

N° d'ordre : 293-2004

Année : 2004-2005

THESE

présentée

devant l'UNIVERSITE CLAUDE BERNARD - LYON 1

pour l'obtention du DIPLOME DE DOCTORAT

(arrêté du 25 avril 2002)

et soutenue publiquement le

20 décembre 2004

par

OLLIER Sébastien

.....
**Des outils pour l'intégration des contraintes spatiales,
temporelles et évolutives en analyse des données écologiques**

Tome 1
.....

Spécialité : biostatistique

JURY : Dominique Pontier, Présidente
Nigel Yoccoz, Rapporteur
Claude Millier, Rapporteur
Jean Thioulouse, Directeur
Pierre Couteron, Co-directeur

DISCIPLINE : biostatistique

RESUME en français :

Cette thèse propose des outils et concepts nouveaux pour l'intégration des contraintes spatiales, temporelles et évolutives en analyse des données écologiques.

On revient sur la question théorique de l'ordination sous contraintes spatiales par une revue des objets permettant l'intégration des proximités spatiales. On introduit ensuite une nouvelle procédure qui généralise, à l'interface des programmations 'spdep' et 'ade4' du logiciel R, l'ACP sous contrainte de Wartenberg. On aborde ensuite le problème de la typologie de structures multiéchelles, ce qui nous amène à préciser la définition des méthodes d'étude de la structure d'une variable à différentes échelles. On propose une solution à la normalisation des échelles. Les illustrations portent sur des données d'altimétrie laser. Enfin, à partir d'une critique des procédures *ad hoc* rencontrées dans la littérature, on définit des procédures canoniques permettant la prise en compte des proximités évolutives en analyse des données. Les bases orthonormées associées aux matrices de proximité phylogénétiques et leur usage en analyse de données sont introduits.

La conclusion porte sur la pratique de la biométrie et les relations qui s'établissent entre donnée expérimentale, langage mathématique et mise en oeuvre informatique.

MOTS-CLES en français : analyse multivariée, analyse multiéchelle, analyse comparative autocorrélation, écologie statistique, logiciel R

TITRE en anglais : Some tools for the integration of spatial, temporal and evolutive dependence in ecological data analysis

RESUME en anglais

We present new tools and concepts for taking spatial, temporal and evolutive dependence into account in ecological data analysis.

We go back over the problem of multivariate analysis of spatial patterns by examining statistical tools permitting the integration of space in data analysis. We then introduce a new statistical method to generalise, at the interface of the 'ade4' and 'spdep' packages of the R software, the multivariate spatial correlation analysis of Wartenberg. The second part deals with typology of multiscale patterns. Methods for multiscale pattern analysis are presented in the same theoretical context, which leads to a solution for normalisation of scale. An illustration is provided on laser altimetry data. In the last part, a revue of *ad hoc* statistical comparative methods is given. We then define canonical procedures to integrate phylogenetic proximities in data analysis: orthonormal basis and phylogenetic proximity matrices are introduced.

The conclusion tackle on biometry practice and the relations taking place between experimental data, mathematical tools, and computer science.

MOTS-CLES en anglais : multivariate analysis, multiscale analysis, comparative analysis, autocorrelation, statistical ecology, R software

INTITULE ET ADRESSE DE L'U.F.R. OU DU LABORATOIRE :

Laboratoire de Biométrie et Biologie Evolutive, UMR 5558

SOMMAIRE

INTRODUCTION.....	1
CHAPITRE 1.....	9
1.1. INTRODUCTION.....	11
1.2. L'ESPACE VUE AU TRAVERS DU VOISINAGE.....	20
1.2.1. Définition	
1.2.2. Relations de voisinage	
1.2.3. Pondérations de voisinage	
1.3. INDICES UNIVARIÉS DE LA STRUCTURE SPATIALE.....	33
1.3.1. L'indice I de Moran (1948, 1950)	
1.3.2. Le coefficient de contiguïté c de Geary (1954)	
1.3.3. Quand les deux écoles se rejoignent ...	
1.3.4. Tests contre l'absence de structure spatiale	
1.4. HESITATIONS METHODOLOGIQUES.....	44
1.4.1. L'école de Lebart : variances et covariances locales	
1.4.2. L'école de l'auto-corrélation spatiale multivariée	
1.5. GÉNÉRALISATION DE L'APPROCHE DE WARTENBERG.....	49
1.5.1. Principes	
1.5.2. Définitions	
1.5.3. La fonction multispati(...)	
1.5.4. Un test de permutation multivarié contre l'absence de structure spatiale	
1.6. ILLUSTRATIONS.....	58
1.6.1. Analyses à composantes cartographiables	
1.6.2. Une information exclusivement cartographiable	
1.6.3. Mélanges entre variance globale et variance locale	
1.7. DISCUSSION ET PERSPECTIVES.....	66
1.8. BIBLIOGRAPHIE.....	69
CHAPITRE 2.....	77
2.1. INTRODUCTION.....	79
2.2. DONNÉES D'ALTIMÉTRIE LASER.....	80
2.2.1. Contexte	
2.2.2. Description de l'expérience	
2.2.3. Les données	
2.3. STRUCTURE D'UNE VARIABLE QUANTITATIVE.....	85
2.4. FAMILLES DE K FORMES BILINÉAIRES SYMÉTRIQUES.....	89
2.4.1. Définitions	
2.4.2. La classe d'objets 'kfbs'	
2.4.3. Formes de Geary/Lebart : le variogramme	
2.4.4. Formes de Moran/Smouse : le corrélogramme	
2.4.5. Formes de Greig-Smith/Noy-Meir : les msbs	
2.4.6. Formes de Hill : les tlv	
2.4.7. Typologie d'un ensemble de formes bilinéaires	
2.5. BASES ORTHONORMÉES ET FAMILLES DE K PROJECTEURS.....	107
2.5.1. Définitions	
2.5.2. La classe d'objets 'orthobasis'	
2.5.3. Les bases associées à la diagonalisation des matrices symétriques	

2.5.4.	Expression analytique des vecteurs propres de l'opérateur de Méot	
2.5.5.	La base associée à l'analyse spectrale à une dimension	
2.5.6.	Les bases d'ondelettes à une dimension	
2.6.	NORMALISATION DES FORMES BILINÉAIRES.....	129
2.6.1.	Introduction	
2.6.2.	Définitions	
2.6.3.	Typologie de structures	
2.7.	APPLICATIONS AUX DONNÉES D'ALTIMÉTRIE LASER.....	139
2.8.	DISCUSSION ET PERSPECTIVES.....	139
2.9.	BIBLIOGRAPHIE.....	140
CHAPITRE 3.....		145
3.1.	INTRODUCTION.....	147
3.2.	LA PHYLOGÉNIE COMME NOUVELLE CLASSE DE DONNÉES.....	151
3.2.1.	Définitions	
3.2.2.	La classe d'objets 'phylog'	
3.3.	REPRÉSENTATION GRAPHIQUE DES DONNÉES.....	158
3.3.1.	La fonction symbols.phylog(...)	
3.3.2.	La fonction dotchart.phylog(...)	
3.3.3.	La fonction table.phylog(...)	
3.4.	LA MÉTHODE DES CONTRASTES.....	162
3.4.1.	Le principe des contrastes phylogénétiques	
3.4.2.	La métrique phylogénétique	
3.4.3.	Usage de la méthode des contrastes	
3.5.	LE TEST D'ABOUHEIF (1999).....	177
3.5.1.	Principe du test d'Abouheif	
3.5.2.	Le cas d'une variable quantitative	
3.5.3.	Le cas d'une variable qualitative	
3.5.4.	La matrice de proximité A	
3.5.5.	Conclusions	
3.6.	DU CORRÉLOGRAMME A L'ORTHOGRAM.....	188
3.7.	DISCUSSION ET PERSPECTIVES.....	190
3.8.	BIBLIOGRAPHIE.....	192
CONCLUSION.....		197
BIBLIOGRAPHIE.....		201

INTRODUCTION

A en croire le titre de la prochaine réunion annuelle conjointe de l'Ecological Society of America (ESA) et de l'International Congress of Ecology (INTECOL), qui se tiendra à Montréal du 7 au 12 août 2005 (<http://abstracts.co.allenpress.com/esa/entrance.html>), les notions de structures (« *pattern* ») et d'échelles (« *scale* ») sont bien des questions centrales de l'écologie (Levin, 1992). De fait, la plupart des systèmes écologiques présentent une importante variabilité dans l'espace et dans le temps de leurs principales caractéristiques (biomasse, composition spécifique, ...), variabilité qui est à la fois déterminant et conséquence de leur dynamique d'ensemble (Hanski, 1994). Selon Frontier et Pichod-Viale (1990), « *une des questions fondamentales de l'analyse actuelle des écosystèmes est précisément leur stratégie d'occupation de l'espace-temps, et ce, à toutes les échelles d'observation* ». Etudier la variabilité spatiale et temporelle qui affecte populations, peuplements et écosystèmes, sur une large gamme d'échelles est donc au cœur des préoccupations des écologues.

Cet engouement a suscité assez vite une demande méthodologique des écologues vis-à-vis des statisticiens, assurant le développement d'échanges interdisciplinaires et favorisant l'émergence de méthodes statistiques aptes à mettre en évidence les principales échelles de variations. Quel que soit l'objectif recherché, la plupart de ces études ont fait l'objet d'un échantillonnage spatialisé de plusieurs unités statistiques de façon répétée dans le temps. Elles ont conduit à l'obtention d'un ensemble complexe de données, généralement multivariées. La caractéristique principale de ces données, hormis leur caractère multivarié, est donc l'ordonnement des unités statistiques (relevés, populations ou organismes ...) selon un critère spatial ou temporel. Par conséquent, chaque unité statistique ne peut être considérée comme indépendante des autres dans la mesure où elle entretient avec elles des relations de proximité spatiale et/ou temporelle. Les relations de voisinage entre stations de mesure sur un réseau hydrographique constituent un exemple de proximités spatiales fréquemment rencontré en écologie des eaux douces (Poizat & Pont, 1996). De même, les relations de parenté entre organismes sont un cas particulier des proximités temporelles, que l'on appellera « proximités évolutives » pour les distinguer des proximités temporelles plus classiquement étudiées. En effet, chaque organisme peut être d'abord perçu comme faisant partie de groupes caractérisés par l'existence d'un ancêtre commun et possédant avec d'autres groupes de même nature des relations de parenté. Ces relations sont généralement traduites

dans des systèmes hiérarchiques obéissant à un certain nombre de règles de construction exprimées par des classifications hiérarchiques, des taxonomies ou des arbres phylogénétiques.

De manière générale, ces ressemblances, qu'elles soient spatiales, temporelles ou évolutives peuvent être vue comme des contraintes : l'existence d'un plan d'organisation commun difficile à contrôler n'autorise pas n'importe quelle variation. Il ne peut être négligé et doit être pris en compte lors de l'analyse des données. Parfois la caractérisation de la structure inhérente à ce plan sous-jacent est l'objectif majeur de l'expérimentation (cartographie de l'abondance d'une espèce, évolution dans le temps d'un indice, ...). D'autres fois, ce plan constitue un facteur de confusion qu'il est souhaitable d'éliminer avant toute analyse des données. De plus, d'un point de vue purement statistique, la non indépendance des unités statistiques invalide l'hypothèse classique selon laquelle les observations individuelles sont des réalisations indépendantes d'une même variable aléatoire, ce qui a des conséquences fâcheuses sur les estimateurs comme la moyenne (Bivand, 1980) ou sur la pertinence des procédures de randomisation (Fortin & Jacquez, 2000).

L'intégration des contraintes spatiales, temporelles et évolutives apparaît donc comme un problème incontournable en écologie statistique. Cet intérêt a été ressenti très tôt en écologie végétale (Greig-Smith, 1952; Hill, 1973), et de nombreuses méthodes d'analyse ont été proposées depuis (Dale, 1999), dans tous les champs de l'écologie, pour traiter des données de nature variée (Dale et al., 2002). Une synthèse assez complète, parue dans *Ecography* (2002), fait le tour des outils et concepts développés autour de la problématique spatiale (Dale et al., 2002; Dungan et al., 2002; Keitt et al., 2002; Koenig & Knops, 1998; Legendre et al., 2002; Liebhold & Gurevitch, 2002; Perry et al., 2002).

Ce travail s'inscrit dans la perspective de chercher de nouvelles méthodes d'analyse de données prenant en compte une notion de proximité entre individus statistiques. Le cadre que nous nous étions fixé au départ était celui de l'ordination multiéchelle (Noy-Meir & Anderson, 1971). Ce problème spécifique n'a pourtant été que partiellement abordé au cours de cette thèse (Couteron & Ollier, sous presse), car son étude nécessitait de revenir en amont sur plusieurs problèmes sous-jacents qui paraissaient résolus sans l'être. Ce mémoire est donc composé de trois parties, chacune d'elle traitant d'un des problèmes sous-jacents.

Dans la première partie, j'aborde le problème de l'ordination sous contrainte spatiale avec comme objectif d'étendre les analyses sous contrainte déjà existantes à l'ensemble des analyses multidimensionnelles utilisées en écologie. Pour cela, une position stratégique a été adoptée quant à la manière d'introduire la contrainte spatiale. En effet, depuis une

cinquantaine d'années, deux écoles coexistent : celle de la variance locale qui s'est développée selon la logique de l'indice de Geary (Geary, 1954), et celle de l'autocorrélation spatiale qui s'est développée selon la logique de l'indice de Moran (Moran, 1948). C'est en repartant de ces débats théoriques exposés dans la bibliographie que l'on a abouti au développement de deux approches assez générales permettant l'intégration des proximités spatiales en analyse multidimensionnelle (Couteron & Ollier, sous presse; Ollier et al., soumis). Les procédures sont décrites puis illustrées à partir de quelques situations expérimentales.

Dans la seconde partie, je relate les résultats obtenus suite à une consultation statistique initiée au cours de mon DEA et poursuivie en thèse. Elle avait pour objectif d'étudier la faisabilité d'une typologie des couverts forestiers à partir de données d'altimétrie laser levées en Guyane Française. Cette étude nous a conduit à aborder le problème de la typologie de structures multiéchelles. En partant cette fois-ci des données, on a donc été amené à aborder les concepts théoriques associés à la définition et la comparaison d'une ou plusieurs métriques de la structure. Les programmes associés à la mise en œuvre de ces procédures sont décrits dans cette partie. Le traitement des données d'altimétrie laser est exposé et discuté (Ollier et al., 2003).

La dernière partie aborde la mesure de la structure d'un trait biologique dans un arbre phylogénétique. L'article fondateur est celui de Felsenstein (Felsenstein, 1985) : il pose clairement le problème de la non indépendance des organismes, supports de mesure des traits d'histoire de vie en écologie évolutive. Les proximités entre les organismes s'expriment au travers d'un arbre phylogénétique, ce qui introduit implicitement une nouvelle classe de données en écologie statistique. Comme souvent en biologie, cette nouvelle classe de données, en définissant de nouveaux besoins, a conduit au développement de pratiques statistiques *ad hoc*, dont l'objectif était de décrire la variabilité d'un trait biologique dans un arbre phylogénétique. Ces pratiques sont implicitement en connexion avec des modèles centraux, ou tentent de l'être, mais les auteurs se sont perdus dans les franges complexes de la statistique. On propose alors de définir des pratiques canoniques (Ollier et al., sous presse), en s'appuyant sur les problèmes soulevés par les pratiques *ad hoc*. Les données sont alors introduites afin de vérifier le réalisme de ces nouveaux outils.

Au cours de cette thèse, je me suis donc intéressé à trois classes de problèmes. Ce travail s'est fait dans le cadre d'échanges interdisciplinaires, à l'interface de la statistique et de l'écologie. Il s'inscrit dans le champ de la biométrie, dont « *l'objectif, est de participer à*

l'élaboration d'une méthodologie nouvelle à la disposition des sciences expérimentales, c'est-à-dire dans les sciences expérimentales (Legay, 1976)». Par conséquent, cette thèse est une illustration parmi beaucoup d'autres (Auda, 1983; Hanafi, 1997; Méot, 1992; Mercier, 1991; Torre, 1996; Yoccoz, 1988) d'une pratique de la biométrie et des relations qui s'établissent entre donnée expérimentale, langage mathématique et mise en oeuvre informatique. En effet, la pratique de la biométrie passe nécessairement par la prise en compte de ces trois éléments et c'est en soit un objet de recherche : « l'ignorance du mathématicien en face des objets biologiques et celle du biologiste en face d'un langage sont une donnée d'expérience, quotidienne et inépuisable (Chessel, 1992) ». Ainsi, j'évoque en conclusion la diversité des interactions entre les trois composantes « données-modèles-programmes », en soulignant le rôle des échanges interdisciplinaires en analyse des données. En particulier, j'insiste sur la structure et la fonction du logiciel R (Ihaka & Gentleman, 1996), en montrant dans quelle mesure ce dernier a constitué un élément central du dialogue. L'essentiel du travail a en effet été réalisé dans le cadre des relations interdisciplinaires établies autour du logiciel R. C'est pourquoi j'ai tenu à présenter l'ensemble des données et programmes développés au cours de cette thèse en annexes. De même, les lignes de commandes à l'origine des figures sont explicitées (en caractères rouges) afin que l'utilisateur puisse s'approprier d'autant plus facilement les outils développés au cours de cette thèse. Les données, ainsi que les fonctions sont en partie intégrées à la librairie ade4 (Chessel et al., soumis), l'autre partie étant disponible sur simple demande, car « la libre circulation des données et des programmes est un facteur décisif de développement (Chessel, 1992)».

BIBLIOGRAPHIE

- Auda, Y.** (1983) Rôle des méthodes graphiques en analyse des données : application au dépouillement des enquêtes écologiques. Thèse de 3^o cycle, Université Lyon 1.
- Bivand, R.** (1980) A Monte Carlo study of correlation estimation with spatially autocorrelated observations. *Quaestiones Geographicae*, 6, 5-10.
- Chessel, D.** (1992) Echanges interdisciplinaires en analyse de données écologiques. Mémoire d'habilitation. Université Lyon 1.
- Chessel, D., Dufour, A.-B., & Thioulouse, J.** (Submitted) The ade4 package. *R News*.
- Couteron, P. & Ollier, S.** (sous presse) A generalized variogram-based framework for multiscale ordination. *Ecology*.
- Dale, M.R.T.** (1999) *Spatial pattern analysis in plant ecology* Cambridge University Press.
- Dale, M.R.T., Dixon, P., Fortin, M.J., Legendre, P., Myers, D., & Rosenberg, M.** (2002) Conceptual and mathematical relationships among methods for spatial analysis. *ecography*, 25, 558-577.
- Dungan, J.L., Perry, J., Dale, M.R.T., Citron-Pousty, S., Fortin, M.J., Jakomulska, A., Legendre, A., Miriti, M., & Rosenberg, M.S.** (2002) A balanced view of scaling in spatial statistical analysis. *Ecography*, 25, 626-640.
- Felsenstein, J.** (1985) Phylogenies and the comparative method. *The American Naturalist*, 125, 1-15.
- Fortin, M.-J. & Jacquez, G.M.** (2000) Randomization tests and spatially autocorrelated data. *Bulletin of the Ecological Society of America*, 81, 201-205.
- Frontier, S. & Pichod-Viale, D.** (1990) *Ecosystèmes. Structure, fonctionnement, evolution*, Second edn. Dunod.
- Geary, R.C.** (1954) The contiguity ratio and statistical mapping. *The incorporated Statistician*, 5, 115-145.
- Greig-Smith, P.** (1952) The use of random and contiguous quadrats in the study of the structure of plant communities. *Annals of Botany, London*, 16, 293-316.
- Hanafi, M.** (1997) Structure de l'ensemble des analyses multivariées des tableaux de données à trois entrées : éléments théoriques et appliqués. Thèse de doctorat, Université Lyon 1.
- Hanski, I.** (1994) Spatial scale, patchiness and population dynamics on land. *Phil. Trans. R. Soc. London*, 343B, 19-25.

Hill, M.O. (1973) The intensity of spatial pattern in plant communities. *Journal of Ecology*, 61, 225-235.

Ihaka, R. & Gentleman, R. (1996) R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5, 299-314.

Keitt, T.H., Bjørnstad, O.N., Dixon, P., & Citron-Pousty, S. (2002) Accounting for spatial pattern when modeling organism-environment interactions. *Ecography*, 25, 616–625.

Koenig, W.D. & Knops, J.M.H. (1998) Testing for spatial autocorrelation in ecological studies. *Ecography*, 21.

Legay, J.M. (1976) Pour une Biométrie. *Statistique et Analyse des Données*, 1, 5-11.

Legendre, P., Dale, M.R.T., Fortin, M.J., Gurevitch, J., Hohn, M., & Myers, D. (2002) The consequences of spatial structure for the design and analysis of ecological surveys. *Ecography*, 25, 601-615.

Levin, S.A. (1992) The problem of pattern and scale in ecology. *Ecology*, 73, 1943-1967.

Liebold, A.M. & Gurevitch, J. (2002) Integrating the statistical analysis of spatial data in ecology. *ecography*, 25, 553-557.

Méot, A. (1992) Explication de contraintes de voisinage en analyse multivariée. Application dans le cadre de problématiques agronomiques. Thèse de 3^o cycle, Université Claude Bernard (Lyon I).

Mercier, P. (1991) Analyses des relations espèces-environnement et étude de la co-structure d'un couple de tableaux. Thèse de doctorat, Université Lyon 1.

Moran, P.A.P. (1948) The interpretation of statistical maps. *Journal of the Royal Statistical Society, B*, 10, 243-251.

Noy-Meir, I. & Anderson, D.J. (1971). Multivariate pattern analysis, or multiscale ordination: towards a vegetation hologram ? In *Statistical Ecology, III Many species populations ecosystems and systems analysis* (eds G.P. Patil, E.C. Pielou & W.E. Waters), pp. 208-231. Pennsylvania State University Press.

Ollier, S., Chessel, D., Couteron, P., Péliissier, R., & Thioulouse, J. (2003) Comparing and classifying one-dimensional spatial patterns: an application to laser altimeter profiles. *Remote Sensing of Environment*, 85, 453-462.

Ollier, S., Couteron, P., & Chessel, D. (sous presse) Orthonormal transforms to describe and test the phylogenetic signal. *Biometrics*.

Ollier, S., Dray, S., & Chessel, D. (soumis) Taking into account spatial dependence in multivariate analysis: a generalization of Wartenberg's multivariate spatial correlation. *Geographical Analysis*.

Perry, J.N., Liebhold, A.M., Rosenberg, M.S., Dungan, J., Miriti, M., Jakomulska, A., & Citron-Pousty, S. (2002) Illustrations and guidelines for selecting statistical methods for quantifying spatial patterns in ecological data. *Ecography*, 25, 578-600.

Poizat, G. & Pont, D. (1996) Multi-scale approach to species-habitat relationships: juvenile fish in a large river section. *Freshwater Biology*, 36, 611-622.

Torre, F. (1996) Analyse de co-structure de deux tableaux totalement appariés : application à la comparaison de deux méthodes d'échantillonnage en écologie. Thèse de doctorat, Université Lyon 1.

Yoccoz, N. (1988) Le rôle du modèle euclidien d'analyse des données en biologie évolutive. Thèse de doctorat, Université Lyon 1.

