinaison offre un cadre très performant pour l'analyse de données spatialisées en raison de la complémentarité de ces différentes approches. Les analyses multivariées permettent d'identifier des structures, les géostatistiques permettent d'appréhender le caractère spatial de ces structures alors que le SIG est un outil de stockage, manipulation, création, et représentation de données spatialisées.

Références

[1] T.A. Banet and L. Lebart. Local and partial principal component analysis (pca) and correspondence analysis (ca). In International Association for Statistical Computing., editor, *COMPSTAT 84*, pages 113–123. Physica-Verlag, Vienna, 1984.
[2] D. Borcard and P. Legendre. Environmental control and spatial structure in ecological communities : an example using oribatid mites (acari, oribatei). *Environmental and Ecological Statistics*, 1 :37–61, 1994.
[3] D. Borcard, P. Legendre, and P. Drapeau. Partialling out the spatial component of ecological variation. *Ecology*, 73 :1045–1055, 1992.
[4] D. Chessel and R. Sabatier. Couplage de triplets statistiques et graphes de voisinage. In B. Asselain, M. Boniface, C. Duby, C. Lopez, J.P. Masson, and J. Tranchefort, editors, *Biométrie et Données spatio-temporelles*, pages 28–37. Société Fran¸caise de Biométrie, ENSAR, Rennes, 1993.
[5] A.D. Cliff. and J.K. Ord. *Spatial autocorrelation*. Pion, London, 1973.
[6] M.R.T. Dale and M.W. Zbigniewicz. The evaluation of multi-species pattern. *Journal of Vegetation Science*, 6 :391–398, 1995.
[7] G. Di Bella and G. Jona-Lasinio. Including spatial contiguity information in the analysis of multispeci.c patterns. *Environmental and Ecological Statistics*, 3 :269–280, 1996.
[8] S. Dolédec and D. Chessel. Co-inertia analysis : an alternative method for studying species-environment relationships. *Freshwater Biology*, 31 :277–294, 1994.

[9] S. Dray. *Eléments d'interface entre analyses multivariées, systèmes d'information géographique et observations écologiques*. Thèse de doctorat, Université Lyon I, 2003.
[10] S. Dray, N. Pettorelli, and D. Chessel. Matching data sets from two di.erent spatial samplings. *Journal of Vegetation Science*, 13 :867–874, 2002.
[11] R.A. Fisher. The precision of discriminant functions. *Annals of Eugenics*, 10 :422–429, 1940.
[12] K.R. Gabriel. The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, 58(3) :453–467, 1971.
[13] R. Gittins. *Canonical analysis, a review with applications in ecology*. Springer Verlag, Berlin, 1985.
[14] D.W. Goodall. Objective methods for the classi.cation of vegetation iii. an essay on the use of factor analysis. *Australian Journal of Botany*, 2 :304–324, 1954.
[15] M.J. Greenacre. *Theory and Applications of Correspondence Analysis*. Academic Press, London, 1984.
[16] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28 :321–377, 1936.
[17] L. Lebart. Analyse statistique de la contigu¨ýté. *Publication de l'Institut de Statistiques de l'Université de Paris*, 28 :81–112, 1969.
[18] L. Lebart, A. Morineau, and N. Tabart. *Techniques de la description statistique, méthodes et logiciels pour la description des grands tableaux*. Dunod, Paris, 1977.
[19] A. Méot, P. Legendre, and D. Borcard. Partialling out the spatial component of ecological variation : questions and propositions in the linear modelling framework. *Environmental and Ecological Statistics*, 5 :1–27, 1998.
[20] I. Noy-Meir and D.J. Anderson. Multivariate pattern analysis, or multiscale ordination : towards a vegetation hologram ? In G.P. Patil, E.C. Pielou, and W.E. Waters, editors, *Statistical Ecology III. Populations, ecosystems, and systems analysis*, pages 207–232. Pennsylvania State University Press, 1971.
[21] D. Poidevin. *La carte moyen d'action : conception - réalisation*. Ellipse, Paris, 1999.
[22] C.R. Rao. The use and interpretation of principal component analysis in applied research. *Sankhya A*, 26 :329–359, 1964.
[23] J.A. Schaefer and F. Messier. A paired-quadrat method for use in multiscale ordination. *Vegetatio*, 113 :9–11, 1994.
[24] M. Tenenhaus and F.W. Young. An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical multivariate data. *Psychometrika*, 50(1) :91–119, 1985.
[25] C.J.F. ter Braak. Canonical correspondence analysis : a new eigenvector technique for multivariate direct gradient analysis. *Ecology*, 67 :1167–1179, 1986.
[26] C.J.F. ter Braak. The analysis of vegetation-environment relationships by canonical correspondence analysis. *Vegetatio*, 69 :69–77, 1987.
[27] J. Thioulouse, D. Chessel, and S. Champely. Multivariate analysis of spatial patterns : a unified approach to local and global structures. *Environmental and Ecological Statistics*, 2 :1–14, 1995.
[28] J.M. Ver Hoef and D.C. Glenn-Lewin. Multiscale ordination : a method for detecting pattern at several scale. *Vegetatio*, 82 :59–67, 1989.
[29] J.M. Ver Hoef, D.C. Glenn-Lewin, and M.J.A. Werger. Relationship between pattern and vertical structure in a chalk grassland. *Vegetatio*, 83 :147–155, 1989.

[30] D. Wartenberg. Canonical trend surface analysis : a method for describing geographic pattern. *Systematic Zoology*, 34(3) :259–279, 1985.
[31] D. Wartenberg. Multivariate spatial correlation : a method for exploratory geographical analysis. *Geographical Analysis*, 17(4) :263–283, 1985.