ORDINATION SOUS CONTRAINTES SPATIALES

Développement méthodologique à partir d'un débat bibliographique

1.	INTRODUCTION.....	11
2.	L'ESPACE VUE AU TRAVERS DU VOISINAGE	20
2.1.	Définition	20
2.2.	Relations de voisinage.....	21
2.3.	Pondérations de voisinage.....	26
3.	INDICES UNIVARIÉS DE LA STRUCTURE SPATIALE.....	33
3.1.	L'indice I de Moran (1948, 1950).....	33
3.2.	Le coefficient de contiguïté c de Geary (1954).....	36
3.3.	Quand les deux écoles se rejoignent	38
3.4.	Tests contre l'absence de structure spatiale	41
4.	HESITATIONS METHODOLOGIQUES.....	44
4.1.	L'école de Lebart : variances et covariances locales	44
4.2.	L'école de l'auto-corrélation spatiale multivariée.....	46
5.	GÉNÉRALISATION DE L'APPROCHE DE WARTENBERG	49
5.1.	Principes	49
5.2.	Définitions.....	53
5.3.	La fonction multispati(...).	54
5.4.	Un test de permutation multivarié contre l'absence de structure spatiale.....	56
6.	ILLUSTRATIONS.....	58
6.1.	Analyses à composantes cartographiables	58
6.2.	Une information exclusivement cartographiable	61
6.3.	Mélanges entre variance globale et variance locale.....	63
7.	DISCUSSION ET PERSPECTIVES	66
8.	BIBLIOGRAPHIE	69

1. INTRODUCTION

La majorité des observations écologiques sont généralement référencées au temps et à l'espace. De plus, elles sont souvent multidimensionnelles, l'information étant disponible pour un grand nombre de descripteurs simultanément. On est donc confronté en écologie à l'omniprésence naturelle du temps et de l'espace associés à des données de nature multivariées. En économie on rencontre également ce genre de problèmes, les données multivariées étant généralement associées à des unités administratives correspondant à des enregistrements surfaciques (Jayet, 1999). C'est ainsi que s'est posé initialement la question au travers d'un des jeux de données les plus célèbres de la statistique spatiale, celui des comtés d'Irlande de Geary (1954). Les données d'origine dans l'article fondateur de Geary sont curieusement multivariées (Figure 1.1, Annexe 1.11).

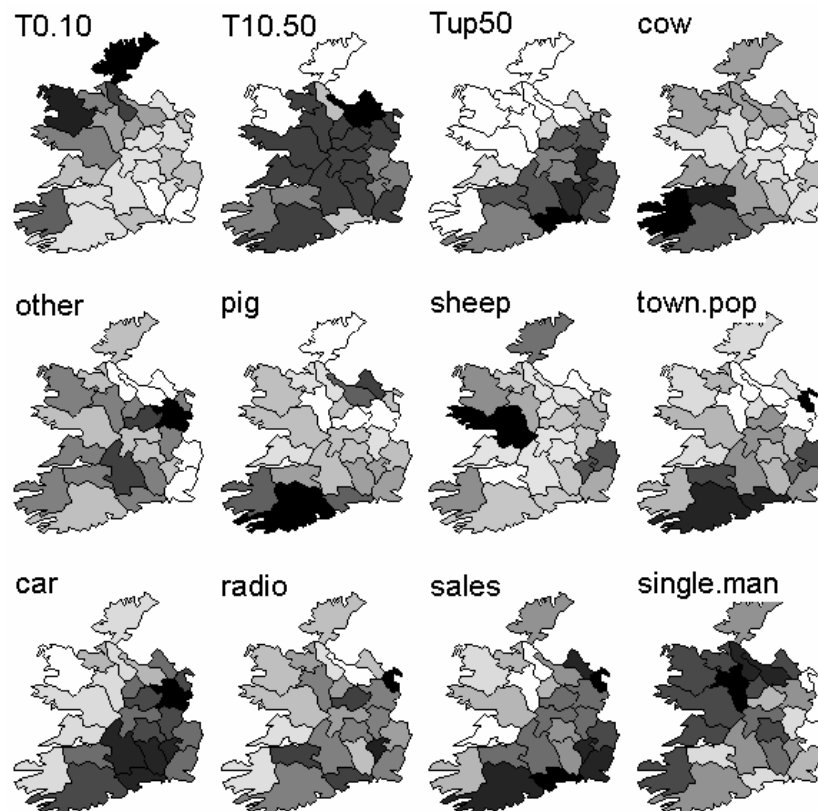


Figure 1.1 : cartographie par unité surfacique pour 25 districts d'Irlande. Code des variables : **1-2-3** répartition (en 1 pour 1000) des propriétés agricoles en 3 groupes d'imposition (T0.10 <10 £, T10.50 10-50 £, Tup50 >50 £). **4-5-6-7** Nombres moyens d'animaux pour 1000 acres de prairies et cultures 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 moyenne par habitant en £ **12- single.man** Pourcentage de célibataires parmi les hommes de 30-34 ans en 1 pour 1000. Données normalisées.

Si l'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. La même question est posée pour des données phytécologiques (Figure 1.2 et Figure 1.3, Annexe 1.13 et Annexe 1.17).

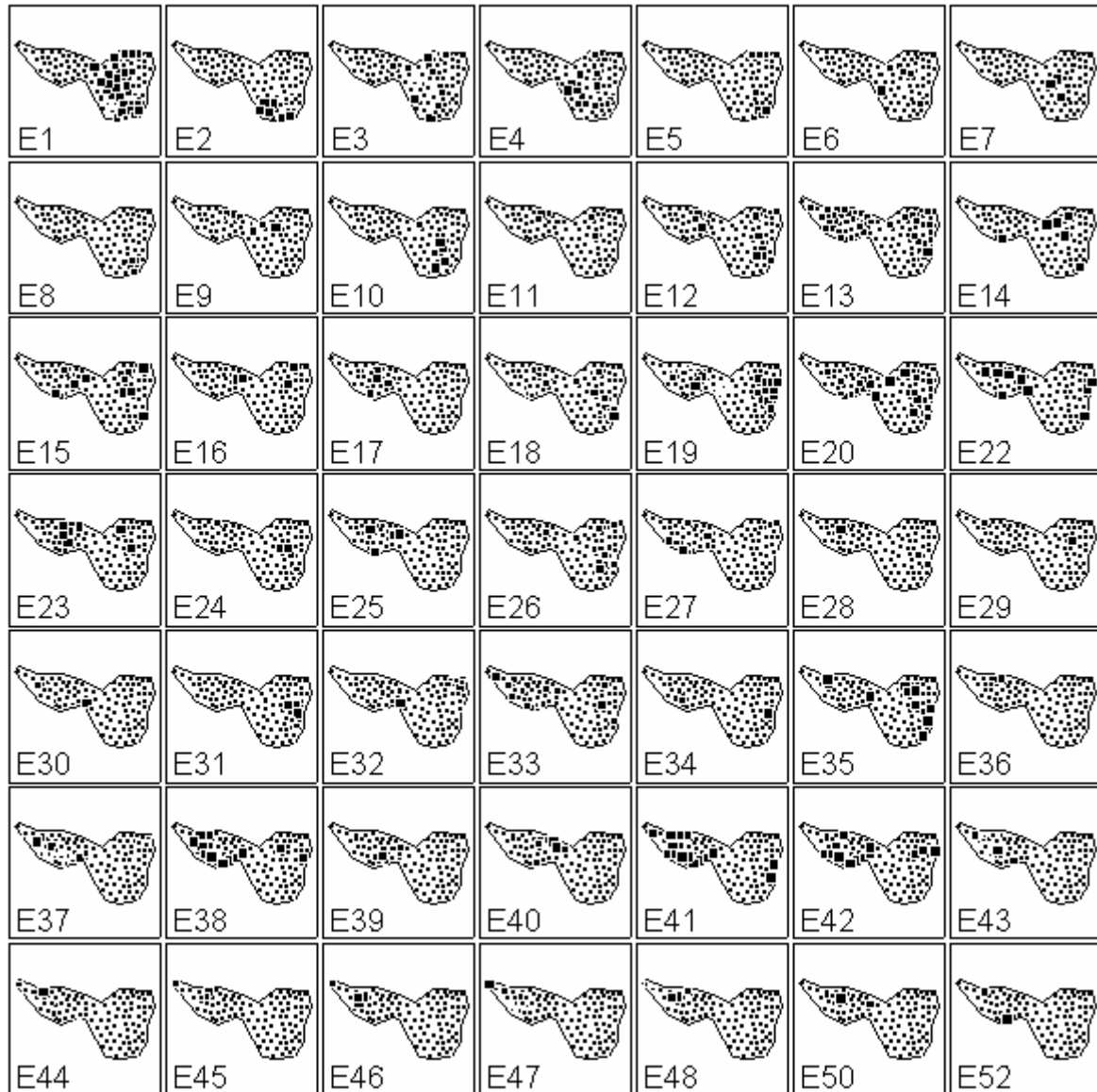


Figure 1.2 : 49 espèces végétales dans une enquête phytécologique sur une plaine côtière marécageuse la Mafragh (Annaba, Algérie) (de Belair, 1981) comportant 97 relevés (16x8 km) floristiques.



Figure 1.3 : 35 espèces d’Oribates dans 70 carottes de sol de 5 cm de diamètre et 10 cm de profondeur (Borcard et al., 1992).

La description phytosociologique fortement multivariée porte sur des mesures élémentaires très simples et diverses (présence-absence 0-1, notes d’abondance-dominance entière 0-7, classe de recouvrement ou codage semi quantitatif). Une autre discipline ou le multivarié et le spatial font un couple particulièrement recherché est celui de la génétique. Multivarié est par essence l’enregistrement de la variabilité génétique (génotype d’un individu ou fréquences alléliques d’un groupe sur un ou plusieurs loci). Dans l’exemple considéré (Annexe 1.2), issu du travail de Fievet et al. (2001), on a par exemple 6 loci avec respectivement 2, 5, 2, 4, 4 et 5 allèles et des fréquences alléliques. Spatialisé est par essence l’échantillonnage des individus. L’espace est cependant celui du fonctionnement biologique. L’exemple de Fievet et al. (2001) est saisissant pour cette plasticité de la notion spatiale, qui

dépasse largement la notion de coordonnées ou même celle de distances. L'analyse porte en effet sur une crevette *Atya innocous* qui vit et se reproduit en eau douce mais dont les larves dévalent les cours d'eau et possèdent une période de croissance en mer. Les stations sont situées sur les rivières de Basse-Terre en Guadeloupe (Figure 1.4).

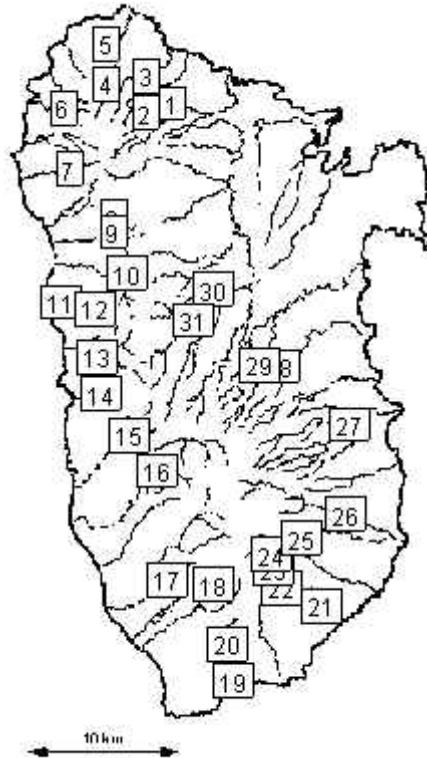


Figure 1.4 : 31 stations situées sur les rivières de Basse-Terre en Guadeloupe (Fievet et al., 2001).

On considérera par rapport aux données de Fievet (2001) que deux stations sont proches si elles sont dans le même bassin versant ou si elles sont dans deux bassins versants voisins (au sens de la distance à parcourir le long de la côte entre les embouchures). L'hypothèse est la suivante:

- soit le mode de fonctionnement génère une seule population avec brassage complet au cours de la migration (il n'y a pas de structure spatiale),
- soit ce dernier induit une structure spatiale dans la composante génétique avec une ressemblance plus forte entre stations plus proches.

Ainsi, au travers de ces divers exemples, on constate que les tableaux de données écologiques peuvent contenir des variables qualitatives, quantitatives ou distributionnelles. Le tableau peut aussi être homogène lorsque dans chaque cellule, à chaque ligne et chaque colonne, les mesures, répétées dans le temps ou l'espace, portent sur la même variable. Par

exemple, les températures moyennes de 30 villes mesurées pendant 12 mois forment un tableau homogène (Figure 1.5, Annexe 1.21). Certes, il fait plus chaud au sud qu'au nord mais que peut-on dire d'autre sur la variabilité annuelle de la structure spatiale des températures en France ?

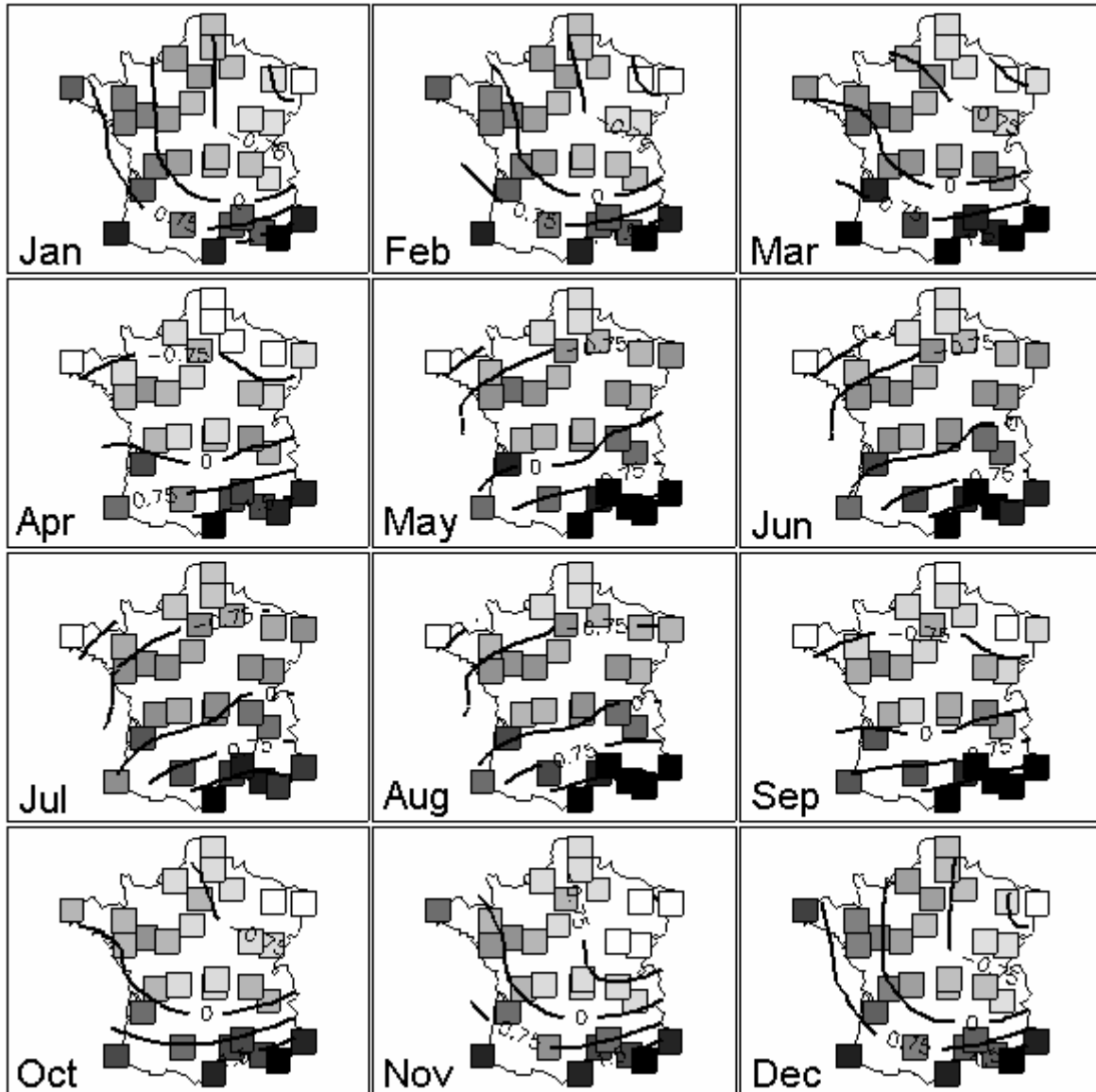


Figure 1.5 : températures moyennes normalisées par mois exprimées en niveau de gris pour 30 villes de France et 12 mois de l'année (Besse, 1979). Sont également représentées les courbes de niveaux estimées par régression polynomiale locale (`fonction loess(...)` de la library `stats`).

De même, la production annuelle de clémentines suivie sur 15 ans pour 20 clémentiniers forme un tableau homogène (Figure 1.6, Annexe 1.6). Certes, la production augmente avec le temps mais que peut-on dire d'autre sur la variabilité de l'évolution temporelle de la production de clémentines ?

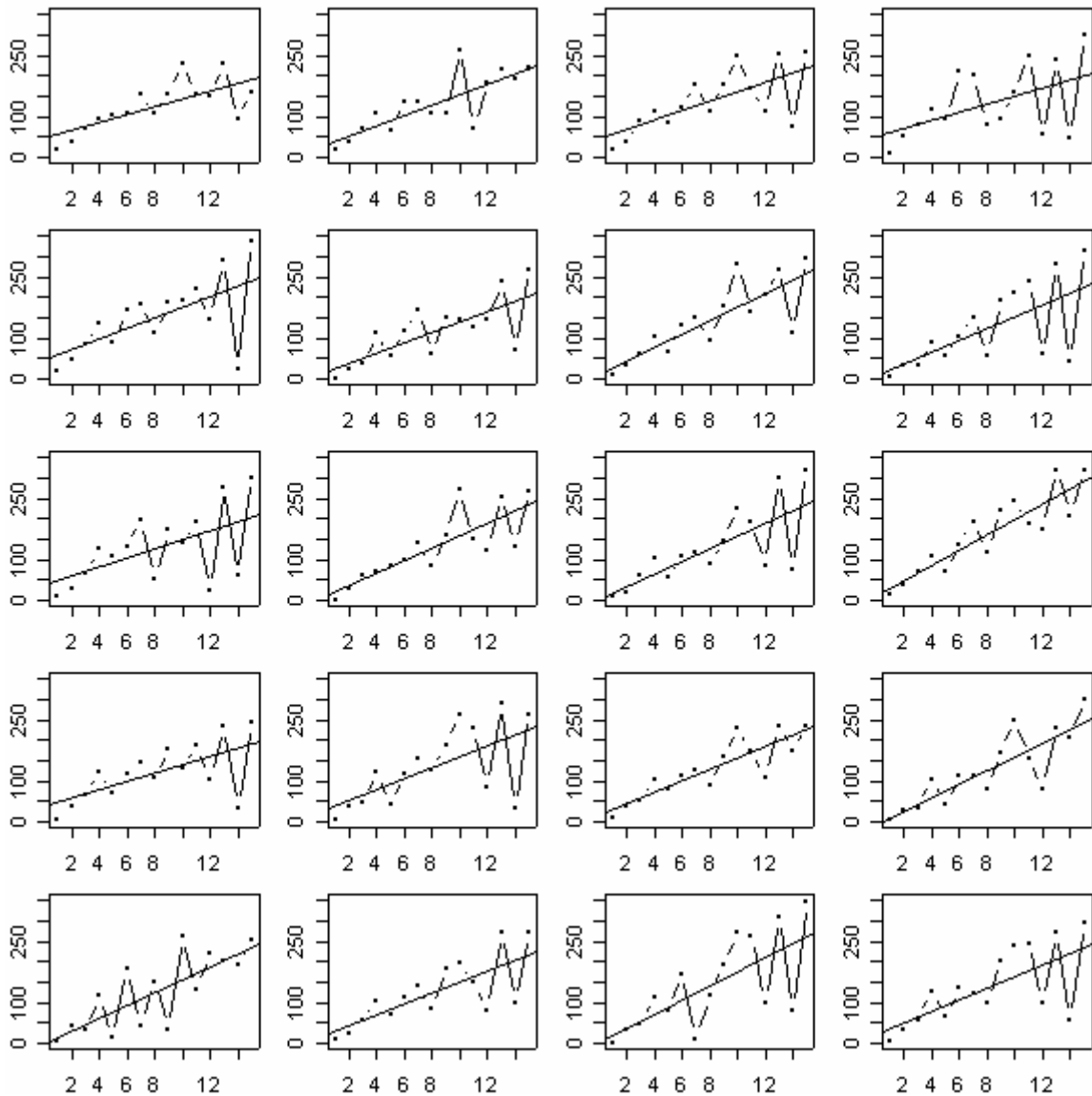


Figure 1.6 : évolution de la production annuelle de 20 clémentiniers pendant 15 ans (Tisné-Agostini, 1988). Sont représentées les droites de régression respectives de la production en fonction du temps.

Bien que l'information temporelle et/ou spatiale soit omniprésente, elle entre rarement de façon explicite dans le traitement des données. Pourtant, elle apparaît dans nombre d'études au moment de l'interprétation. Le premier article sur l'ACP en écologie (Goodall, 1954) comme l'un des premiers articles sur l'AFC en écologie (Hatheway, 1971) cartographient des coordonnées factorielles et notent l'efficacité de cette pratique. Hill (Hill, 1974) puis Estève (Estève, 1978) représentent des coordonnées factorielles le long d'un transect tout comme Dessier et Laurec (Dessier & Laurec, 1978) les représentent en fonction du temps. Dans tous les cas, sans introduire la structure du plan d'observation (stations sur une carte, placettes sur un transect, prélèvements dans une chronique), on obtient avec les analyses classiques une expression parfaitement satisfaisante des résultats exprimés dans cette

structure. On peut pour s'en convaincre reprendre l'exemple traité par J. Estève dans l'article précité (Annexe 1.20). La Figure 1.7 restitue l'évolution, le long du transect, de la présence des 15 espèces principales puis celle de la première coordonnée de chaque analyse (ACP et AFC).

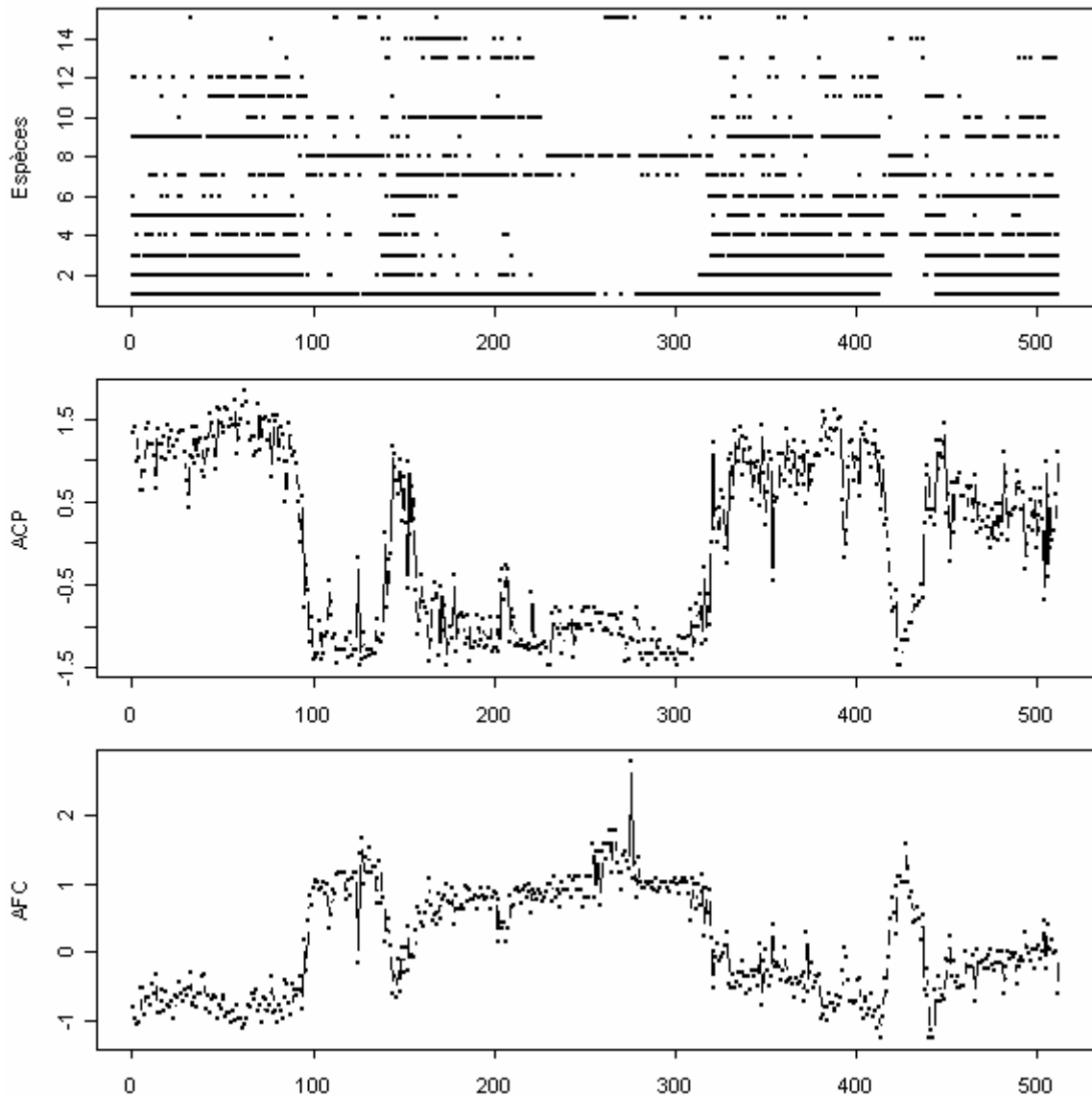


Figure 1.7 : présence-absence de 15 espèces le long d'un transect de 512 placettes en steppe semi-aride (**en haut**). Premières coordonnées factorielles (ACP : **au milieu**, AFC : **en bas**).

On note l'étroite similitude des deux résultats et la possibilité de faire dans un cas comme dans l'autre un découpage de l'espace qui intègre la structure multispécifique du tapis végétal. Tout se passe comme si la structure spatiale sous-jacente intervenait directement, alors qu'il n'en est rien. C'est ce que soulignent Grunsky et Agteberg (Grunsky & Agteberg, 1991) selon qui « *a conventional principal component analysis is sometime useful for enhancing information within multivariate data that are spatially related* ». L'idée est reprise

dans l'article de Solow (Solow, 1994) qui remarque également, « *in many applications, the first few principal components account for a large proportion of the total variance and are taken to represent trend* ». En fait, lorsque l'essentiel de la variance et de la covariance est une conséquence des variations spatiales (ou temporelles), une analyse simple telle que l'analyse en composante principale extrait la structure spatiale (ou temporelle) sous-jacente puisque c'est la source de toutes les variations.

Toutefois, ce n'est pas toujours le cas et deux objections majeures peuvent être soulignées. La première vient simplement du fait que la variabilité des données résulte parfois de la superposition d'une couche spatiale (ou temporelle) et d'une autre qui ne l'est pas. On peut vouloir éliminer l'information spatiale pour mettre en valeur celle qui ne l'est pas. C'est le point de vue défendu dans Borcard et al. (1992) qui donne une partition des variations de l'abondance d'espèces en quatre composantes indépendantes (spatiale, environnementale, environnementale structurée dans l'espace, indéfinie). On peut également rechercher l'objectif contraire à savoir extraire de manière optimale l'information spatiale. C'est ce que recherche Nielsen (1994) qui souhaite assurer une compression des données (images géoréférencées issus de la teledetection) tout en optimisant la qualité des images (rapport signal/bruit). Il remarque à juste titre que les « *principal components will not always produce components that show decreasing image quality with increasing component number. It is perfectly imaginable that certain types of noise have higher variance than certain types of signal components* ». La deuxième objection est liée à la complexité des structures spatiales (ou temporelles) en jeu. En effet, lorsque plusieurs processus spatiaux se superposent, les structures s'expriment souvent à de multiples échelles. Dans ce cas, les analyses factorielles classiques, bien qu'elles permettent généralement de dégager une information fortement structurée dans l'espace (ou dans le temps), ne permettent pas d'appréhender différentes échelles simultanément. Cette deuxième objection a fait l'objet de nombreux développements, définissant ainsi le champ de l'ordination multiéchelle (« *multi-scale ordination, MSO* »). La voie a été ouverte par Noy-Meir et Anderson (1971) qui résume parfaitement la situation : « *multivariate methods for the analysis of vegetation describe the patterns of covariation of species, but only at a single predetermined scale. Pattern analysis (such as block size variance analysis) describes variation of pattern over a wide range of scales, but only for one species at a time. A method is proposed which combines information from all species at all scales to produce an integrated representation of total pattern* ». Cette seconde objection est intimement liée à la première dans la mesure où elle ne pourra être sérieusement abordée

qu'une fois la première résolue. C'est pourquoi on s'est intéressé dans un premier temps, au problème d'ordination intégrant l'espace (ou le temps) en essayant d'être aussi général que possible, tant du point de vue des méthodes d'ordination envisagées que des manières d'intégrer l'espace (ou le temps).

Cette intervention active de l'espace est le fait des méthodes d'ordination locale et globale mais aussi des approches géostatistiques multivariées telles que l'analyse factorielle krigéante (Sandjivy & Galli, 1984). Toutefois, comme le font remarquer Royer (1984) puis Goulard et al. (1987), *« tant au niveau de la description que de l'estimation, les méthodes géostatistiques se révèlent conceptuellement très adaptées mais leur application pratique n'est efficace que dans le cas d'un petit nombre de variables régionalisées stationnaires. Pour un nombre important de variables, il apparaît que les méthodes multidimensionnelles d'analyse des données sont encore les seules utilisables d'un point de vue concret. Elles peuvent être utilisées pour dégager les variables qui seront soumises ensuite à l'étude géostatistique »*. Dans ces problèmes largement multivariés, on a d'abord besoin de trier les variables en fonction de leur *« pattern »* ou mode de variation dans l'espace, pour le moins sur l'existence d'une information spatialisée. Mais, plus profondément, une méthode multivariée a pour fonction essentielle de réduire le nombre de variables en faisant des combinaisons linéaires qui optimisent ce que l'on sait faire en univarié. La régression multiple donne la combinaison qui offre la meilleure régression simple, l'analyse discriminante donne la combinaison qui offre la meilleure analyse de variance, l'analyse canonique donne la combinaison des variables du premier tableau et celle du second tableau qui optimise la corrélation. En multivarié spatial, il en sera de même, l'objectif étant alors d'intégrer un critère lié à la structure spatiale dans la maximisation. La première difficulté vient du fait que si les points de mesure se suivent le long d'un transect, le seul numéro d'ordre des lignes des tableaux de données contient toute l'information de proximité entre points, qu'on s'en serve ou non. Dans tous les autres cas cette information doit être intégrée explicitement et il existe de multiples manières d'intégrer l'espace. La deuxième difficulté est liée à l'existence de deux critères en compétition depuis 50 ans. Il y a donc deux écoles (au moins) de statistiques multivariées spatialisées. C'est au travers des débats bibliographiques qui ont eu lieu autour de ces deux difficultés et que l'on retrace par la suite (paragraphes 2 et 3), qu'a émergé le concept d'une nouvelle ordination sous contrainte qui généralise celle définie par Wartenberg (1985). Ce choix est discuté dans le quatrième et le cinquième paragraphe de ce chapitre. Dans la suite, on parlera seulement de la notion d'espace, abandonnant les contraintes

temporelles. Bien que ces deux notions soient profondément différentes, les méthodes descriptives décrites par la suite sont facilement transposables de l'espace vers le temps, ce que nous illustrerons dans le sixième paragraphe. Par contre, la réciproque serait loin d'être vraie...

2. L'ESPACE VUE AU TRAVERS DU VOISINAGE

2.1. Définition

En écologie statistique, on peut intégrer l'espace de multiples manières. Une des plus simples est de prendre deux coordonnées (x_i, y_i) pour chaque unité statistique, ce qui associe à chaque couple de points une distance, par exemple $d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$. Plusieurs classes de données, telle que les enregistrements surfaciques (Figure 1.1), supportent mal cette réduction. De même, en hydrobiologie, les unités statistiques sont des tronçons de rivière et la distance n'a pratiquement aucun sens pour mesurer des proximités spatiales. Dans tous les cas, on peut par contre introduire l'espace en quantifiant comme on le désire le voisinage. En effet, sur l'ensemble des unités statistiques il existe une structure de contiguïté. Il peut s'agir d'une proximité plane si par exemple des individus sont des enregistrements surfaciques. Sont alors voisins dans le cas le plus simple deux unités surfaciques ayant une frontière commune. Les séries chronologiques induisent également des structures de proximité, chacun des moments pouvant être relié au suivant. Une contiguïté peut par ailleurs être définie sur des régions d'un espace euclidien de dimension quelconque, défini, par exemple, par un ensemble de variables.

De manière générale, on définit la contiguïté sur un ensemble de n unités statistiques I par un graphe dont les sommets sont les éléments de I et les arêtes relient un sommet i à ses voisins. L'ensemble des voisins de i , qui est contenu dans I est noté $V(i)$. On note \mathbf{W} la matrice carrée associée au graphe, de dimension égale au cardinal de I . On dira qu'il s'agit d'une matrice de pondérations de voisinage. Son terme général $w_{ii'}$, (positif ou nul), est le poids de l'élément i' dans le voisinage de i . Si i' n'est pas voisin de i , ce poids est nul. Si l'on veut faire jouer le même rôle à tous les voisins de i , le poids de chacun des voisins est réduit à 1. Le graphe devient alors un graphe non pondéré et la matrice de voisinage est directement l'expression du graphe de voisinage correspondant. On la notera \mathbf{M} . La possibilité de pondérer le graphe assouplit la notion de contiguïté et permet d'introduire des notions de proximité en donnant par exemple un poids plus important aux surfaces voisines

dont les frontières sont les plus grandes. De plus, un graphe n'est pas forcément symétrique, notamment lorsque l'on souhaite tenir compte des relations amont-aval pour des données spatiales ou des relations avant-après pour des données temporelles. Cette définition de la contiguïté par un graphe quelconque est assez générale pour recouvrir le plupart des situations auxquelles on peut être confronté en écologie. Les éléments de statistique spatiale s'appuieront donc sur la quantification du voisinage et la structure spatiale s'exprimera comme une relation quantitative mesurée sur chaque couple de points au travers du graphe de voisinage et des matrices qui lui sont associées. L'intégration de l'espace vue au travers du voisinage peut se faire de multiples manières, tant les façons de définir un graphe et les manières de pondérer les relations de voisinage sont variées. Cette extrême souplesse s'exprime parfaitement au travers des diverses fonctions de la librairie `spdep` développée dans R (Ihaka & Gentleman, 1996) par R. Bivand. En effet, selon R. Bivand, `spdep` « *is a collection of functions to create spatial weights matrix objects from polygon contiguities, from point patterns by distance and tessellations, for summarising these objects, and for permitting their use in spatial data analysis* ». C'est au travers de ces fonctions que l'on va présenter successivement les différentes manières de créer un graphe de voisinage et les différentes options pour les pondérer.

2.2. Relations de voisinage

Bien que la librairie des graphes de voisinage dans R soit la librairie `spdep` de R. Bivand, l'objet graphe de voisinage est conservé dans la librairie `ade4` comme une liste d'arêtes et forme la classe d'objet 'neig'. Cela permet de récupérer les graphes de voisinage éventuellement implantés dans l'ancienne version d'ADE-4. Dans `spdep`, les graphes de voisinages sont conservés comme liste de voisins et forment des objets de la classe 'nb'. Les deux classes d'objet sont équivalentes dans la mesure où elles contiennent la même information mais sous des formes différentes. Les fonctions `neig2nb(...)` et `nb2neig(...)` permettent de passer d'une classe d'objets à l'autre. Une remarque est très importante : la librairie de R. Bivand ne contient jamais de matrices et aucune des fonctions présentes ne manipule des matrices de voisinages (qui contiennent énormément de valeurs nulles). Ces fonctions n'ont donc pratiquement pas de limites en nombre de points, car elles n'utilisent que des listes de voisins et des listes de poids de voisinage (comparer les fonctions `moran.test(...)` et `geary.test(...)` de la librairie `spdep` avec la fonction `gearymoran(...)` (Annexe 2.3) de la librairie `ade4`). Les notations matricielles seront donc ici purement conceptuelles. On peut implémenter un graphe de voisinage dans R de multiples manières.

- La manière la plus simple est de définir manuellement la liste des arêtes, la liste des voisins ou la matrice du graphe de voisinage (Figure 1.8).

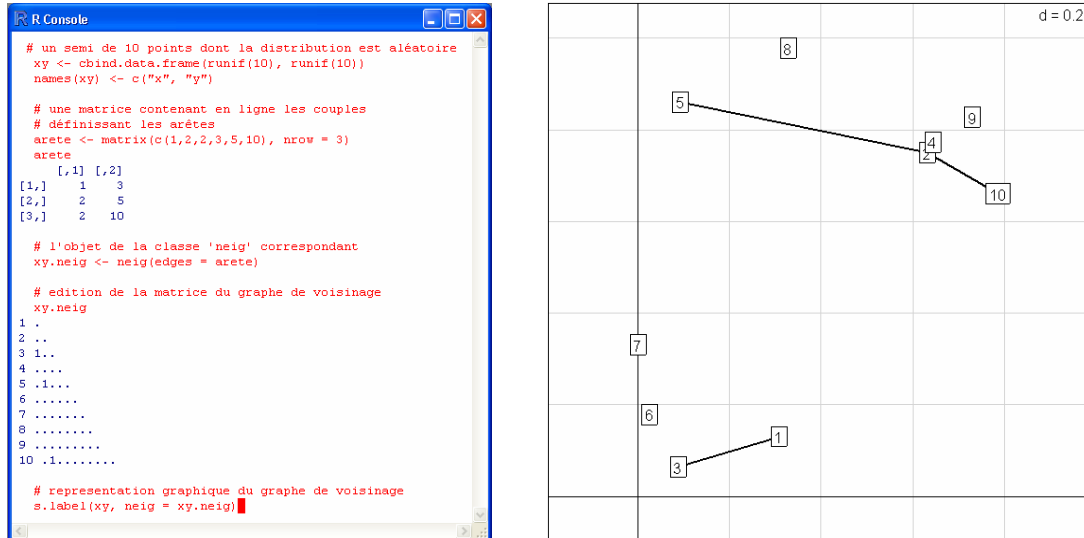


Figure 1.8 : console de commandes du logiciel R (à gauche). On peut y lire les instructions permettant de définir manuellement un graphe de voisinage (à droite) à partir de la liste des arêtes du graphe. La fonction centrale est la fonction `neig(...)`.

- On peut définir un certain nombre de graphes réguliers comme le graphe linéaire, le graphe circulaire, et les graphes sur une grille régulière (Figure 1.9).

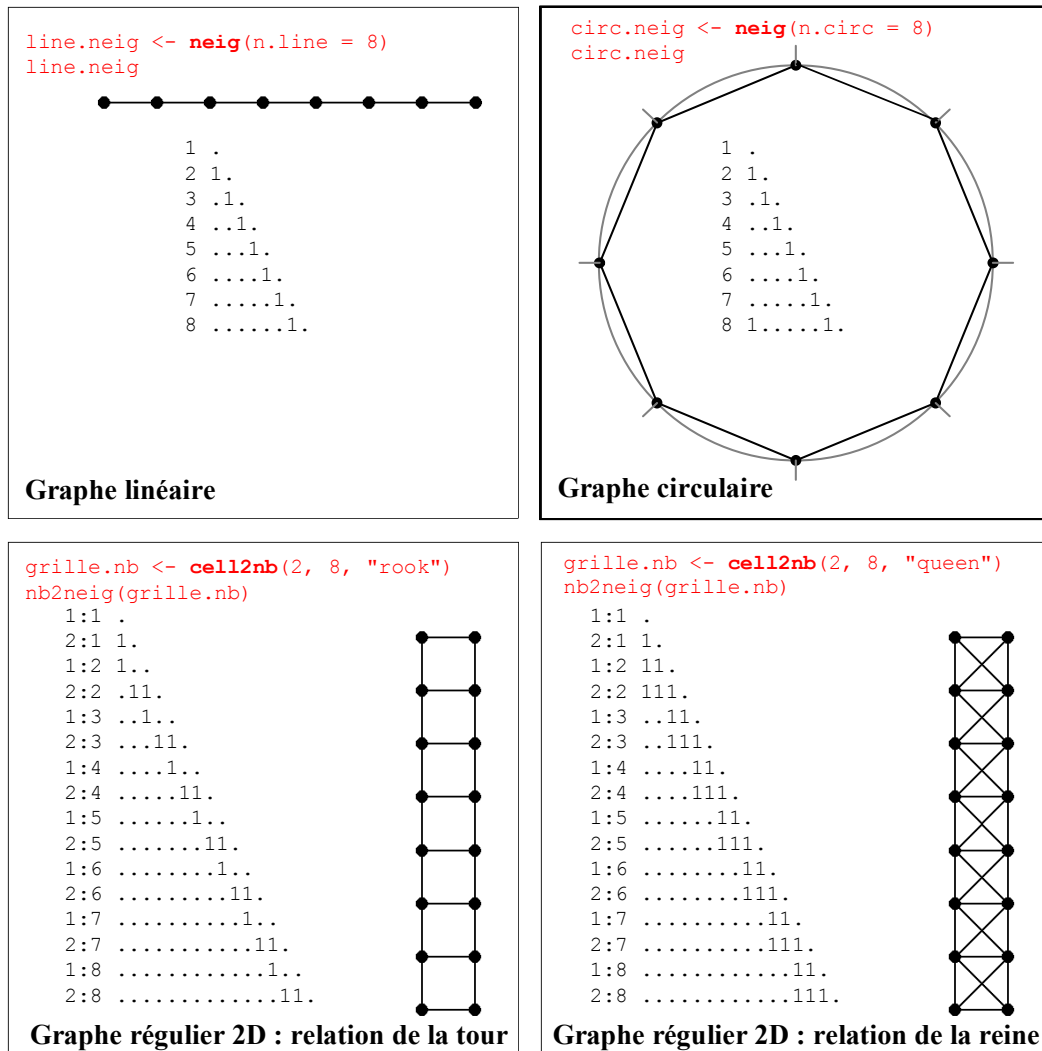


Figure 1.9 : graphes réguliers. En rouge, les instructions avec en gras, les fonctions permettant de définir le graphe régulier correspondant. Au centre de chaque image, les matrices du graphe de voisinage.

- On peut définir également des graphes variés à partir des coordonnées des unités ponctuelles. Le graphe de voisinage peut être notamment dérivé du diagramme de Voronoï (Upton & Fingleton, 1985), construction géométrique dans le plan correspondant à une partition de l'espace en polygones, chacun étant défini autour d'un point. Le graphe dual du diagramme de Voronoï est la triangulation de Delaunay. Il s'agit du graphe reliant les points générateurs des polygones contigus. La librairie `tripack` (code fortran de R.J. Renka, fonctions R de A. Gebhardt et contributions de S. Eglen et S. Zuyev) est entièrement dédiée à la triangulation des données spatiales. Les voisins au sens de la triangulation de Delaunay ne sont pas forcément les plus proches voisins d'un point au sens de la distance euclidienne qui les séparent, par contre ils lui sont contigus. Le graphe des plus proches voisins par la distance euclidienne n'est donc pas le même que celui

défini par la triangulation de Delaunay. De plus, le voisinage par les plus proches voisins, contrairement au voisinage induit par la triangulation de Delaunay, conduit à un nombre constant de voisins, ce qui est un avantage (pondération uniforme par unité ponctuelle) que l'on paye par la non symétrie (Pace & Zou, 2000). Le graphe induit par la triangulation de Delaunay est également différent du graphe définissant les voisins par un couple de distances (d_1 , d_2) : deux points sont voisins si et seulement si leur distance est supérieure à d_1 et inférieure à d_2 (Figure 1.10).

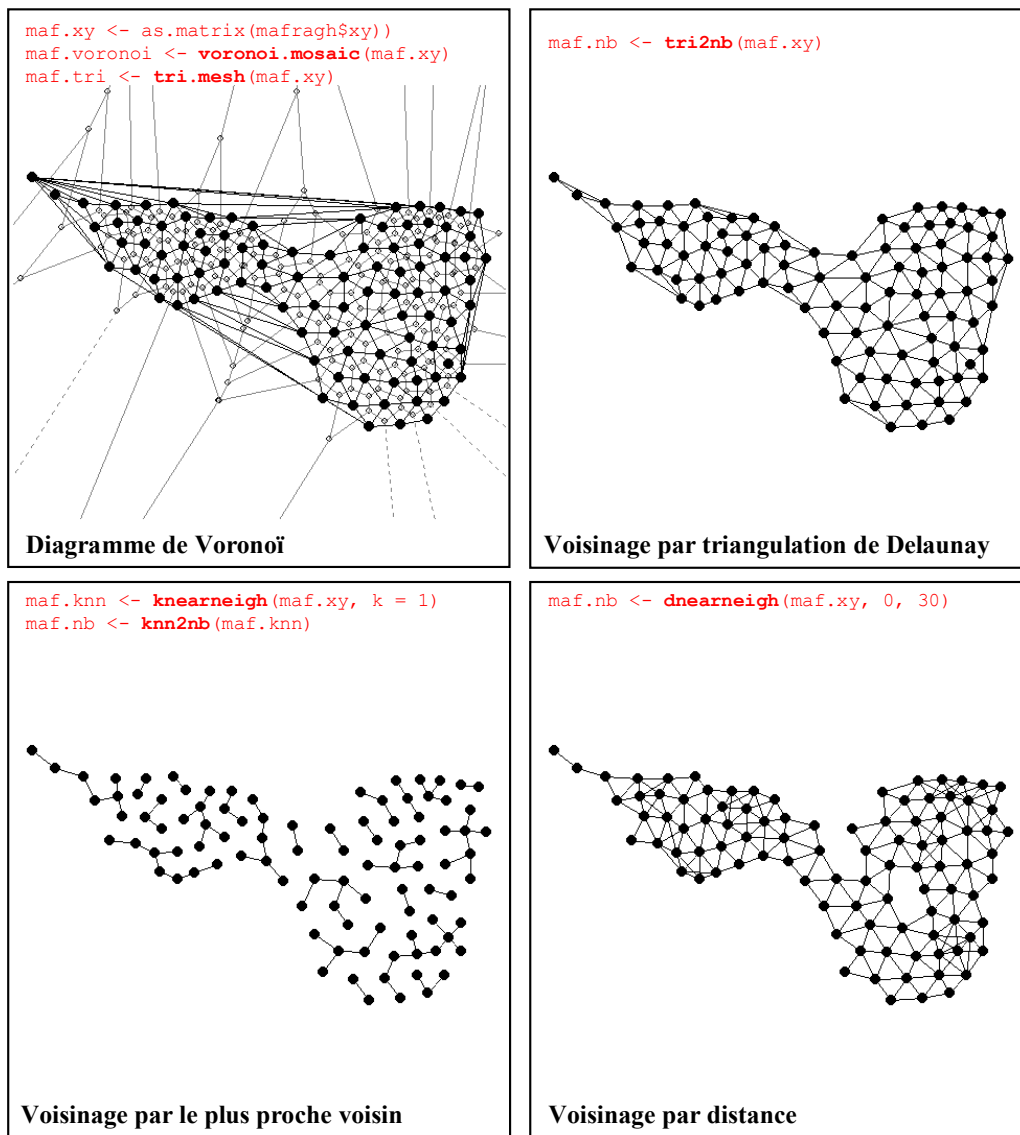


Figure 1.10 : diagramme de Voronoï et triangulation de Delaunay associés aux relevés floristiques du jeu de données mafragh (Annexe 1.13) (**en haut, à gauche**). Graphe de Voronoï : entre la triangulation de Delaunay et le graphe final il y a un ajustement manuel nécessaire réalisable grâce à la fonction `edit.nb(...)` (**en haut, à droite**). Graphe de voisinage par le plus proche voisin : un point a un seul plus proche voisin mais peut être le plus proche voisin de plusieurs autres (**en bas, à gauche**). Graphe de voisinage par distance : deux point sont voisins s'ils sont distants de 30 unités au plus (**en bas, à droite**). En rouge, les instructions avec en gras, les fonctions permettant de définir les graphes correspondants.

- Par ailleurs, le graphe de Delaunay comporte l'information de plusieurs autres sous graphes. Le graphe de Gabriel, sous-graphe du graphe de Voronoï est défini par : i et j sont voisins s'ils le sont au sens de la triangulation de Delaunay et si $d_{ij} \leq \min_k \left(\sqrt{d_{ik}^2 + d_{jk}^2} \right)$. Deux points sont donc connectés si aucun autre point ne se trouve à l'intérieur du cercle de diamètre défini par ces deux points (Gabriel & Sokal, 1969). De même, le graphe du voisinage relatif est un sous-graphe du graphe de Delaunay. Dans ce graphe, deux points i et j sont voisins s'ils le sont au sens de la triangulation de Delaunay et si $d_{ij} \leq \min_k \left(\max(d_{ik}, d_{jk}) \right)$. Enfin, le graphe de longueur minimale reliant l'ensemble des unités ponctuelles est également un sous graphe du graphe de Voronoï (Figure 1.11).

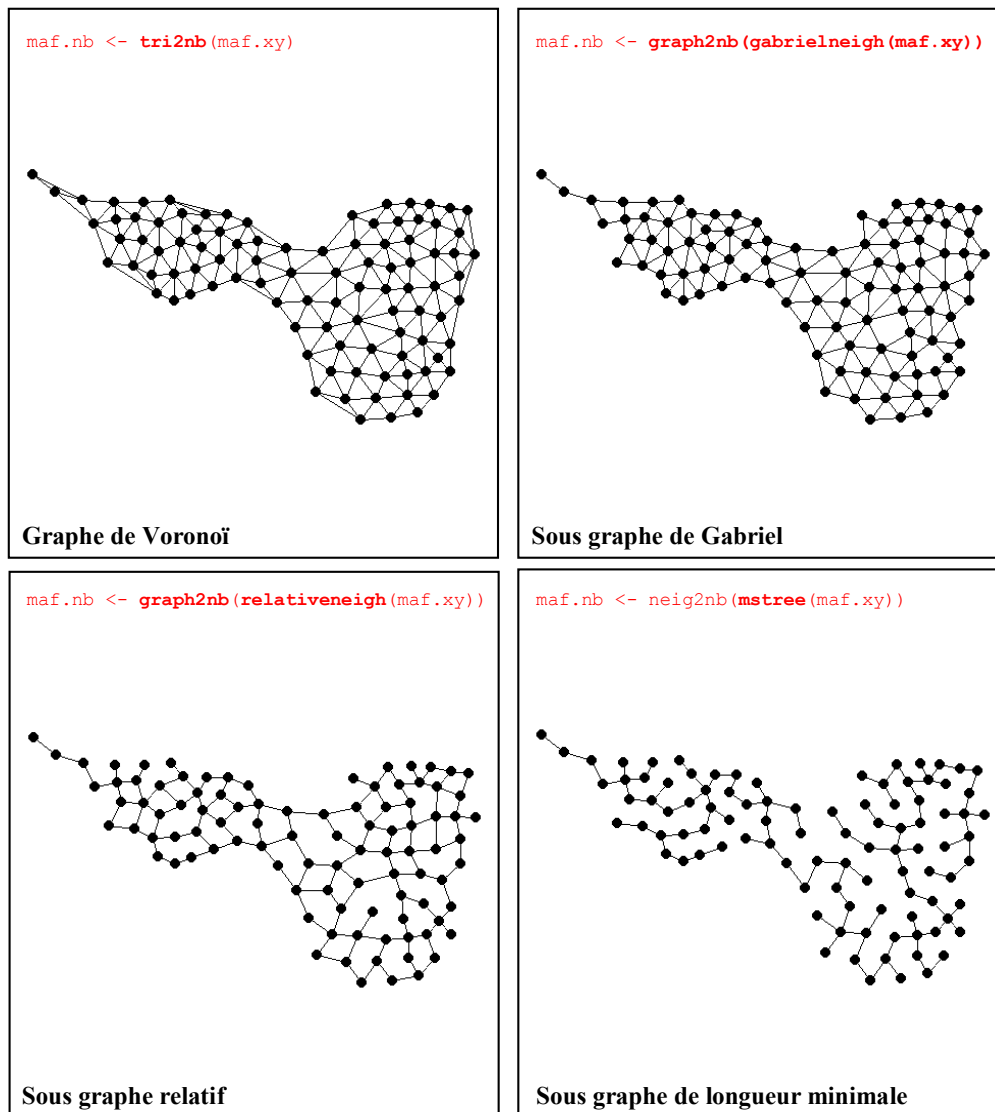


Figure 1.11 : Graphe de Voronoï et sous graphes associés aux relevés floristiques du jeu de données mafragh. En rouge, les instructions avec en gras, les fonctions permettant de définir les graphes correspondants.

- On peut finalement définir des graphes de voisinage à partir des unités surfaciques. Deux unités sont alors considérées comme voisines si elles partagent une frontière commune (Figure 1.12).

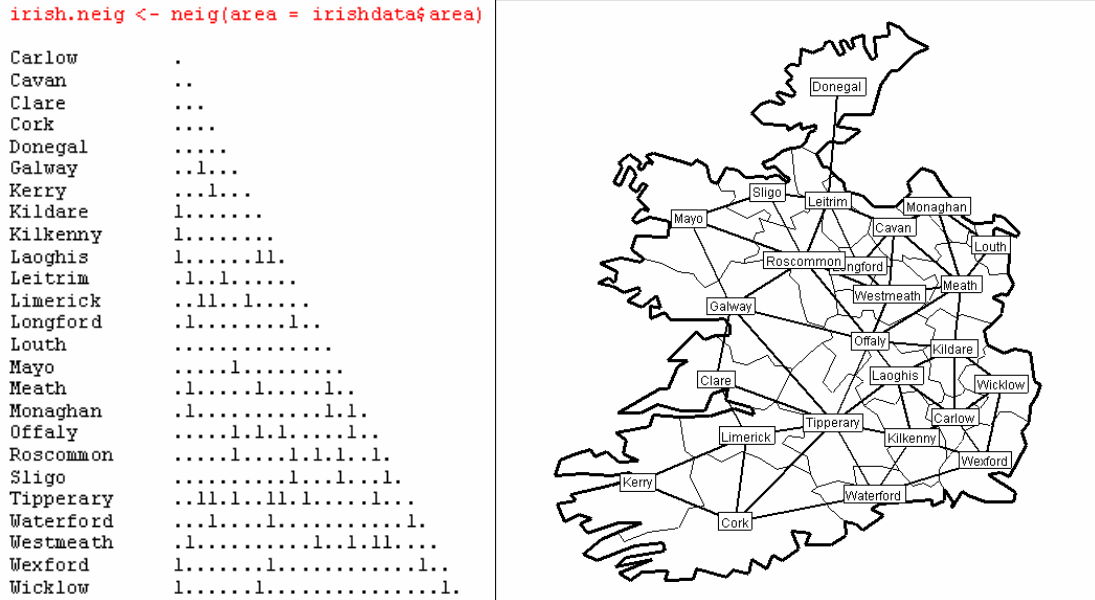


Figure 1.12 : représentation du graphe de voisinage (à droite) et de la matrice de voisinage binaire (à gauche) associés aux comtés d'Irlande (Annexe 1.11). Deux comtés sont voisins s'ils partagent une frontière commune.

Le graphe de voisinage constitue le matériel de base des ordinations sous contraintes spatiales. On est capable de le construire pour des supports spatiaux de nature diverse. Les graphes de la classe 'nb' sont orientés : i peut être voisin de j sans que j soit voisin de i . C'est vrai pour deux stations sur un cours d'eau : l'amont influence l'aval et non l'inverse. La matrice du graphe \mathbf{M} est définie par $m_{ij} = 1 \Leftrightarrow j$ est voisin de i . La plupart des auteurs de l'école de Lebart (Cox & Lewis, 1969) travaillant sur la variance locale n'envisagent que des graphes symétriques et ne manipulent pas de pondération de voisinage autre que directement induites par la relation binaire. Pour beaucoup d'autres, et on va voir pourquoi, le second élément de la prise en compte de l'espace est la pondération de voisinage. Un même graphe de la classe 'nb' peut donner plusieurs pondérations de la classe 'listw'. La matrice des pondérations de voisinage du graphe est notée de manière générale \mathbf{W} telles que $w_{ij} \geq 0$ et $w_{ij} = 0 \Leftrightarrow m_{ij} = 0$. On rentre alors dans la tradition des géographes et des économètres.

2.3. Pondérations de voisinage

Une pondération de voisinage est toujours associée à un graphe de voisinage. Ce qui est pondéré c'est le lien entre voisins. R. Bivand a représenté les principales options dans ses procédures. Dans un objet de la classe 'listw' on a d'abord une liste à n composantes qui sont des vecteurs donnant les numéros des voisins (on peut ou non tolérer des points sans voisin) puis une liste à n composantes qui sont des vecteurs donnant les poids des voisins. Le premier élément nous donne implicitement la matrice \mathbf{M} alors que le second correspond à la matrice \mathbf{W} . Il y a au moins deux manières principales de pondérer les voisinages (Cliff & Ord, 1973). Le plus simple est de laisser agir la fonction `nb2listw(...)`. Prenons comme exemple le graphe de voisinage associé aux comtés d'Irlande (Figure 1.12).

```
is.matrix(irish.neig) # irish.neig est un graphe de la classe neig
[1] TRUE
```

```
irish.neig[1:3,]
  [,1] [,2]
[1,]   3   6 # l'arête 1 relie le point 3 au point 6
[2,]   4   7 # l'arête 2 relie le point 4 au point 7
[3,]   1   8
. . .

dim(irish.neig)
[1] 54  2 # Il y a 54 arêtes dans ce graphe

attributes(irish.neig)
$dim
[1] 54  2

$degrees
  Carlow   Cavan   Clare   Cork   Donegal   Galway   Kerry   Kildare
      5       5       3       4       1       5       2       5
. . .

$call
neig(area = irishdata$area.utm)

$class
[1] "neig"
```

```
is.list(irish.nb) # irish.nb est un objet de la classe nb
[1] TRUE
```

```
irish.nb
$"1"
[1] 8 9 10 24 25 # la liste des voisins de 1

$"2"
[1] 11 13 16 17 23 # la liste des voisins de 2
. . .
$"25"
[1] 1 8 24

attributes(irish.nb)
$names
[1] "1" "2" "3" "4" "5" "6" "7" "8" "9" "10" "11" "12" "13" "14" "15"
[16] "16" "17" "18" "19" "20" "21" "22" "23" "24" "25"

$region.id
[1] "Carlow" "Cavan" "Clare" "Cork" "Donegal" "Galway"
```

```

. . .
[19] "Roscommon" "Sligo"      "Tipperary" "Waterford" "Westmeath" "Wexford"
[25] "Wicklow"

$gal
[1] FALSE

$call
neig2nb(neig = irish.neig)

$class
[1] "nb"

```

Les deux objets contiennent la même information dans des formats différents mais seul le second fournit des pondérations de voisinages de la classe ‘listw’ :

```
nb2listw(nb, glist = NULL, style = "W", zero.policy = FALSE)
```

```

pond.w <- nb2listw(irish.nb, style="W")
pond.b <- nb2listw(irish.nb, style="B")
pond.c <- nb2listw(irish.nb, style="C")
pond.u <- nb2listw(irish.nb, style="U")
pond.s <- nb2listw(irish.nb, style="S")
names(pond.w)
[1] "style"      "neighbours" "weights"

```

La fonction reprend le graphe et donne des poids aux arêtes. Il y a 5 options :

W *row standardised* : l'option W, **par défaut**, donne un poids égal à l'inverse du nombre de voisins. La matrice **W** est alors de somme unité par ligne et nous l'appellerons **L** (pour profils lignes) :

```

pond.w$weights[1] # 0.2 = 1/5
[[1]]
[1] 0.2 0.2 0.2 0.2 0.2

unlist(lapply(pond.w$weights, sum))
[1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

sum(unlist(pond.w$weights))
[1] 25

```

L'option W donne une *row standardized spatial weights matrix* comme dans Cliff and Ord (1973) ou Anselin and Hudak (1992) .

B *basic binary coding* : l'option B donne un poids unité à chaque couple de voisins, c'est-à-dire la matrice **M** de Lebart (1969) :

```

pond.b$weights[1] # Chaque arête du graphe a le même poids de voisinage
[[1]]
[1] 1 1 1 1 1

unique(unlist(pond.b$weights))
[1] 1

```

```

unlist(lapply(pond.b$weights, sum))
[1] 5 5 3 4 1 5 2 5 5 5 5 4 4 2 3 5 3 6 7 3 8 4 5 4 3

sum(unlist(pond.b$weights))
[1] 106

```

C *globally standardised* : l'option C donne le même poids unité à chaque couple de voisins, égal au nombre de points divisé par le nombre de couples de voisins (n/a) :

```

pond.c$weights[1]
[[1]]
[1] 0.2358 0.2358 0.2358 0.2358 0.2358

unique(unlist(pond.c$weights))
[1] 0.2358

sum(unlist(pond.c$weights))
[1] 25

```

Cette option donne n fois les (*doubly standardized spatial weights matrix* comme dans Wartenberg (1985) ou Anselin et al (2002). Nous écrirons ces matrices $n\mathbf{F}$ avec \mathbf{F} une distribution de fréquences bivariées, la somme de tous les éléments faisant l'unité.

U *globally standardised* : *U is equal to C divided by the number of neighbours (sums over all links to unity)*. C'est la précédente divisée par n donc \mathbf{F} :

```

pond.u$weights[1]
[[1]]
[1] 0.00926 0.00926 0.00926 0.00926 0.00926

unique(unlist(pond.u$weights))
[1] 0.00926

sum(unlist(pond.u$weights))
[1] 1

```

S *variance-stabilizing coding scheme* : l'option S est due à Tiefelsdorf et al. (1999). Dans ce schéma chaque ligne de \mathbf{M} est normalisée comme un vecteur pour la métrique canonique, donc divisée par la racine du nombre de voisins, puis divisée par la somme totale du résultat, ce qui donne une distribution de fréquences non symétriques, puis multipliée par n pour que la somme soit, comme pour les autres égale au nombre de points.

```

pond.s$weights[1]
[[1]]
[1] 0.2212 0.2212 0.2212 0.2212 0.2212

deg <- unlist(lapply(iris.nb, length))
deg
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
 5  5  3  4  1  5  2  5  5  5  5  4  4  2  3  5  3  6  7  3  8  4  5  4  3
(1/sqrt(5))/sum(sqrt(deg))*25
[1] 0.2212

sum(unlist(pond.s$weights))
[1] 25

```


Pour éviter les complications nous dirons que cette matrice est encore de type nF .

On peut de plus importer directement une liste de poids, comme celle des longueurs de frontières et transformer le résultat.

```
irish.list.w <- apply(irishdata$link.utm, 1, function(x) x[x!=0])
pond.ext.w <- nb2listw(irish.nb, glist = irish.list.w, style = "W")
pond.ext.b <- nb2listw(irish.nb, glist = irish.list.w, style = "B")
pond.ext.c <- nb2listw(irish.nb, glist = irish.list.w, style = "C")
pond.ext.u <- nb2listw(irish.nb, glist = irish.list.w, style = "U")
pond.ext.s <- nb2listw(irish.nb, glist = irish.list.w, style = "S")
```

W *row standardised* : l'option W passe la matrice de poids en distribution de fréquences par point, c'est-à-dire la matrice **L** de terme général $w_{ij}/w_{i\cdot}$:

```
pond.ext.w$weights[1]
[[1]]
  Kildare Kilkenny  Laoghis  Wexford  Wicklow
  0.1031  0.2528  0.1008  0.2445  0.2988

unlist(lapply(pond.ext.w$weights, sum))
[1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

sum(unlist(pond.ext.w$weights))
[1] 25
```

B *basic binary coding* : l'option B repasse en binaire la relation et on obtient strictement le même résultat avec le graphe de voisinage simple :

```
pond.ext.b$weights[1]
[[1]]
[1] 1 1 1 1 1

pond.b$weights[1] # l'introduction des poids ne modifie pas le résultat
[[1]]
[1] 1 1 1 1 1
```

C *globally standardised* : l'option C donne la distribution de fréquence multipliée par le nombre de points :

```
pond.ext.c$weights[[1]]
  Kildare Kilkenny  Laoghis  Wexford  Wicklow
  0.09632 0.23629 0.09422 0.22846 0.27928

unlist(lapply(pond.ext.c$weights, sum))
[1] 0.93457 1.17501 0.61166 1.29065 0.05327 1.49448 0.72411 0.99412 1.11880
[10] 1.14033 1.02783 1.17664 0.77922 0.33354 0.97816 1.26971 0.50156 1.53714
[19] 1.55220 0.88029 2.02735 0.88106 1.17693 0.60439 0.73698

sum(unlist(pond.ext.c$weights))
[1] 25
```

C'est typiquement nF .

U *globally standardised* : l'option U donne la distribution de fréquence non modifiée, typiquement **F** :

```
pond.ext.u$weights[[1]]
Kildare Kilkenny  Laoghis  Wexford  Wicklow
0.003832 0.009622 0.003324 0.007736 0.011813

sum(unlist(lapply(pond.ext.u$weights, sum)))
[1] 1
```

La transformation *S* est encore disponible. Pour le passage au multivarié, nous retiendrons les deux cas fondamentaux :

W normalisée par ligne (option W) : matrice de type **L**

W globalement normalisée (option U) : matrice de type **F**

$$l_{ij} = \frac{w_{ij}}{\sum_j w_{ij}} = \frac{w_{ij}}{w_{i\bullet}} \text{ soit } \mathbf{L} = [f_{j/i}] \Rightarrow \mathbf{L}\mathbf{1}_n = 1 \text{ et } f_{ij} = \frac{w_{ij}}{\sum_{i,j} w_{ij}} \text{ soit } \mathbf{F} = [f_{ij}] \Rightarrow \mathbf{1}'_n \mathbf{F} \mathbf{1}_n = 1$$

La question des poids de voisinage est une source d'ambiguïté remarquable dans les présentations des méthodes multivariées intégrant l'espace. On utilise en général le terme générique **W** pour parler de **F**, $n\mathbf{F}$, **M**, **L** ou $n\mathbf{L}$, ce qui ne simplifie pas les choses. Les dénominations sont tellement instables qu'il faut vérifier systématiquement ce qui s'est fait. Dans Cliff et Ord (1973), les auteurs proposent un système quelconque où le poids est une fonction des longueurs des frontières communes et des distances entre centres, ou encore l'inverse de la distance entre centres, mais aussi une matrice **W** de somme unité par lignes et non symétrique pour faire en sorte que **Wz** calcule les moyennes des voisins d'une unité statistique. Cet aspect sans pratique canonique, ou pour le moins à deux options dominantes, ouvre la porte à toutes les manipulations arbitraires. Ceci se retrouve dans l'usage qui est fait de l'indice de Moran dans l'étude de la relation entre un trait biologique et une phylogénie ((Gittleman & Kot, 1990), nous y reviendrons dans le troisième chapitre). Cependant, deux pratiques distinctes sont utilisées constamment par la plupart des auteurs, la normalisation par ligne et la double normalisation.

Pour les tenants de l'autocorrélation spatiale, et donc de l'école fondée par Moran la **normalisation par ligne** est fréquente. Pour le voisinage par distance maximum, la pondération de voisinage définie par Pace et Barry (1997) est $d_{ij} \leq d_{\max} \Rightarrow w_{ij} = 1$ où d_{\max} est la distance maximum d'influence fixée. Comme indiqué par les auteurs cette pondération est

ensuite normalisée par ligne et donc du type **L**. La relation de voisinage des m plus proches voisins définie par Pace et al (Abramovich et al., 2003) s'écrit : $0 < d_{ij} < d_i^m \Rightarrow w_{ij} = \frac{1}{m}$ où d_i^m est la distance de i à son m -ième plus proche voisin. Cette pondération de voisinage est non symétrique mais « *row-stochastic* », c'est-à-dire de somme unité par lignes, et donc du type **L**. Pour Bavaud (1998) la définition d'une pondération de voisinage est très précise. Soit $S = \{1, \dots, n\}$ un ensemble de points. Une matrice de pondération de voisinage est une matrice **W** à n lignes et n colonnes telle que a) $w_{jk} \geq 0$ b) $\sum_{k=1}^n w_{jk} = 1 \quad \forall j \in S$. Les poids w_{ij} ne sont pas forcément nuls. On utilise les termes équivalents de « *contiguity, connectivity, adjacency, association* » pour ce type de matrices. La matrice **W** n'est pas forcément symétrique car elle donne une indication sur l'influence potentielle de i sur j . Il s'agit alors de la matrice de transition d'un processus de Markov. Elle est de type **L**. Certes, une proposition remarquable est faite encore par Pace et LeSage (2002) pour combiner les relations aux k premiers voisins pour obtenir une matrice de poids bi-stochastique dites *the doubly spatial model* (distributions de fréquences par lignes et par colonnes), ce qui est très judicieux, mais encore loin d'être devenu un standard. Les matrices de type **L**, stochastiques par ligne, ou markoviennes, décrivent la répartition de l'influence du point i sur l'ensemble des autres par une distribution de fréquence. C'est l'option par défaut dans **spdep** et ce n'est sûrement pas un hasard.

Pour les praticiens de la variance locale et de l'école, fondée par Geary puis reprise par Lebart, c'est au contraire les matrices de type **F** qui s'imposent. La double normalisation est simplement la division par la somme de toutes ses valeurs qui donne une matrice de poids de voisinage. Le poids de voisinage n'a guère de sens pour le couple (i, i) et pour éviter les sommes pour $i = j$ on simplifie en posant $w_{ii} = 0$. Une matrice de poids de voisinage est donc une matrice carrée, symétrique, à diagonale nulle et somme unité. A partir de maintenant nous utiliserons **L** ou **F** pour désigner des pondérations de voisinage et **W** quand les deux cas sont concernés. L'usage des matrices **F** en autocorrélation est aussi largement répandu. P. Aubry (2000) qui fait une analyse bibliographique hors du commun utilise directement sans notion de voisinage les pondérations :

$$w_{ij} = 1 - \frac{d_{ij}}{\max(d_{ij})} \text{ et } w_{ij} = \frac{1}{d_{ij}}$$

La question des poids de voisinage n'est donc pas fixée et ne le sera sans doute jamais, une solution universelle pour tous les problèmes et tous les types de données n'ayant pas de sens. C'est dans ce contexte que l'on se pose la question de l'analyse multivariée en introduisant la notion de voisinage comme contrainte. L'intérêt est d'aborder des tableaux massivement multivariés comme le sont par exemple des relevés de faune ou de flore. Les méthodes existantes telle que l'ACP de Wartenberg (1985), s'appuyant sur les éléments univariés de base, on propose, avant de s'attaquer au problème multivarié, d'étudier les deux principaux indices de la structure spatiale.

3. INDICES UNIVARIÉS DE LA STRUCTURE SPATIALE

Les indices de Geary (1954) et de Moran (1948; 1950) sont à la base de deux écoles de statistiques spatiales. On reprendra directement la présentation de Cliff et Ord (1973). On rappelle que n est le nombre d'unités statistiques et \mathbf{W} est la matrice des poids de voisinage.

x_i est la valeur de l'unité statistique i et $z_i = x_i - \bar{x}$ avec $\bar{x} = \frac{1}{n} \sum_i x_i$. La notation classique est

:

$$\sum_{(2)} y_{ij} = \sum_{\substack{1 \leq i \leq n \\ 1 \leq j \leq n \\ i \neq j}} y_{ij}$$

3.1. L'indice I de Moran (1948, 1950)

Le I de Moran est en général défini par:

$$I = \frac{n \sum_{(2)} w_{ij} z_i z_j}{\sum_{(2)} w_{ij} \sum_{i=1}^n z_i^2}$$

ce qui désigne quelquefois (définition de \mathbf{F}) :

$$I = \frac{\mathbf{z}' \mathbf{F} \mathbf{z}}{\sum_{i=1}^n z_i^2 / n}$$

mais le plus souvent (les sommes par lignes de \mathbf{L} sont égales à 1 et la somme vaut n) :

$$I = \frac{\mathbf{z}' \mathbf{L} \mathbf{z} / n}{\sum_{i=1}^n z_i^2 / n}$$

L'indice de Moran est en général utilisé dans une des trois possibilités :

- graphe de voisinage binaire, non orienté, symétrique de matrice d'incidence \mathbf{M} avec m arêtes, soit $2m$ couples de voisins ou encore $\mathbf{1}'_n \mathbf{M} \mathbf{1}_n = 2m$:

$$I = \frac{1}{2m} \frac{\mathbf{z}' \mathbf{M} \mathbf{z}}{\sum_{i=1}^n z_i^2 / n} = \frac{\mathbf{z}' \mathbf{F} \mathbf{z}}{\sum_{i=1}^n z_i^2 / n}$$

- pondération de voisinage binaire, symétrique de matrice \mathbf{F} après normalisation globale :

$$I = \frac{\mathbf{z}' \mathbf{F} \mathbf{z}}{\sum_{i=1}^n z_i^2 / n}$$

- pondération de voisinage markovienne, normalisée par lignes, de matrice \mathbf{L} :

$$I = \frac{1}{n} \frac{\mathbf{z}' \mathbf{L} \mathbf{z}}{\sum_{i=1}^n z_i^2 / n} = \frac{1}{n} \frac{\sum_{i=1}^n z_i z_{v(i)}}{\sum_{i=1}^n z_i^2 / n} = \frac{\langle \mathbf{z} | \mathbf{L} \mathbf{z} \rangle_{\frac{1}{n}}}{\sum_{i=1}^n z_i^2 / n}$$

$z_{v(i)}$ est la moyenne des valeurs de la variable calculée sur les points voisins avec les poids relatifs des voisins. On appelle cette quantité un coefficient d'autocorrélation bien que ce ne soit pas un coefficient de corrélation (il faudrait que \mathbf{z} et $\mathbf{L} \mathbf{z}$ soient normées, ce qui est vrai pour la première mais pas pour la seconde) ni même une covariance (il faudrait que \mathbf{z} et $\mathbf{L} \mathbf{z}$ soit centrées, ce qui est vrai pour la première mais pas pour la seconde). C'est simplement le produit scalaire entre la variable mesurée et la variable obtenue par l'opération \mathbf{L} (moyenne sur les voisins).

Dans cette dernière optique, la fonction importante est `lag.listw(...)` : elle calcule pour un vecteur \mathbf{x} de longueur n et de composantes x_i le vecteur de composantes

$$y_i = \sum_{j \text{ voisin de } i} w_{ij} x_j \text{ ou encore } \mathbf{y} = \mathbf{L} \mathbf{x} \text{ dit } \textit{lag vector}.$$

```
print(lag.listw(pond.ext.w, rep(1, 25)))
[1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

Dans l'option W, le *lag vector* est simplement le vecteur des moyennes des valeurs prises par les voisins. Pour une variable constante, on trouve la même variable.

```
print(lag.listw(pond.ext.b, rep(1, 25)))
[1] 5 5 3 4 1 5 2 5 5 5 5 4 4 2 3 6 3 7 7 3 8 4 5 4 3
```

Dans l'option B, on trouve le nombre de voisins.

```
print(lag.listw(pond.ext.u, rep(1, 25)))
[1] 0.037383 0.047000 0.024466 0.051626 0.002131 0.059779 0.028964 0.039765
[9] 0.044752 0.045613 0.041113 0.047065 0.031169 0.013342 0.039127 0.050788
[17] 0.020062 0.061485 0.062088 0.035212 0.081094 0.035242 0.047077 0.024175
[25] 0.029479
```

Dans l'option U, on a directement les poids de voisinage.

```
print(lag.listw(pond.ext.c,rep(1,25)))
[1] 0.93457 1.17501 0.61166 1.29065 0.05327 1.49448 0.72411 0.99412 1.11880
[10] 1.14033 1.02783 1.17664 0.77922 0.33354 0.97816 1.26971 0.50156 1.53714
[19] 1.55220 0.88029 2.02735 0.88106 1.17693 0.60439 0.73698
```

Dans l'option C, on trouve les moyennes par voisin déformées par le rapport du poids de voisinage sur le poids uniforme. Nous n'utiliserons par la suite que l'option W pour laquelle l'opération à un sens ordinaire en analyse des données.

De plus, à cette définition de l'indice de Moran est associé un graphe canonique appelé le Moran scatterplot (Anselin, 1996). Il s'agit de la représentation graphique du « lag-vector » $y = Lx$ en fonction de x (Figure 1.13). La pente de la droite de régression de y en fonction de x reflète alors l'autocorrelation spatiale, et la position des unités statistiques sur le plan nous indique si l'on a affaire à une regroupement spatial (la valeur en un point et les valeurs des voisins se ressemblent : les points sont tous situés dans les quadrants 2-4), à une aberration spatiale (la valeur en un point et les valeurs des voisins s'opposent : les points sont tous situés dans les quadrants 1-3) ou à une absence de structure (pas de lien systématique entre la valeur en un point et les valeurs des voisins : distribution aléatoire des points sur l'ensemble du plan).

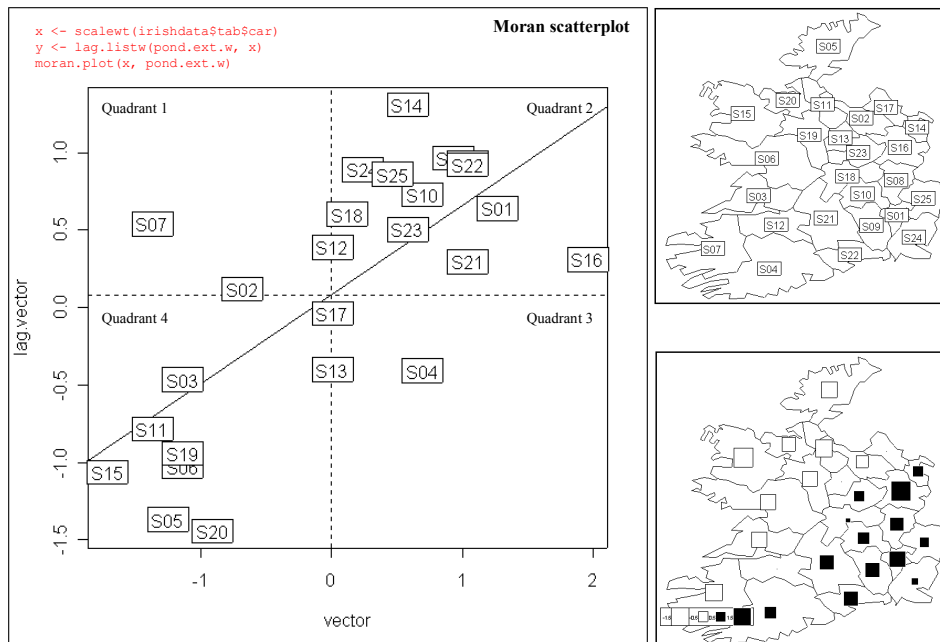


Figure 1.13 : Scatterplot de Moran. En abscisse les valeurs d'une variable x : ici il s'agit de la variable **car** du jeu de donnée irishdata (Annexe 1.11) qui a été préalablement centrée puis normée. Elle est représentée sur le fond de carte en bas à droite. En ordonnée, la variable y correspondant à la moyenne des valeurs des voisins. La droite est l'estimation du modèle $y = ax + b$. La pente reflète l'autocorrelation. Les deux droites en pointillés passent par les moyennes. La position des unités statistiques sur le plan, ici les comtés d'Irlande, est informative : quadrants 1-3=aberration spatiale, quadrants 2-4=regroupement spatiale.

3.2. Le coefficient de contiguïté c de Geary (1954)

Le c de Geary vaut :

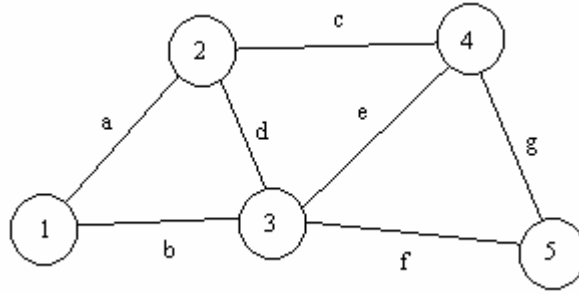
$$c = \frac{\sum_{(2)} w_{ij} (x_i - x_j)^2}{2 \sum_{(2)} w_{ij} \sum_{i=1}^n z_i^2 / (n-1)}$$

qui semble plutôt utilisé comme :

$$c = \frac{\sum_{(2)} f_{ij} (x_i - x_j)^2}{2 \sum_{i=1}^n z_i^2 / (n-1)} \text{ ou } c = \frac{\frac{1}{2m} \sum_{(2)} m_{ij} (x_i - x_j)^2}{2 \sum_{i=1}^n z_i^2 / (n-1)}$$

Cette fois-ci, est introduite la variance en $1/(n-1)$, contrairement à l'indice de Moran qui utilise la variance en $1/n$. C'est une différence mineure mais les deux indices introduisent deux définitions principales et deux usages de la matrice \mathbf{W} , dont on peut se demander s'ils ont la même signification. Notons que, dans tous les cas, on retrouve la division par $\sum_{(2)} w_{ij}$ et $\sum_{i=1}^n z_i^2 / n$ ou l'équivalent en $n-1$ et que le centrage à pondération uniforme est préalable dans l'indice de Moran et sans effet dans l'indice de Geary puisque $z_i - z_j = x_i - \bar{x} - x_j + \bar{x} = x_i - x_j$. La signification de ces indices ne pose pas de problème majeur. Ils ont été abondamment commentés. Si l'on fait l'impasse sur le $1/n$ ou $1/(n-1)$, les deux indices utilisent une variable \mathbf{z} sous sa forme centrée (moyenne nulle). La présence du carré de la différence qui ne distingue pas les couples (i, j) et (j, i) fait que l'indice de Geary n'a de sens que pour des matrices \mathbf{W} symétriques. Dans ce cas, les deux expressions jumelles $\sum_{(2)} w_{ij} z_i z_j$ (Moran) et $\sum_{(2)} w_{ij} (x_i - x_j)^2$ (Geary) peuvent être intimement liées (voir paragraphe suivant).

Pour redéfinir la famille des indices de Geary de manière plus efficace, on utilise la remarque fondamentale dans Banet et Lebart (1984). Soit un graphe de voisinage entre n points comportant m arêtes.



Soit \mathbf{O} la matrice à m lignes et n colonnes croisant les arêtes et les sommets. Pour l'arête i qui relie les sommets k et l avec $k < l$ on a $\mathbf{O}_{ik} = 1$, $\mathbf{O}_{il} = -1$ et $\mathbf{O}_{ij} = 0$ ailleurs. L'écriture est unique dès que la numérotation des sommets est donnée. Soit \mathbf{M} la matrice de voisinage (n lignes et n colonnes) et \mathbf{N} la matrice diagonale des degrés des sommets (nombre de voisins). Dans l'exemple :

$$\mathbf{O} = \begin{matrix} a \\ b \\ c \\ d \\ e \\ f \\ g \end{matrix} \begin{bmatrix} 1 & -1 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 \\ 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 1 & -1 \end{bmatrix} \quad \mathbf{M} = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{bmatrix} \quad \mathbf{N} = \begin{bmatrix} 2 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 \\ 0 & 0 & 4 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 2 \end{bmatrix}$$

On a :

$$\mathbf{O}'\mathbf{O} = \mathbf{N} - \mathbf{M}$$

$\mathbf{O}'\mathbf{O}$ est une matrice symétrique et non négative ($\mathbf{x}'\mathbf{O}'\mathbf{O}\mathbf{x} \geq 0$). Les poids de voisinage des points sont sur la diagonale de $\mathbf{P} = \frac{1}{2m}\mathbf{N}$ (les arêtes sont comptées deux fois) et les poids de voisinage des arêtes sont dans $\mathbf{F} = \frac{1}{2m}\mathbf{M}$:

$$\sum_{(2)} f_{ij} (x_i - x_j)^2 = 2\mathbf{x}'(\mathbf{P} - \mathbf{F})\mathbf{x} = 2\mathbf{z}'(\mathbf{P} - \mathbf{F})\mathbf{z}$$

D'où :

$$c = \frac{\mathbf{z}'(\mathbf{P} - \mathbf{F})\mathbf{z}}{\sum_{i=1}^n z_i^2 / (n-1)}$$

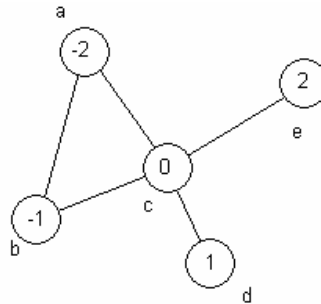
Ces propriétés restent vraies pour une matrice de poids de voisinage \mathbf{W} quelconque. On peut simplifier en normalisant a priori. Une modification mineure donne alors :

$$c^* = \mathbf{y}^t (\mathbf{P} - \mathbf{F}) \mathbf{y} \text{ avec } \mathbf{y} = \frac{\mathbf{z}}{\sqrt{\sum_{i=1}^n z_i^2 / (n-1)}}$$

On a par ailleurs $I^* = \mathbf{y}^t \mathbf{F} \mathbf{y}$. Ces relations extrêmement simples cachent en fait de nombreux problèmes qui ont beaucoup nui à leur usage effectif.

3.3. Quand les deux écoles se rejoignent ...

Pour comprendre la signification des deux indices, une réécriture de la notion de variance est indispensable. Elle a été faite par Lebart (1969) et le procédé a été utilisé indépendamment par Light & Margolin (1971) dans un autre problème. Soit un exemple numérique très simple comportant 5 observations a, b, c, d et e. Supposons la relation de voisinage suivante :



Dans les cercles on trouve la valeur de la variable en chacun des points. En supposant une pondération uniforme des 5 mesures, la moyenne vaut $m = 0$ et la variance vaut

$$\frac{(-2)^2 + (-1)^2 + (0)^2 + (1)^2 + (2)^2}{5} = 2$$

En général pour n observations $x_1, \dots, x_i \dots x_n$ de poids $p_1, \dots, p_i \dots p_n$ la moyenne et la variance sont définies par :

$$\bar{x} = \sum_{i=1}^n p_i x_i \text{ et } var(\mathbf{x}) = \sum_{i=1}^n p_i (x_i - \bar{x})^2 = \sum_{i=1}^n p_i z_i^2$$

Cette même variance peut se concevoir comme une fonction de toutes les différences entre les n mesures prises deux à deux.

	a	b	c	d	e
a	0	-1	-2	-3	-4
b	1	0	-1	-2	-3
c	2	1	0	-1	-2
d	3	2	1	0	-1
e	4	3	2	1	0

La moyenne (sur les 25 couples) des carrés de toutes les différences deux à deux vaut $100/25 = 4$ soit deux fois la variance. En général :

$$\text{var}(\mathbf{x}) = \frac{1}{2} \sum_{(2)} p_i p_j (x_i - x_j)^2$$

On retiendra la relation fondamentale :

$$\sum_{(2)} p_i p_j (x_i - x_j)^2 = 2 \sum_{i=1}^n p_i (x_i - \bar{x})^2$$

La variance à pondération quelconque est *la moitié* de la moyenne des carrés des différences élémentaires. Pour une vraie pondération de voisinage, l'indice de Geary mesure donc la variabilité locale et l'indice de Moran mesure la covariance locale (ou autocorrélation). Ces deux approches sont presque complémentaires sans l'être tout à fait. En effet, l'indice de Geary, contrairement à l'indice de Moran, semble supprimer toute notion de moyenne. En outre il est, comme rapport de deux sommes de carrés, toujours positif. La moyenne de la variable, en revanche, intervient fortement dans I . Or la moyenne intervient dans la définition ordinaire de la variance. En effet, si on cherche le nombre α qui minimise :

$$\frac{1}{n} \sum_{i=1}^n (x_i - \alpha)^2$$

on trouve $\alpha = \bar{x}$ et le minimum atteint est la variance. Le numérateur et le dénominateur de l'indice de Moran n'ont donc pas un statut aussi voisin que le numérateur et le dénominateur de celui de l'indice de Geary. Si on cherche le nombre α qui minimise :

$$\frac{1}{m} \sum_{(2)} m_{ij} (x_i - \alpha)(x_j - \alpha)$$

on ne trouve pas $\alpha = \bar{x}$. En effet, un calcul simple sur un polynôme du second degré conduit à :

$$\alpha = \frac{\mathbf{x}^t \mathbf{M} \mathbf{1}_n}{\mathbf{1}_n^t \mathbf{M} \mathbf{1}_n} = \frac{1}{m} \sum_{i=1}^n m_i x_i = m_v(\mathbf{x})$$

où m_v désigne la moyenne de voisinage de la variable \mathbf{x} calculée avec un poids d'une observation i proportionnel à son nombre de voisins. C'est précisément l'écart entre la

moyenne ordinaire et la moyenne de voisinage qui sépare les deux approches. En effet, si on réécrit l'indice de Moran en utilisant la moyenne de voisinage :

$$I^* = \frac{\frac{1}{2m} \sum_{(2)} m_{ij} (x_i - m_v(\mathbf{x}))(x_j - m_v(\mathbf{x}))}{\frac{1}{n} \sum_{i=1}^n (x_i - m_v(\mathbf{x}))} = \mathbf{y}^t \mathbf{F} \mathbf{y} \text{ avec } \mathbf{y} = \frac{x_i - m_v(\mathbf{x})}{\frac{1}{n} \sum_{i=1}^n (x_i - m_v(\mathbf{x}))}$$

et si on réécrit l'indice de Geary sous la forme :

$$c^* = \frac{\frac{1}{2m} \sum_{(2)} m_{ij} (x_i - x_j)^2}{\frac{1}{n} \sum_{i=1}^n (x_i - m_v(\mathbf{x}))} = \mathbf{y}^t (\mathbf{P} - \mathbf{F}) \mathbf{y}$$

on a alors simplement $I^* + c^* = \mathbf{y}^t \mathbf{F} \mathbf{y} + \mathbf{y}^t (\mathbf{P} - \mathbf{F}) \mathbf{y} = \mathbf{y}^t \mathbf{P} \mathbf{y} = 1$. Cette décomposition est curieuse car seuls deux termes sur les trois sont toujours positifs. Elle est abondamment commentée par Durand et al. (1999) et Ghertsos et al. (2001). Pour un processus "lisse" donc fortement cartographiable, la variance locale est faible mais positive et la covariance locale est positive et forte. Pour un processus à forte variation entre voisins, la variance locale est plus forte que la variance d'ensemble et l'autocovariance est négative. Les deux statistiques disent la même chose tandis que leur somme est constante. On pourrait croire la question résolue mais ce point de vue cache un gros inconvénient. Pour une approche inférentielle, la pondération non uniforme qui intervient dans le calcul de la moyenne et de la variance fait que cette moyenne et cette variance ne sont pas des invariants dans l'espace des $n!$ permutations définies sur les données. La pondération de voisinage peut cependant être uniforme dans quelques cas. Elle est uniforme si l'on travaille avec la pondération de voisinage \mathbf{L} mais dans ce cas \mathbf{W} n'est plus symétrique et l'indice de Geary perd son sens. De même, pour le voisinage défini par le plus proche voisin, le nombre de voisins étant identique, la pondération est uniforme mais l'on perd à nouveau la symétrie. L'idéal est d'avoir une matrice symétrique dont les pondérations lignes et colonnes sont uniformes. Ces matrices, appelés doublement stochastiques (Pace & LeSage, 2003), ont par ailleurs des propriétés canoniques intéressantes. On peut les obtenir par des transformations particulières, telles que celles définies par Pace et Le Sage (2003). De plus, comme on le verra dans le troisième chapitre, certains graphes comme les phylogénies introduisent des matrices doublement

stochastiques qui leur sont canoniquement associées. On retrouvera en multivarié les quatre cas fondamentaux :

I classiquement défini par :
$$I = \frac{n \sum_{(2)} w_{ij} z_i z_j}{\sum_{(2)} w_{ij} \sum_{i=1}^n z_i^2} = \mathbf{y}' \mathbf{W} \mathbf{y}$$
 avec \mathbf{W} plutôt sous la forme \mathbf{L}

c classiquement défini par :
$$c = \frac{\sum_{(2)} f_{ij} (x_i - x_j)^2}{2 \sum_{i=1}^n z_i^2 / n} = \mathbf{y}' (\mathbf{P} - \mathbf{F}) \mathbf{y}$$

I et c sont liés par la relation $I^* + c^* = \mathbf{y}' \mathbf{F} \mathbf{y} + \mathbf{y}' (\mathbf{P} - \mathbf{F}) \mathbf{y} = \mathbf{y}' \mathbf{P} \mathbf{y} = 1$ lorsque \mathbf{x} est centré et normé pour la pondération de voisinage

En particulier, si \mathbf{W} est une matrice bistochastique, la relation reste vraie pour la pondération uniforme classiquement utilisée : $I + c = \mathbf{y}' \mathbf{F} \mathbf{y} + \mathbf{y}' (\mathbf{Id}_n - \mathbf{F}) \mathbf{y} = \mathbf{y}' \mathbf{y} = 1$

3.4. Tests contre l'absence de structure spatiale

L'absence de structure spatiale est décrite par l'hypothèse nulle « z_i est la réalisation d'une variable aléatoire gaussienne de loi $N(\mu, \sigma^2)$ » (modèle gaussien), ou par l'hypothèse nulle « les observations sont distribuées dans l'espace par tirage au hasard dans l'espace des $n!$ permutations des n premiers entiers » (modèle non paramétrique). Dans ce dernier cas, on peut soit utiliser une approximation de la loi de la statistique basée sur les moments ou générer des tirages aléatoires (test de randomisation). La librairie `spdep` propose des fonctions mettant en œuvre l'ensemble de ces tests. La documentation des fonctions est explicite.

Pour faire les tests de Moran (respectivement Geary) dans le modèle gaussien ou le modèle non paramétrique de l'équiprobabilité des $n!$ permutations des données, en utilisant une approximation de la loi de la statistique basée sur les moments, il faut utiliser la fonction `moran.test(...)` (respectivement `geary.test(...)`) :

```

R 'moran.test' help
moran.test                package:spdep                R Documentation
Moran's I test for spatial autocorrelation

Description:

Moran's test for spatial autocorrelation using a spatial weights
matrix in weights list form. The assumptions underlying the test
are sensitive to the form of the graph of neighbour relationships
and other factors, and results may be checked against those of
'moran.mc' permutations.

Usage:

moran.test(x, listw, randomisation=TRUE, zero.policy=FALSE,
           alternative="greater", rank = FALSE, na.action=na.fail, spChk=NULL)

```

L'option 'randomisation' définit avec quel type d'hypothèse nulle on travaille :

```

randomisation: variance of I calculated under the assumption of
randomisation, if FALSE normality

```

```
unclass(moran.test(irishdata$tab$car, pond.ext.w))
```

```
$statistic
```

```
Moran I statistic standard deviate
                                4.21
```

```
$p.value
```

```
  Kildare
1.277e-05
```

```
$estimate
```

```
Moran I statistic      Expectation      Variance
0.57665                -0.04167         0.02157
```

```
$alternative
```

```
[1] "greater"
```

```
$method
```

```
[1] "Moran's I test under randomisation"
```

```
$data.name
```

```
[1] "irishdata$tab$car \nweights: pond.ext.w \n"
```

```
unclass(moran.test(irishdata$tab$car, pond.ext.w, randomisation = FALSE))
```

```
$statistic
```

```
Moran I statistic standard deviate
                                4.301
```

```
$p.value
```

```
  Kildare
8.495e-06
```

```
$estimate
```

```
Moran I statistic      Expectation      Variance
0.57665                -0.04167         0.02067
```

```
$alternative
```

```
[1] "greater"
```

```
$method
```

```
[1] "Moran's I test under normality"
```

```
$data.name
```

```
[1] "irishdata$tab$car \nweights: pond.ext.w \n"
```

Pour faire les tests de Moran (respectivement Geary) dans le modèle non paramétrique de l'équiprobabilité des $n!$ permutations des données, en générant des tirages aléatoires, on utilise la fonction `moran.mc(...)` (respectivement `geary.mc(...)`) :

```

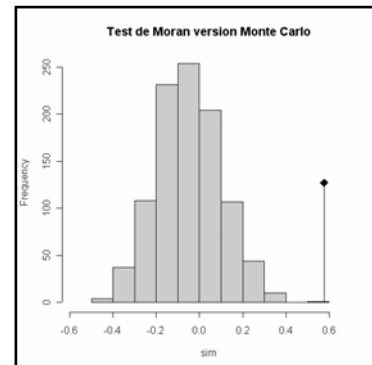
R 'moran.mc' help
moran.mc                package:spdep                R Documentation
Permutation test for Moran's I statistic
Description:
  A permutation test for Moran's I statistic calculated by using
  nsim random permutations of x for the given spatial weighting
  scheme, to establish the rank of the observed statistic in
  relation to the nsim simulated values. The examples show how
  'boot(sim="permutation")' can replicate this function (thanks to
  Virgilio Gómez Rubio and the DCluster package).
Usage:
  moran.mc(x, listw, nsim, zero.policy=FALSE, alternative="greater", na.acti$

```

```

test.moran.mc <- moran.mc(irisdata$tab$car, pond.ext.w, nsim = 999)
unclass(test.moran.mc)
$statistic
statistic
  0.5766
$parameter
observed rank
      1000
$parameter
observed rank
      1000
$p.value
[1] 0.001
$alternative
[1] "greater"
$method
[1] "Monte-Carlo simulation of Moran's I"
$data.name
[1] "irisdata$tab$car \nweights: pond.ext.w \nnumber of simulations + 1: 1000 \n"
$res
[1] -0.0799985 ...
[997] -0.1725840 -0.0300413 -0.1187202  0.5766499

```



```

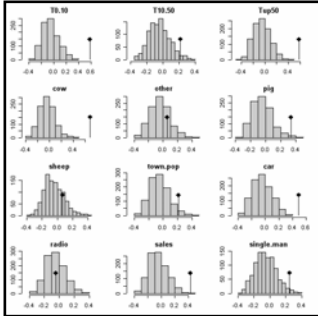
plot(as.randtest(test.moran.mc$res, test.moran.mc$statistic), main = "Test de Moran
version Monte Carlo")

```

Les tests de permutations sont les plus robustes : cette dernière technique l'emporte sur les autres et limite les discussions byzantines. De plus, on a beaucoup discuté de la puissance de ces tests. Globalement le I de Moran l'emporte sur le c de Geary (1973). Ce débat disparaît dès lors que l'on utilise les indices avec la pondération de voisinage, ou encore mieux, directement avec une matrice de poids de voisinage bi-stochastique. Dans ce cas, les deux tests sont de même puissance. Une version commune de ces deux tests dans sa version Monte

Carlo est proposée (Annexe 2.3). Elle est définie directement sur un tableau mais le test s'effectue variable par variable :

```
test.gearymoran <- gearymoran(listw2mat(nb2listw(iris.nb)), irisdata$tab)
test.gearymoran
class: krandttest
test number: 12
permutation number: 999
  test      obs      P(X<=obs) P(X>=obs)
1  T0.10    0.582      1          0.001
2  T10.50   0.215      0.966      0.036
3  Tup50    0.585      1          0.001
4  cow      0.682      1          0.001
5  other   0.054 0.791 0.211
6  pig      0.335      0.994      0.008
7  sheep  0.07  0.81  0.192
8  town.pop 0.214      0.966      0.036
9  car      0.491      1          0.001
10 radio  -0.046 0.529 0.473
11 sales    0.424      1          0.002
12 single.man 0.246      0.98      0.022
```



```
plot(test.gearymoran)

# seules les variables 'radio' et 'other' et 'sheep' ne sont pas spatialement
# structurées
```

Les éléments théoriques ainsi que les données et les outils techniques pour aborder la question de l'ordination sous contrainte spatiale sont désormais définis. On peut dès lors aborder la question de fond.

4. HESITATIONS METHODOLOGIQUES

n unités statistiques portent une pondération de voisinage du type \mathbf{W} . Un schéma de dualité, objet de la classe 'dudi' ou triplet statistique à n lignes et p colonnes s'écrit $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$. Le but est d'introduire les contraintes de voisinage \mathbf{W} dans l'analyse de $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$. On peut le faire en choisissant le point de vue de voisinage \mathbf{L} et en partant de l'indice de Moran. Ce choix est loin d'être naturel et Lebart (1969) a fait le contraire en partant de l'indice de Geary et du point de vue de voisinage \mathbf{F} .

4.1. L'école de Lebart : variances et covariances locales

Une remarque s'impose sur le choix par Lebart de c au détriment de I dans l'approche multivariée. On comprend facilement que $c^* = \mathbf{z}'(\mathbf{P} - \mathbf{F})\mathbf{z}$ mesure la variance locale. Elle s'écrit :

$$c^* = \frac{1}{2m} \mathbf{z}'(\mathbf{N} - \mathbf{M})\mathbf{z} = \mathbf{z}'(\mathbf{P} - \mathbf{F})\mathbf{z} = \frac{1}{2} \sum_{(2)} f_{ij} (x_i - x_j)^2 = \frac{1}{2} \sum_{(2)} m_{ij} (z_i - z_j)^2$$

La généralisation de Lebart (1969) introduit la matrice de covariance spatiale $\mathbf{X}'(\mathbf{P}-\mathbf{F})\mathbf{X}$ à partir des graphes de voisinage non pondérés, qu'il appelle matrice de contiguïté en référence à l'indice de Geary. L'idée a été reprise par Monestiez (1978) et généralisée aux pondérations de voisinage quelconques dans le cadre de l'ACP par Le Foll (1982). Mom (1998) admet une pondération extérieure \mathbf{D} , qui sommée sur les voisins, donne une surpondération de voisinage \mathbf{D}_* , l'opérateur de lissage $\mathbf{H} = \mathbf{D}_*^{-1}\mathbf{M}\mathbf{D}$ et l'analyse de $((\mathbf{I}_n - \mathbf{H})\mathbf{X}, \mathbf{Q}, \mathbf{D})$. Méot et al (1993), dans la même situation, introduisent l'opérateur \mathbf{D} -symétrique $\mathbf{D}_* - \mathbf{M}\mathbf{D}$ mais tous conservent des formes quadratiques positives, donc le point de vue initial de la variance locale, qui donne pour deux tableaux l'analyse de covariance locale (Chessel & Mercier, 1993) et l'ACPVI locale (Cornillon & Sabatier, 1998). Toutes ces approches, dites encore analyses locales, portent sur la variabilité de voisinage que Benali et Escofier (1990) relie à l'analyse factorielle des différences locales.

Pourquoi donc l'idée de Lebart ne s'est-elle pas imposée concrètement ? Le doute s'installe quand Benali et Escofier (1990), dans le même article, mettent en avant l'existence de l'objectif inverse sous la forme de l'analyse factorielle lissée, c'est-à-dire l'analyse de $(\mathbf{L}\mathbf{X}, \mathbf{Q}, \mathbf{D})$, donc du tableau des moyennes de voisinage. On diagonalise encore un opérateur positif. De même, Royer (1984) puis Faraj et Cailly (2001) défendent le point de vue inverse au travers des analyses de proximités. L'objectif est parfaitement clair : il s'agit selon Royer (1984), « *de définir un indice de proximité entre échantillons puis d'optimiser le rapport de la variance locale sur la variance totale calculée sur l'ensemble des variables disponibles à l'aide des matrices de variance covariance locale et totale. Des combinaisons linéaires de variables appelées facteurs de proximité sont ainsi calculés : les premiers facteurs décrivent les composantes régionales lentement variables, les derniers facteurs représentent les anomalies locales (faible rapport signal sur bruit)* ». Les auteurs viennent de la géostatistique. Ils travaillent donc sur des variables régionalisées, et sont amenés à rechercher des composantes cartographiables d'où la nécessité de minimiser la variance locale plutôt que de la maximiser comme dans toutes les analyses locales. L'objectif est clairement antinomique à celui défini par Lebart, bien qu'il utilise le même opérateur, dans sa version d'analyse discriminante. L. Lebart a certainement eu le mérite d'ouvrir le débat et de connecter le multivarié et le spatial. Il l'a fait sur la base du multivarié en introduisant le spatial par le biais d'une métrique euclidienne. De manière générale, cette communauté est restée fermée comme en témoigne encore l'intervention de Aaufaure et al. (2000) et la

généralisation aux cubes de données de Cornillon et al. (1999), à l'exception cependant des contacts avec l'école italienne : Di Bella et Jona-Lasinio (1996) l'utilisent dans le champ de l'ordination multi-échelles ouvert par Ver Hoef et Glenn-Lewin (1989). Cette idée est reprise par Wagner (2003; 2004) et généralisée dans Couteron et Ollier (Couteron & Ollier, sous presse) (Annexe 3.2) à partir de l'opérateur défini par Méot.

Le c de Geary est indépendant du centrage puisqu'on ne prend en compte que des différences de valeurs. C'est une forme quadratique positive qui donne une métrique :

$$\langle \mathbf{y} | \mathbf{z} \rangle_c = \mathbf{y}' (\mathbf{P} - \mathbf{F}) \mathbf{z} = \frac{1}{2} \sum_{i,j} f_{ij} (y_i - y_j)(z_i - z_j)$$

La norme associée est la variance locale, le produit scalaire est la covariance locale et en introduisant en analyse de données cette métrique on obtient la famille des analyses locales. C'est simple et mathématiquement élégant, malheureusement ces analyses locales maximisent la variance locale et cet objectif est contraire à la majorité des intentions des expérimentateurs. En effet que cherche-t-on en général? Des combinaisons de variables les plus cartographiables, les plus lissées (des modèles spatiaux) donc des variables avec un *minimum* de variance locale (entre voisins). Que faire d'une analyse élégante qui est opposée au besoin le plus répandu. Évidemment, les analyses locales sont peu utilisées.

4.2. L'école de l'auto-corrélation spatiale multivariée

Seul Wartenberg (1985) a osé casser la contrainte qu'une analyse doit donner des valeurs propres positives. Il diagonalise $\mathbf{R} = \mathbf{X}'\mathbf{W}\mathbf{X} = \mathbf{X}'\mathbf{F}\mathbf{X}$ non sans précaution : « *an important difference between this approach and PCA must be pointed out. Unlike \mathbf{C} , the product-moment correlation matrix that is decomposed in PCA, \mathbf{R} is not positive definite. That is, \mathbf{R} can have negative eigenvalues, which \mathbf{C} cannot. These negative eigenvalues are as important as positive eigenvalues but are of a qualitatively different type. They represent spatial interaction (covariance) that is more important than spatial pattern (variance). ... To avoid this situation, data yielding negative eigenvalues are not used in this paper. All examples have large eigenvalues that are positive only* ».

Il sait que son analyse pourrait donner de grandes valeurs propres négatives ayant du sens mais le cache provisoirement. Il y a cependant une contradiction dans la mesure où l'indice de Moran prend tout son intérêt sur un lien \mathbf{L} et que l'analyse utilise l'indice de Moran sur un lien \mathbf{F} . Ces hésitations font qu'il y a peu d'utilisateurs de ces propositions auxquelles on

préfère les classifications sous contraintes spatiales (spatial clustering) ou les méthodes géostatistiques multivariées (Wackernagel, 2003) comme dans Monestiez et al. (1994).

Mais en géologie, en particulier en minéralogie, la situation est différente. Si on appelle **MSC** pour *Multivariate Spatial Correlation* l'analyse de Wartenberg, la **MSC** est alors voisine de variantes nées à la même époque. Elle n'est pas isolée conceptuellement mais le développement de méthodes nouvelles se fait souvent sur des idées voisines dans des environnements séparés. Est souvent mentionnée **SFA** pour *Spatial Factor Analysis* une analyse proposée et défendue par Grunsky et Agterberg (Grunsky & Agterberg, 1988; Grunsky & Agterberg, 1989, 1991; Grunsky et al., 1996) alors que le terme "spatial factor analysis" renvoie souvent à **MAF** pour *Min/Max Autocorrelation Factor Analysis* créé par Switzer et Green (Switzer & Green, 1984) dans un rapport souvent cité qui a été ensuite redécrit et utilisé à plusieurs reprises par Nielsen et son équipe (Conradsen et al., 1985; Ersbll, 1989; Nielsen, 1995a, b; Nielsen, 1999; Nielsen & Conradsen, 1997; Nielsen et al., 1997; Nielsen et al., 1998).

Dans les trois cas, on utilise une matrice d'autocorrélation croisée entre variables. Pour **MSC**, la plus simple, il s'agit du produit scalaire entre une variable et la moyenne de l'autre sur l'ensemble des points voisins. Pour **MAF**, la seule relation de voisinage envisagée est celle qui relie deux pixels au pas h en x ou y . On a un coefficient d'autocorrélation spatiale au pas h dans un modèle anisotrope. Pour **SFA**, méthode la plus compliquée, la relation de voisinage est définie par un rayon D au delà duquel il n'y a plus de relation de voisinage, et une fonction d'influence qui pondère le voisinage avec une quantité du type $a + bd_{ij} + cd_{ij}^2$ entre deux points i et j tels que leur distance vérifie $d_{ij} < D$. Dans tous les cas, une matrice \mathbf{R} mesure l'association spatiale entre variables. Dans **MAF**, il s'agit théoriquement de corrélation et de questions d'estimation sur les bords. Dans les deux méthodes, il s'agit de produits scalaires et donc de coefficients d'association au sens large. Les trois méthodes n'ont envisagé que des variables quantitatives, normalisées au préalable.

Le lien spatial utilisé dans **SFA** en fait plutôt une curiosité (Grunsky, 2002). Le lien de Wartenberg est le plus général et convient parfaitement en écologie et en économie. Le lien de Nielsen est celui qui est adaptée à l'analyse des images de télédétection comme nous allons le voir ci-après.

Mais il y a quand même entre les deux méthodes une différence de taille. La **MSC** diagonalise directement \mathbf{R} alors que **MAF** est basée sur la diagonalisation de $\mathbf{C}^{-1}\mathbf{R}$, où \mathbf{C} est la matrice de corrélation ordinaire. La **MSC** est une **ACP** (pour augmenter la covariation

spatiale, on doit d'abord augmenter la variance) sous contrainte (la variance ne doit pas augmenter trop au détriment de l'indice de Moran). La **MAF** est apparentée aux analyses discriminantes, et fournit des scores canoniques de moyenne 0 et variance 1 qui maximisent strictement l'indice de Moran. Elle est invariante par combinaisons linéaires de rang plein des données départ. Ceci n'est possible, sans rencontrer d'énormes problèmes de stabilité numérique, que pour des nombres de lignes considérables, ce qui est le cas en analyse d'images. De la télédétection à l'imagerie du cerveau, en passant par les réseaux de stations écologiques, il n'y aurait guère de sens à réclamer une méthode unique, pas plus qu'il n'y a unicité de la définition des pondérations de voisinage.

La variance locale est une forme quadratique et a été intégrée naturellement en analyse des données. La notion d'autocorrélation spatiale ne l'est pas. Mais la signification de l'indice de Moran est parfaitement claire pour des variables centrées :

$$I^* = \mathbf{z}'\mathbf{F}\mathbf{z} = \sum_{ij} f_{ij} z_i z_j$$

Cette quantité est d'autant plus grande (respectivement petite) que de grandes valeurs positives (respectivement négatives) se trouvent associées sur des couples d'observations ayant un grand poids de voisinage. Wartenberg (Blondel, 1985) a utilisé l'autocorrélation spatiale dans l'interprétation d'une analyse ordinaire en diagonalisant la matrice des covariances spatiale définie par les produits (Wartenberg, 1985) :

$$\langle \mathbf{y} | \mathbf{z} \rangle_I = \mathbf{y}'\mathbf{F}\mathbf{z} = \sum_{i,j} f_{ij} y_i z_j$$

Cette quantité ne peut être, malheureusement, un coefficient de corrélation au sens strict du terme, que si le centrage est fait avec une moyenne calculée pour les poids de voisinage des points, et si la normalisation est faite en divisant par un écart-type calculé avec la même pondération. En outre, cette forme quadratique n'est pas positive, et l'analyse peut avoir des valeurs propres négatives. Cette insertion n'est pas optimum du point de vue mathématique, tout en étant très légitime du point de vue expérimental. C'est moins beau, mais c'est beaucoup plus utile. Comme on l'a vu en univarié, on peut concilier les deux points de vue (Thioulouse et al., 1995) en n'utilisant que des données centrées et normées pour la pondération de voisinage. Toutefois, l'introduction d'une pondération de voisinage souvent non uniforme est une contrainte très forte. C'est donc élégant mais peu utilisé. Aussi, afin de définir une analyse sous contrainte sur l'ensemble des triplets statistiques, on s'en est tenu au point de vue de Wartenberg dans la lignée de Moran.

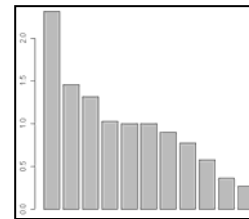
5. GÉNÉRALISATION DE L'APPROCHE DE WARTENBERG

5.1. Principes

$(\mathbf{X}, \mathbf{Q}, \mathbf{D})$ est un schéma de dualité ou analyse de premier niveau (Annexe). \mathbf{X} est un tableau, \mathbf{Q} une pondération de ses colonnes, et \mathbf{D} une pondération de ses lignes. Le plus utile des objets de ce type dérive d'un tableau quelconque contenant des variables quantitatives (*numeric*) et des variables qualitatives (*factor*) voire même des qualitatives à modalités ordonnées (*ordered*). Les quantitatives sont généralement centrées et réduites, les qualitatives décomposées en indicatrices de classes puis centrées correctement et les pondérations font en sorte que chaque variable ait le même poids que les autres. La fonction `dudi.mix(...)` qui met en œuvre l'analyse mixte (1994) assure cette opération. On l'applique sur le jeu de données oribatid (Annexe 1.17) afin d'illustrer les principales propriétés de cette méthode :

```
data(orbitaid)
ori.mix <- dudi.mix(orbitaid$envir)
Select the number of axes: 3
```

```
ori.mix
Duality diagramm
class: mix dudi
$call: dudi.mix(df = orbitaid$envir)
```



```
$nf: 3 axis-components saved
# Une valeur propre intéressante et beaucoup d'inertie désorganisée
```

```
$rank: 11
eigen values: 2.312 1.456 1.316 1.031 1 ...
```

```
vector length mode content
1 $cw 14 numeric column weights
2 $lw 70 numeric row weights
3 $eig 11 numeric eigen values
```

```
data.frame nrow ncol content
1 $stab 70 14 modified array
2 $li 70 3 row coordinates
3 $li 70 3 row normed scores
4 $co 14 3 column coordinates
5 $cl 14 3 column normed scores
other elements: assign index cr
```

Les poids des colonnes sont 1 pour les quantitatives et les fréquences des modalités pour les qualitatives :

```
ori.mix$cw
subst.inter subst.litter subst.peat subst.sph1 subst.sph2 subst.sph3
0.38571 0.02857 0.02857 0.35714 0.15714 0.01429
subst.sph4 shrub.few shrub.many shrub.none topo.blanket topo.hummock
0.02857 0.37143 0.35714 0.27143 0.62857 0.37143
density water
```

```

      1.00000      1.00000
sum(ori.mix$cw)
[1] 5

```

La pondération des lignes est uniforme :

```

unique(ori.mix$lw)
[1] 0.01429
1/nrow(ori.mix$tab)
[1] 0.01429

```

Les axes principaux ordinaires sont des vecteurs en colonnes dans une matrice \mathbf{U}_r (r est le nombre de facteurs conservés dans l'analyse simple) \mathbf{Q} -orthonormés :

$$\mathbf{U}_r^t \mathbf{Q} \mathbf{U}_r = \mathbf{I}_r$$

Les coordonnées de l'analyse simple $\mathbf{L}_r = \mathbf{X} \mathbf{Q} \mathbf{U}_r$ maximisent successivement l'inertie projetée sur un axe \mathbf{u} soit $\|\mathbf{X} \mathbf{Q} \mathbf{u}\|_{\mathbf{p}}^2$. Les maxima successifs sont les valeurs propres de l'analyse simple qu'on notera $\lambda_1, \dots, \lambda_r$. Dans cet exemple, cela signifie que la première coordonnée est un score numérique \mathbf{z} qui maximise la somme des carrés de corrélation $\text{corr}^2(\mathbf{z}, \mathbf{X}[, j])$ quand la variable $\mathbf{X}[, j]$ est quantitative et la somme des rapports de corrélation $\eta^2(\mathbf{z}, \mathbf{X}[, j])$ quand elle est qualitative. Cette analyse est une Analyse en Composantes Principales normée sur matrice de corrélation quand il n'y a que des variables quantitatives et une Analyse des Correspondances Multiples quand il n'y a que des variables qualitatives. Cette propriété se retrouve par l'intermédiaire des listings ainsi que sur le graphe canonique associée à l'analyse (Figure 1.14).

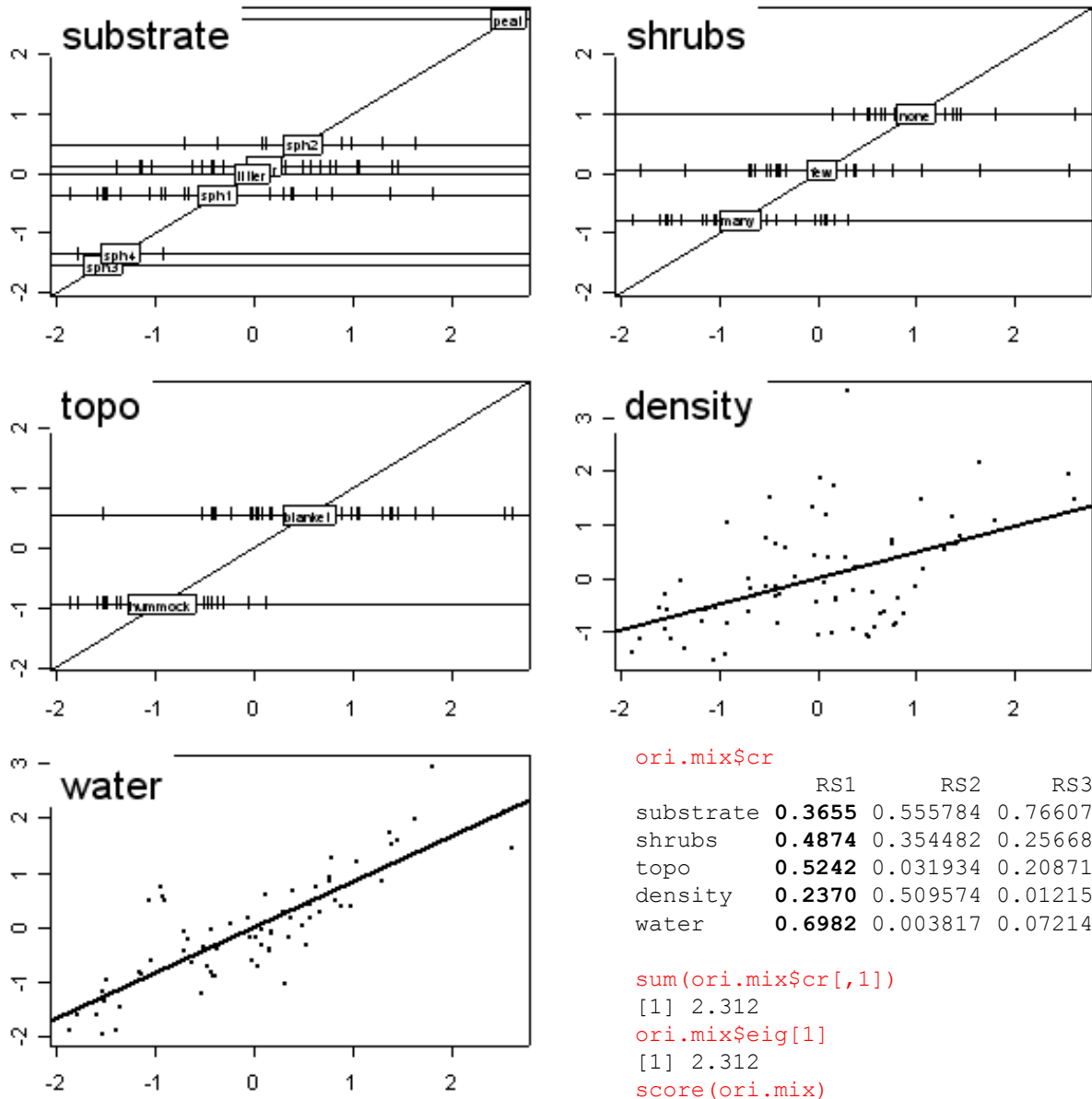


Figure 1.14 : Expression du lien entre les variables environnementales du jeu de données oribatid (Annexe) avec le premier score de synthèse de l'analyse mixte.

Au triplet étudié est associé un opérateur de lissage L qui permet de calculer $Y = LX$ où chaque valeur initiale au point i de la variable j est remplacée par la moyenne des valeurs des voisins de i pour la même variable j . Pour une variable on a $y = Lx$ et le graphe du couple (x,y) est le scatterplot de Moran (Figure 1.13). Ainsi étendue, l'opération génère un deuxième tableau totalement apparié au premier et donc un deuxième nuage de n points de \mathbb{R}^p qu'on peut projeter sur les axes principaux. On peut donc calculer l'autocorrélation des coordonnées et représenter leurs scatterplots de Moran séparément (Figure 1.15).

```

ori.dn <- dnearneigh(as.matrix(orbitid$xy),0,1.5)
ori.listw <- nb2listw(ori.dn)
u <- lapply(ori.mix$li,moran.mc,listw = ori.listw, nsim = 999)
plot(as.randtest(u[[1]]$res, u[[1]]$statistic), main = "test de Moran: score 1")
plot(as.randtest(u[[2]]$res, u[[2]]$statistic), main = "test de Moran: score 2")
moran.plot(ori.mix$li[,1], ori.listw, pch = 20, xlab = "score 1", ylab = "lag.score
1", main = "scatterplot de Moran: score 1")
moran.plot(ori.mix$li[,2], ori.listw, pch = 20, xlab = "score 2", ylab = "lag.score
2", main = "scatterplot de Moran: score 2")

```

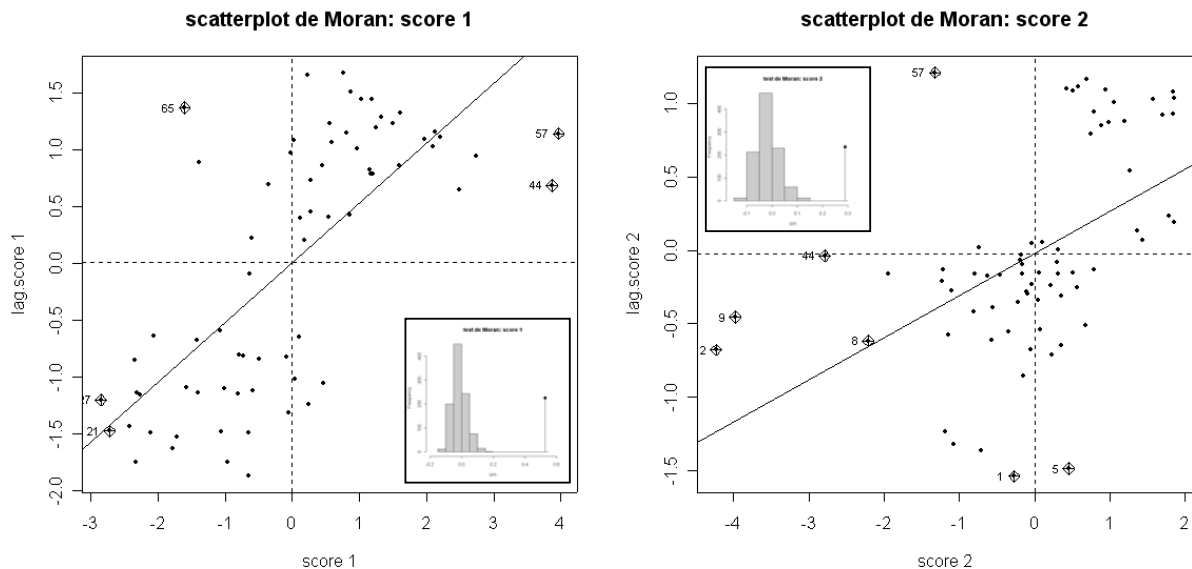


Figure 1.15 : scatterplot de Moran des deux premières coordonnées de l'analyse mixte. Toutes les deux présentent une autocorrelation positive bien que l'analyse n'optimise pas *a priori* leurs propriétés d'autocorrelation spatiale.

Les coordonnées sont de variance maximale et définissent la carte factorielle ordinaire, seconde représentation canonique associée à l'analyse mixte du triplet statistique. On peut alors représenter simultanément sur la carte factorielle les deux scatterplots de Moran (Figure 1.16). Cette figure est une image canonique de l'analyse mixte dans la mesure où c'est elle qui représente le plan pour lequel la représentation du nuage projeté est optimale. Toutefois, les coordonnées n'ayant pas *a priori* de propriétés particulières d'autocorrelation spatiale, cette figure n'est pas optimale du point de vue de la longueur des flèches qui relient les individus projetés sur les axes principaux aux positions moyennes des voisins de ces individus. L'analyse que l'on propose a justement pour objectif de faire de ce graphique un graphique canonique en gardant une part des propriétés de l'analyse classique et en intégrant le voisinage.

```

w <- as.data.frame(apply(ori.mix$li, 2, lag.listw, x = ori.listw))
row.names(w) <- row.names(ori.mix$li)
s.match(ori.mix$li,w, clab = 0.75)

```

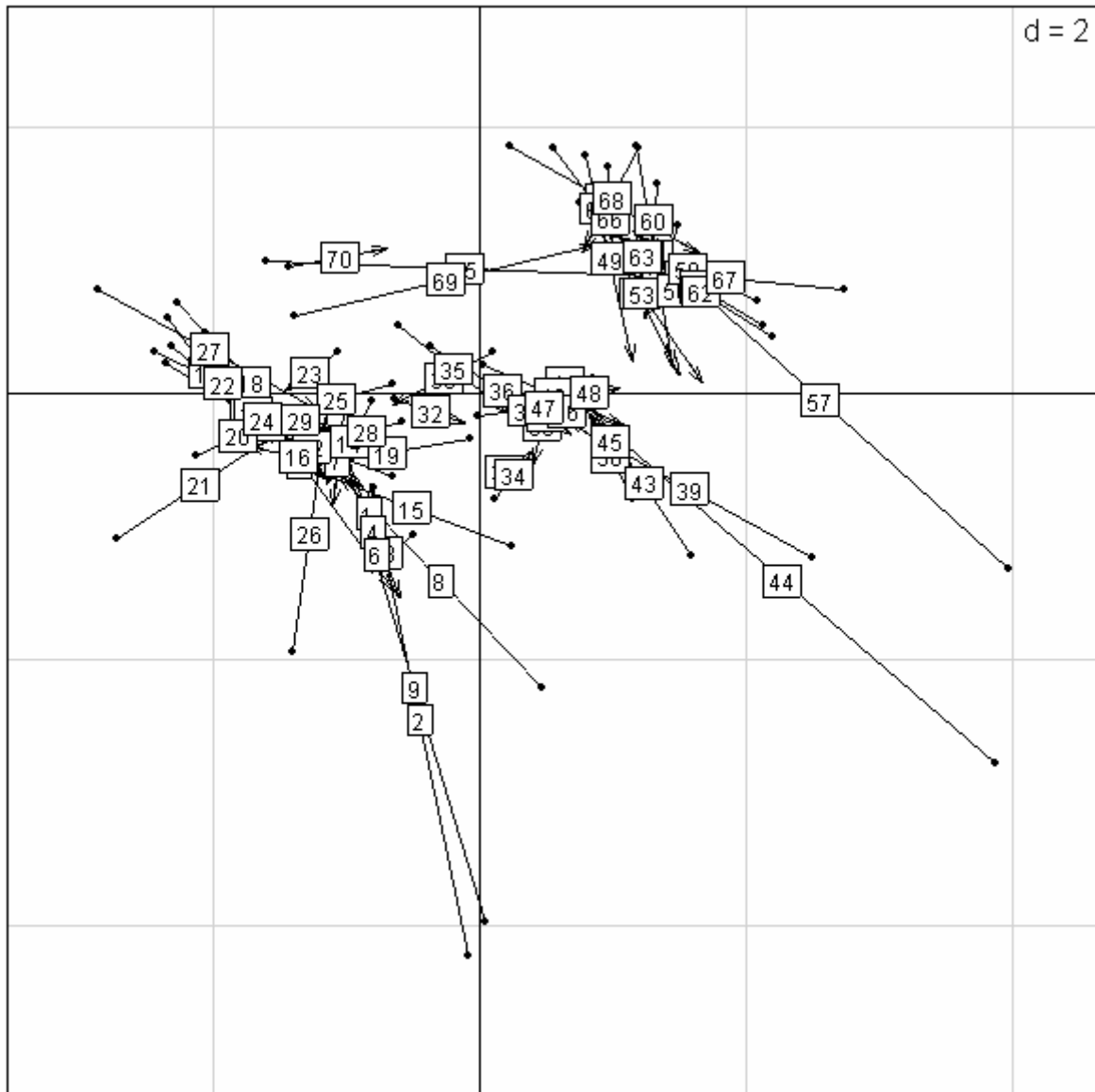


Figure 1.16 : La carte factorielle ordinaire est celle des points à l'origine des flèches. L'extrémité de la flèche est la position moyenne des voisins du point. En quelque sorte, ce graphe est une généralisation à deux dimensions du scatterplot de Moran.

5.2. Définitions

Quelle que soit l'analyse considérée (Tableau 1.1), on comprend que chaque axe de \mathbb{R}^p définit un système de coordonnées qui est plus ou moins autocorrélé.

Fonction	Analyse	Référence
dudi.pca	Principal component Analysis	1
dudi.coa	Correspondence Analysis	2
dudi.acm	Multiple Correspondence Analysis	3
dudi.fca	Fuzzy Correspondence Analysis	4
dudi.mix	Mixture of numeric and factors	5
dudi.nsc	Non Symetric Correspondence Analysis	6
dudi.dec	Decentred Correspondence Analysis	7

Tableau 1.1 : Les différentes analyses d'un triplet statistique (Escoufier, 1987) implémentées dans ade4. 1—Pearson(1901), 2—Greenacre(1984), 3—Tenenhaus and Young(1985), 4—Chevenet et al.(1994), 5—Hill and Smith(1976), Kiers(1994), 6—Kroonenberg and Lombardo(1999), 7—Dolédec et al.(1995).

Les axes principaux de l'analyse simple maximisent l'inertie projetée et n'ont aucune propriété d'autocorrélation particulière. On cherche alors ceux qui maximisent l'autocorrélation. La solution n'est pas ordinaire car le critère est :

$$I(\mathbf{XQ}\mathbf{u}) = \frac{\mathbf{u}'\mathbf{Q}'\mathbf{X}'\mathbf{D}\mathbf{L}\mathbf{X}\mathbf{Q}\mathbf{u}}{\mathbf{u}'\mathbf{Q}'\mathbf{X}'\mathbf{D}\mathbf{X}\mathbf{Q}\mathbf{u}}$$

Considérons la matrice $\mathbf{H} = \frac{1}{2}\mathbf{X}'(\mathbf{L}'\mathbf{D} + \mathbf{D}\mathbf{L})\mathbf{X}\mathbf{Q}$. Elle est \mathbf{Q} -symétrique et possède une base de vecteurs propres \mathbf{Q} -orthonormés. Le premier vecteur propre \mathbf{u}_1 associé à la plus grande valeur λ_1 réalise le maximum de :

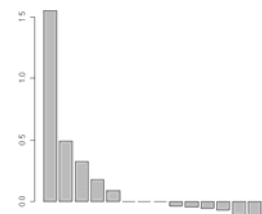
$$\langle \mathbf{H}\mathbf{u} | \mathbf{u} \rangle_{\mathbf{Q}} = \mathbf{u}'\mathbf{Q}'\mathbf{X}'\mathbf{D}\mathbf{L}\mathbf{X}\mathbf{Q}\mathbf{u} = \|\mathbf{X}\mathbf{Q}\mathbf{u}\|_{\mathbf{D}}^2 I(\mathbf{u}) = \text{var}(\mathbf{X}\mathbf{Q}\mathbf{u}) I(\mathbf{X}\mathbf{Q}\mathbf{u}) \text{ avec } \|\mathbf{u}\|_{\mathbf{Q}} = 1$$

Le cas particulier pour une ACP normée est l'analyse de Wartenberg (1985) quand on utilise un lien normalisé par ligne ou la **MAF** de Switzer et Green (1984) étendue à une pondération de voisinage quelconque mais sans inversion de métrique. On appellera **MS** pour multivarié spatial la recherche de la base de vecteurs propres de \mathbf{H} et son usage. La fonction `multispati(...)` (Annexe 2.12) fait cela.

5.3. La fonction `multispati(...)`

La fonction `multispati(...)` (Annexe 2.12) utilise un quadruplet $\left(\begin{matrix} \mathbf{X}, & \mathbf{Q}, & \mathbf{D}, & \mathbf{L} \\ n \times p, & p \times p, & n \times n, & n \times n \end{matrix} \right)$ dont les dimensions sont indiquées en associant une pondération de voisinage (objet de la classe 'listw') à un schéma de dualité (objet de la classe 'dudi').

```
ori.mix.ms <- multispati(ori.mix, ori.listw)
```



Select the first number of axes (≥ 1): 2 # valeurs positives
 Select the second number of axes (≥ 0): 0 # valeurs négatives

La fonction calcule puis diagonalise la matrice \mathbf{H} afin de définir les r axes principaux \mathbf{U}_r de l'analyse qui maximisent $\mathbf{u}'\mathbf{Q}'\mathbf{X}'\mathbf{D}\mathbf{L}\mathbf{X}\mathbf{Q}\mathbf{u} = \text{var}(\mathbf{X}\mathbf{Q}\mathbf{u})/I(\mathbf{X}\mathbf{Q}\mathbf{u})$ donc l'autocovariance des r coordonnées $\mathbf{X}\mathbf{Q}\mathbf{U}_r$, sous la contrainte $\|\mathbf{u}\|_{\mathbf{Q}} = 1$. Peuvent être conservés aussi bien les axes dont l'autocovariance des coordonnées est positive que ceux dont l'autocovariance est négative, d'où le double choix du nombre d'axes.

```
Ur <- as.matrix(ori.mix.ms$c1) # axes principaux
Q <- diag(ori.mix.ms$cw)
round(t(Ur) %*% Q %*% Ur)
```

```
      CS1 CS2
CS1    1  0
CS2    0  1
```

```
s.arrow(ori.mix.ms$c1, clab = 0.65)
```



La fonction calcule ensuite le tableau apparié $\mathbf{Y} = \mathbf{L}\mathbf{X}$ et les coordonnées des lignes de \mathbf{Y} sur les axes principaux de l'analyse. La représentation simultanée des individus de \mathbf{X} et de \mathbf{Y} par leurs coordonnées sur les axes principaux est une représentation canonique de l'analyse : c'est l'image de la maximisation du compromis entre inertie projetée et autocorrelation des coordonnées. C'est un compromis entre la carte factorielle et les scatterplots de Moran :

```
Vr <- ori.mix.ms$li # coordonnées de X
Vr.lag <- ori.mix.ms$ls # coordonnées de LX
s.match(Vr, Vr.lag)
Vr <- as.matrix(Vr)
Vr.lag <- as.matrix(Vr.lag)
D <- diag(ori.mix.ms$lw)
t(Vr) %*% D %*% Vr # var(XQu)

      CS1  CS2
CS1 2.0892 0.2201
CS2 0.2201 1.1561

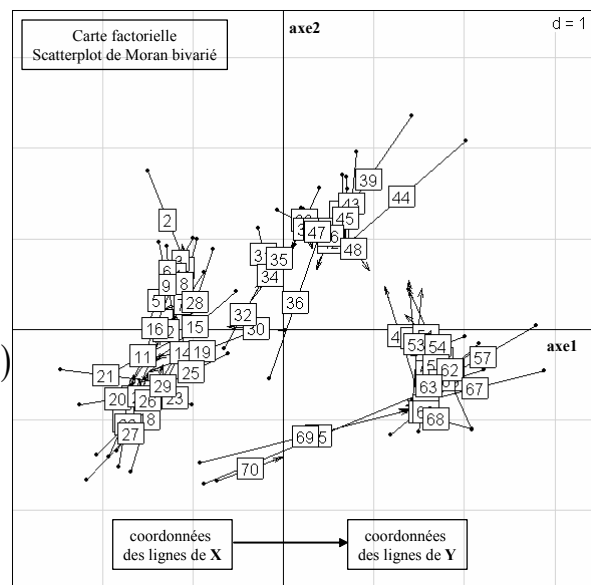
(t(Vr) %*% D %*% Vr.lag) / t(Vr) %*% D %*% Vr # I(XQu)

      CS1  CS2
CS1 0.7418 0.1404
CS2 -0.1404 0.4263

ori.mix.ms$eig[1:2]
[1] 1.5499 0.4928

diag((t(Vr) %*% D %*% Vr.lag)) # u'Q'X'DLXQu

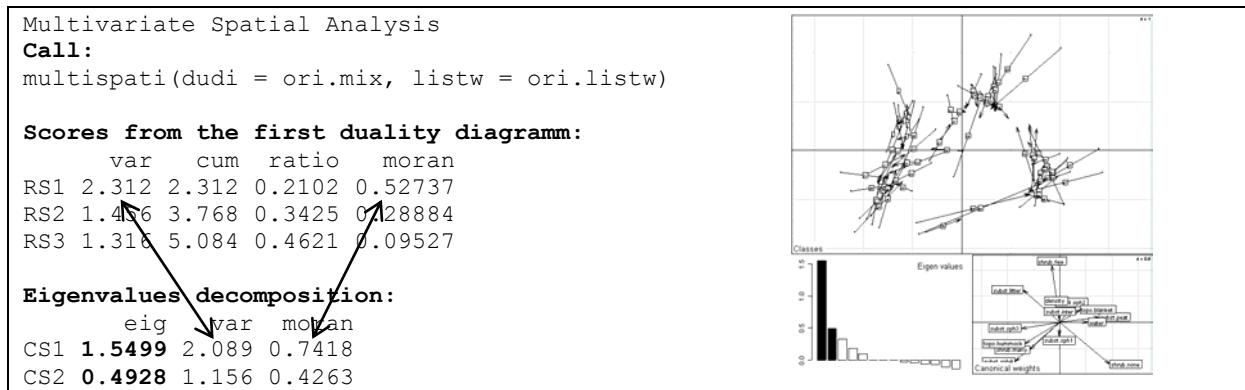
      CS1  CS2
1.5499 0.4928
```



L'ensemble des ces informations peut être obtenu directement par les deux fonctions génériques `summary.multispati(...)` et `plot.multispati(...)` (Annexe 2.12). La notion de compromis est apparente dans les résultats : on y perd du point de vue de l'inertie conservée mais on y gagne du points de vue de l'information spatiale comme on peut le voir en comparant les valeurs liées par les flèches :

`summary.multispati(ori.mix.ms)`

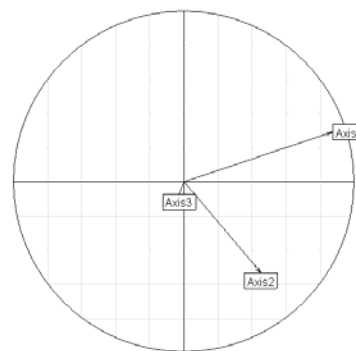
`plot.multispati(ori.mix.ms)`



Afin d'avoir une visualisation des changements apportés par l'analyse sous contrainte, on projette finalement les axes principaux de l'analyse simple sur les axes principaux de l'analyse sous contrainte :

```
t(as.matrix(ori.mix$cl))%*%Q%*%Ur
      CS1      CS2
CS1  0.87487  0.29423
CS2  0.45212 -0.53496
CS3 -0.02875 -0.07265

ori.mix.ms$as
      CS1      CS2
Axis1 0.87487  0.29423
Axis2 0.45212 -0.53496
Axis3 -0.02875 -0.07265
```

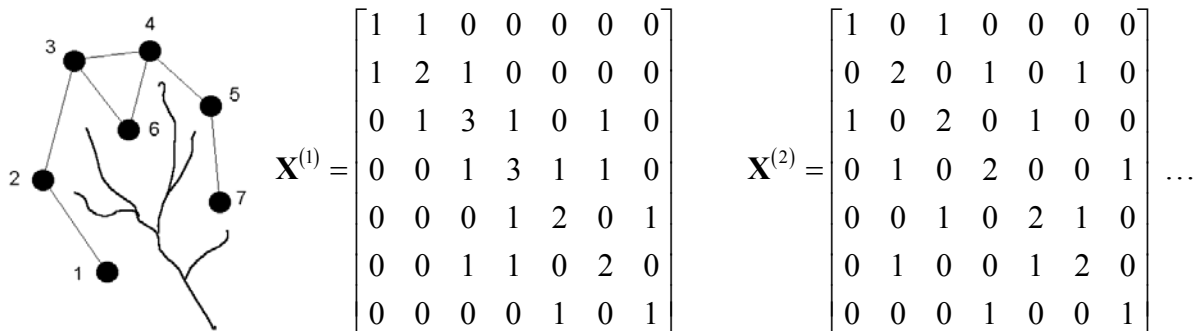


5.4. Un test de permutation multivarié contre l'absence de structure spatiale

Le test effectué variable par variable, tel que défini précédemment est le test de permutations de l'indice de Moran dans sa version avec pondération de voisinage. Ce test apparaît intéressant de prime abord pour se faire une idée sur la structure spatiale de chaque variable, donc sur la structure du tableau. Toutefois, un test global est nécessaire : l'argument donné par Smouse et Peakall (1999), qui introduisent le corrélogramme multivarié en

génétique, est exemplaire : « *population genetic theory predicts that plant populations will exhibit internal spatial autocorrelation when propagule flow is restricted, but as an empirical reality, spatial structure is rarely consistent across loci or sites, and is generally weak. A lack of sensitivity in the statistical procedures may explain the discrepancy. Most work to date, based on allozymes, has involved pattern analysis for individual alleles, but new PCR-based genetic markers are coming in vogue, with vastly increased number of alleles. The field is badly in need of an explicitly multivariate approach that is applicable to multiallelic codominant, multilocus array. The procedure treats the genetic data set as a whole, strengthening the spatial signal and reducing the stochastic (allele-to-allele, and locus-to-locus) noise* ».

Il s'agit donc de coupler un tableau multivarié avec l'espace. On peut reprendre la proposition de Smouse et Peakall (1999) : « *we (i) develop a very general multivariate method, based on genetic distance methods, (ii) illustrate it for multiallelic codominant loci, and (iii) provide non parametric permutational testing procedures for the full correlogram* ». Les individus statistiques sont des organismes ayant subi un typage multilocus. La première partie porte donc sur l'approche des données. Est ensuite abordée l'insertion de l'espace par le biais d'un graphe de voisinage avec une notion d'échelle. La figure explicative du choix est explicite :



On reconnaît les matrices du type $\mathbf{N} + \mathbf{M}$ des relations de voisinage au pas 1 (points reliés par une arêtes) puis au pas 2 (points reliés par un chemin de longueur 2) ... Ceci permet de définir une autocorrélation spatiale au pas h par (formule (15) p. 566) :

$$r^{(h)} = \left(\sum_{i \neq j}^N x_{ij}^{(h)} c_{ij} \right) / \sum_{i=1}^N x_{ii}^{(h)} c_{ii}$$

Cette quantité s'écrit, parce que toute les matrices sont symétriques :

$$r^{(h)} = \frac{\text{Trace}(\mathbf{MXX}^t)}{\text{Trace}(\mathbf{NXX}^t)} = \frac{\text{Trace}(\mathbf{X'FX})}{\text{Trace}(\mathbf{X'PX})}$$

La corrélation de Smouse et Peakall peut donc s'écrire dans le cas général d'une pondération de voisinage quelconque :

$$r = \frac{\text{Trace}(\mathbf{X'DLXQ})}{\text{Trace}(\mathbf{X'DXQ})}$$

La définition de Smouse et Peakall est donc étendue à toute pondération de voisinage et à tout type d'analyse élémentaire. Le test de permutations associé considère que les lignes du tableau et leur poids dans l'analyse sont attribués au hasard dans l'espace. La fonction `multispati.randtest(...)` (Annexe 2.13) fait le calcul. Le test global portant sur la trace de l'analyse sous contrainte, il sera vraisemblablement pris en défaut par un mélange de variables (dans le même tableau) respectivement à variance locale forte et à autocorrélation spatiale forte.

6. ILLUSTRATIONS

On possède désormais les outils pour introduire le voisinage dans l'analyse d'un tableau multivarié. On peut donc reprendre les exemples présentés dans l'introduction. Le jeu de données sur les comtés d'Irlande est traité en détail dans Ollier et al. ((soumis), Annexe 3.1).

6.1. Analyses à composantes cartographiables

Quel que soit le type d'analyse envisagé, on retrouve la même structure pour les cinq exemples considérés (Figure 1.17).

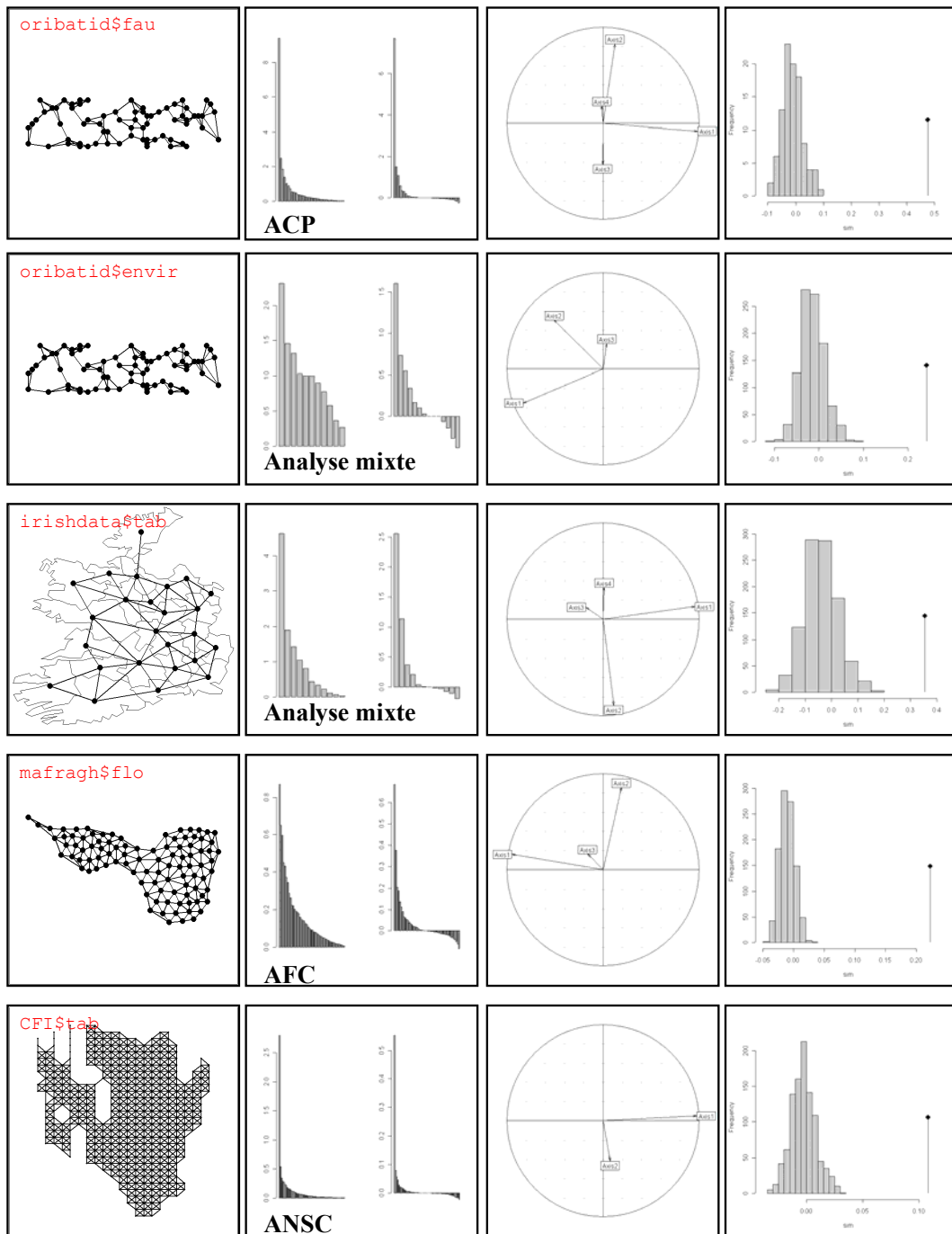


Figure 1.17 : Analyses simples et analyses sous contrainte spatiale pour 5 jeux de données différents (Annexe 1.17, 1.11, 1.13, 1.5). **La première colonne** représente les graphes de voisinage associés à chaque jeu de données. **La deuxième colonne** représente les graphes des valeurs propres. A gauche l'analyse simple, à droite l'analyse sous contrainte. Les valeurs propres, obtenues par deux logiques différentes ne sont pas comparables. Toute l'information spatiale s'exprime sur les premiers axes de l'analyse sous contrainte. La présence de valeurs propres négatives n'a ici aucune signification : l'ordre de grandeur en valeur absolue est celui des axes négligeables de l'autre signe et indique un bruit de fond aléatoire. **La troisième colonne** représente les projections sur le plan défini par les deux premiers axes principaux de l'analyse sous contrainte des premiers axes principaux de l'analyse simple. C'est le moyen le plus efficace de mesurer rapidement l'effet de la contrainte spatiale. On voit ici que les plans 1-2 des deux analyses sont voisins et que la contrainte spatiale fait une correction technique sans changer l'interprétation de la structure multivariée. **La dernière colonne** représente les résultats du test global. Bien évidemment, les données sont fortement structurées dans l'espace pour les cinq jeux de données considérés.

Globalement, la variabilité de chacune des variables est d'ordre spatial et la covariance des variables est entièrement une conséquence des variations spatiales. Quand la structure spatiale est forte, l'analyse simple donne le résultat presque optimum pour les composantes cartographiables mais il y a des exemples plus complexes. Les plans définis par les premiers axes principaux sont donc largement conservés. Toutefois, l'ordre d'intervention des axes peut parfois être un peu modifié par l'analyse sous contrainte, les axes les plus structurés opérant en premier. De même, la lecture des résultats d'un point de vue spatial est simplifiée. Le petit plus est sensible dans le scatterplot de Moran, nouveau type de carte factorielle qui justifie l'analyse sous contrainte (Figure 1.18).

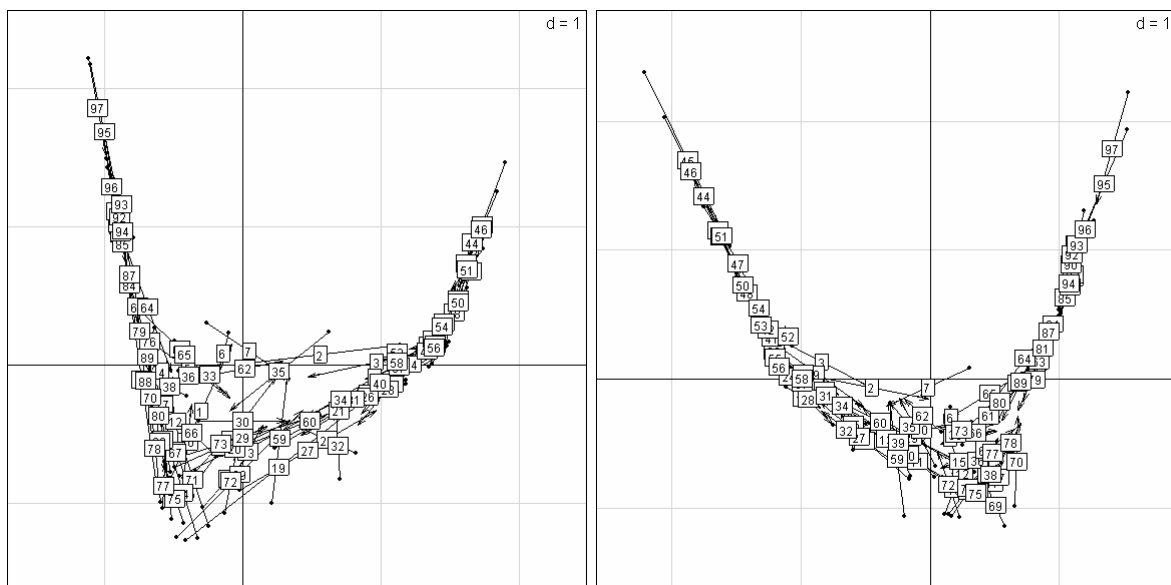


Figure 1.18 : Généralisation à deux dimensions du Moran-plot (données `mafragh$flo`). Chaque point est positionné par ses coordonnées et relié par un vecteur à la position moyenne des voisins pondérée par les poids de voisinage. A gauche analyse simple, à droite analyse sous contrainte. A gauche maximisation de la variance qui, comme les variables sont structurées dans l'espace, est essentiellement spatiale. A droite, on conserve bien la typologie tout en améliorant l'autocorrelation de chacune des coordonnées.

De manière générale, la contrainte spatiale simplifie l'interprétation. Dans le cas où les composantes principales sont fortement cartographiables, on fait une typologie de cartes en associant les variables qui ont même structure spatiale pour créer des cartes de synthèse (Figure 1.19).

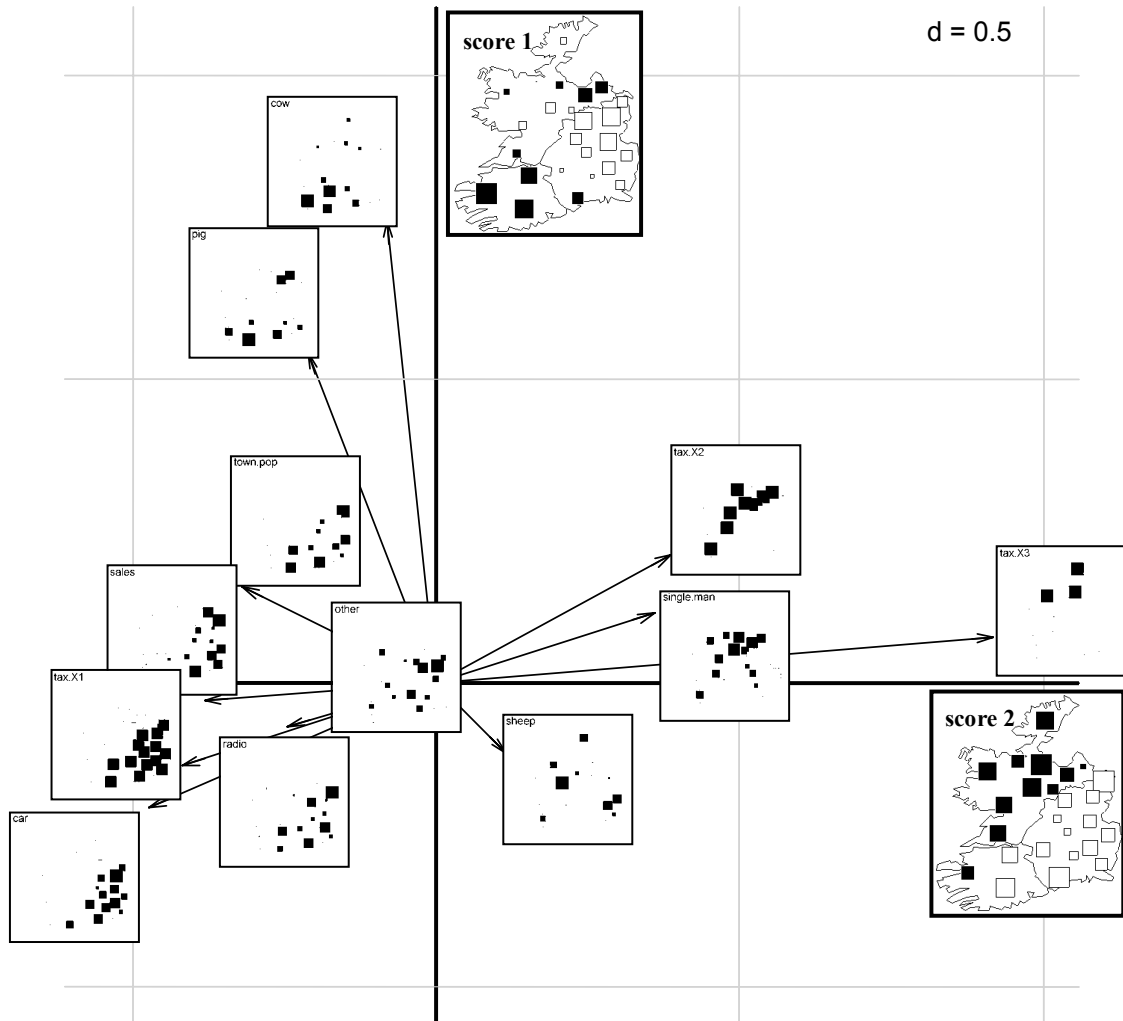


Figure 1.19 : Projection sur le plan formé par le couple d'axes principaux. Les coordonnées des variables sont les coordonnées dans l'approche classique. Ceci forme une sorte de typologie de cartes exactement comme le cercle des corrélations forme une typologie de variables. Au centre du plan, on retrouve les trois variables 'radio', 'other' et 'sheep' non structurées spatialement. Il existe deux structures cartographiables, celle de la richesse totale (nord-est, sud-ouest : axe 1) liée à la plus grande partie des variables et celle de l'élevage bovin et porcin (axe 2) largement indépendant de la précédente.

Contrairement à l'exemple suivant, l'information n'est toutefois pas exclusivement de nature spatialisée et il se pourrait que l'espace soit une contrainte énorme qui cache peut-être des relations d'une autre nature ayant un intérêt écologique. D'où les travaux qui visent à débarrasser les données des composantes spatiales (Borcard & Legendre, 1994; Borcard et al., 1992; Meot et al., 1993).

6.2. Une information exclusivement cartographiable

Le jeu de données t3012 (Annexe 1.21) fournit un exemple où la variabilité est exclusivement d'ordre spatiale, la variance locale étant de plus inexistante (Figure 1.20).

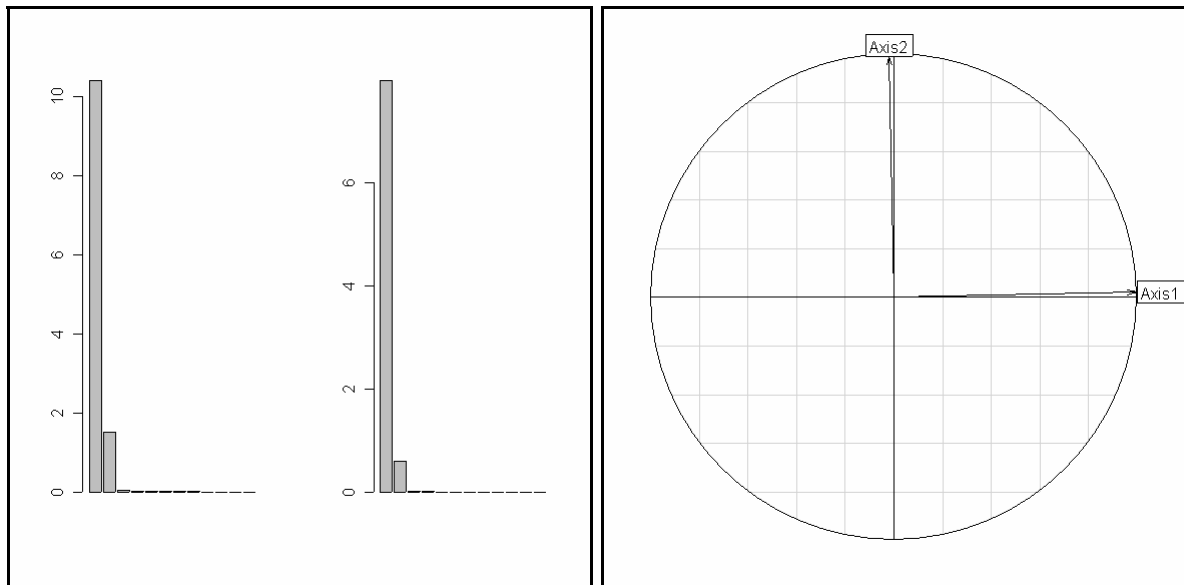


Figure 1.20 : Graphes des valeurs propres (à gauche). À gauche l'analyse simple, à droite l'analyse sous contrainte. Les valeurs propres, issues de logiques différentes ne sont pas comparables. Il n'y a même pas de valeurs propres négatives. Projections sur le plan défini par les deux premiers axes principaux de l'analyse sous contrainte des deux premiers axes principaux de l'analyse simple (à droite). On voit ici que les plans 1-2 des deux analyses sont quasiment identiques. La variabilité est donc exclusivement de nature spatiale et la variance locale n'existe pas vu l'importance des valeurs propres négatives. L'ACP rend donc parfaitement compte de l'information spatiale car dans ce cas, maximiser la variance revient à maximiser l'autocorrelation spatiale positive.

L'ACP rend parfaitement compte des deux composantes cartographiables (Figure 1.21). La première correspond au gradient Nord-Sud : quelle que soit la période de l'année, il est bien connu qu'il fait plus chaud au Sud qu'au Nord. La deuxième composante correspond au gradient Est-Ouest, traduisant l'opposition entre la façade océanique et l'intérieur continental principalement liée à l'influence des courants marins (douceur océanique/froid continental en hiver et fraîcheur océanique/chaleur continentale en été).

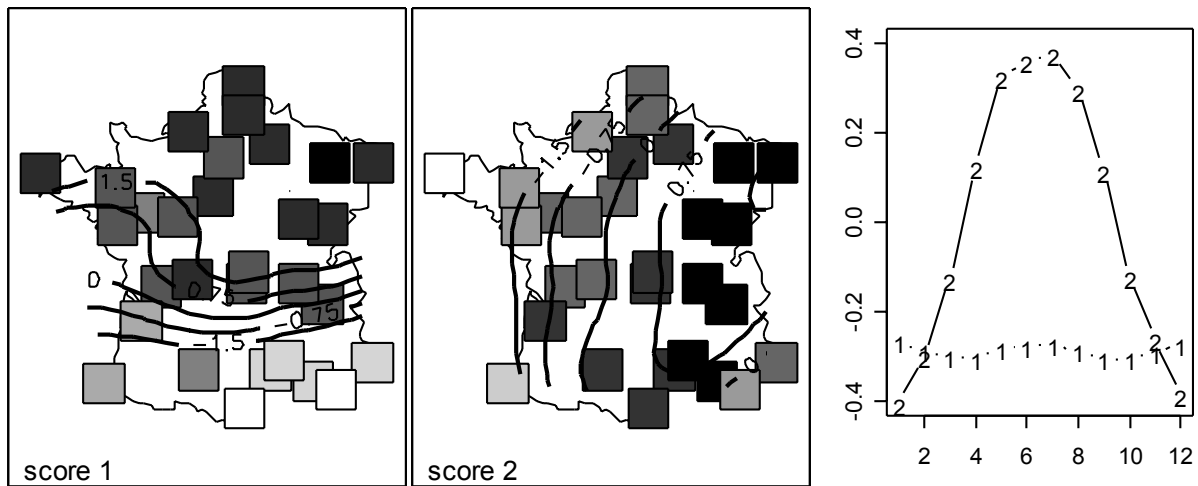


Figure 1.21 : Cartographie des deux premiers scores de l'ACP du tableau t3012\$temp (Annexe). Projection des variables sur les deux axes principaux (à droite).

6.3. Mélanges entre variance globale et variance locale

Contrairement aux analyses locales, les valeurs propres de l'analyse proposée peuvent être négatives car la matrice diagonalisée n'est pas définie positive. Selon Wartenberg (1985), « *these negative eigenvalues are as important as positive eigenvalues but are of a qualitatively different type. They represent spatial interaction (covariance) that is more important than spatial pattern (variance)* ». En effet, les valeurs propres négatives, au même titre que les valeurs propres positives, définissent une composante particulière de la variabilité spatiale dont l'autocorrelation est optimale. Les valeurs propres positives définissent des structures spatiales cartographiables. Les valeurs propres négatives définissent à l'inverse des structures spatiales dont la variabilité locale est maximale. Ce n'est pas le point de vue de Grunsky et Agteberg (Allain & Cloitre, 1991) qui considèrent l'obtention de valeurs propres négatives comme une aberration du point de vue mathématique. Selon eux, « *for the approach to be valid, matrices C and R must be positive definite* ». Contrairement à Wartenberg, ils choisissent un point de vue sans en assumer les conséquences et éliminent de fait une partie de l'information mise en évidence par l'analyse en essayant d'obtenir par des procédures *ad hoc* des matrices définies positives.

Pourtant, on rencontre parfois des valeurs propres négatives qui ne se cachent pas ! (Figure 1.22 et Figure 1.23) Dans le premier exemple (Figure 1.22), la croissance et l'alternance sont deux composantes de la variabilité. L'analyse spatiale les sépare clairement et identifie le groupe des croissances les plus régulières (1,10,12,15 et 16) et des alternances les plus marquées (4,9,5,13,8,11,20,13). A travers cet exemple, on met également en évidence la

faiblesse du test multivarié pris en défaut par un mélange de variables (dans le même tableau) respectivement à variance locale forte et à autocorrélation spatiale forte.

Dans le deuxième exemple également (Figure 1.23), on constate que l'analyse spatiale sépare explicitement les composantes spatiales selon le signe de leur autocorrélation alors que l'analyse classique les mélange allégrement. Cette fois-ci, la variabilité locale est la composante essentielle de la variabilité.

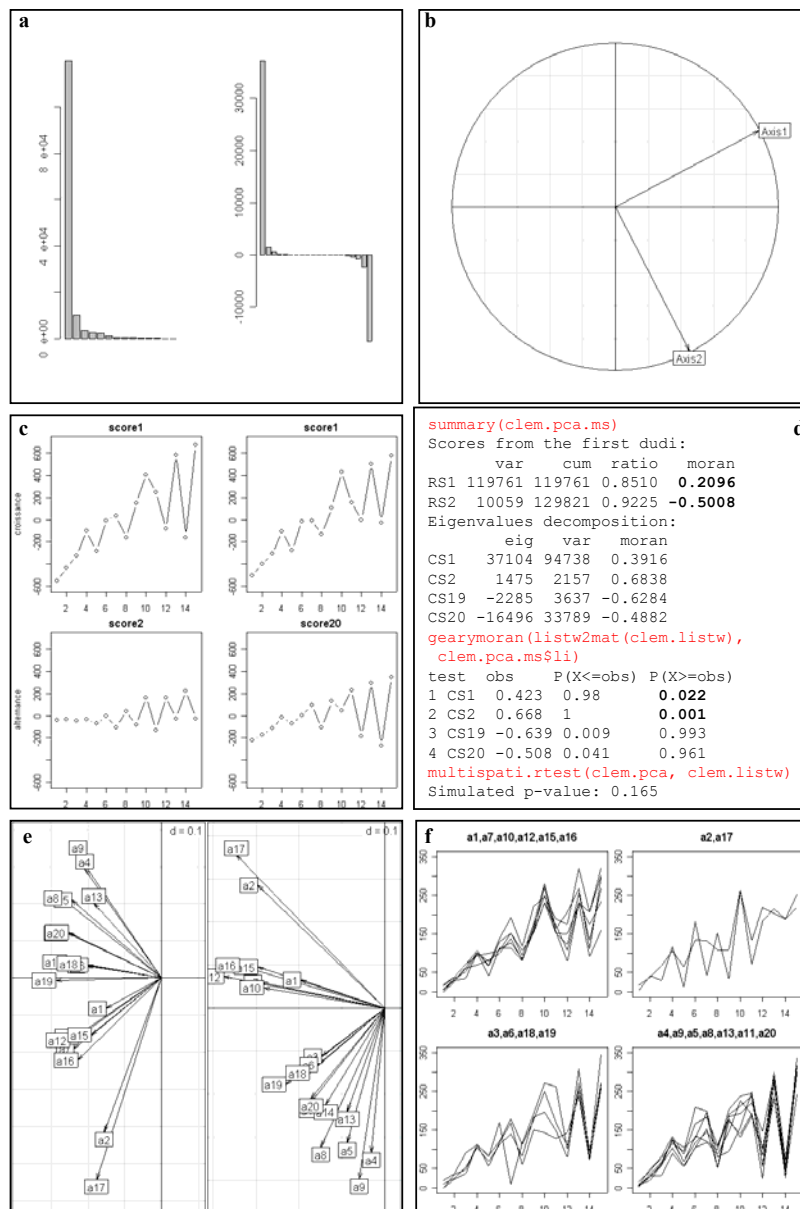


Figure 1.22 : Valeurs propres de l'ACP du jeu de données clementines (Annexe 1.6) (à gauche) et de l'analyse sous contrainte (à droite) (a). Projection des axes principaux de l'ACP sur les axes principaux de l'analyse sous contrainte (b). Scores de l'ACP (à gauche) et de l'analyse sous contrainte (à droite) (c). Listing (d). Projection des variables sur les axes principaux de l'ACP (à gauche) et de l'analyse sous contrainte (à droite) (e). Regroupement des chroniques d'évolution de la production des clémentiniers en fonction de leur position sur le plan des axes principaux (f).

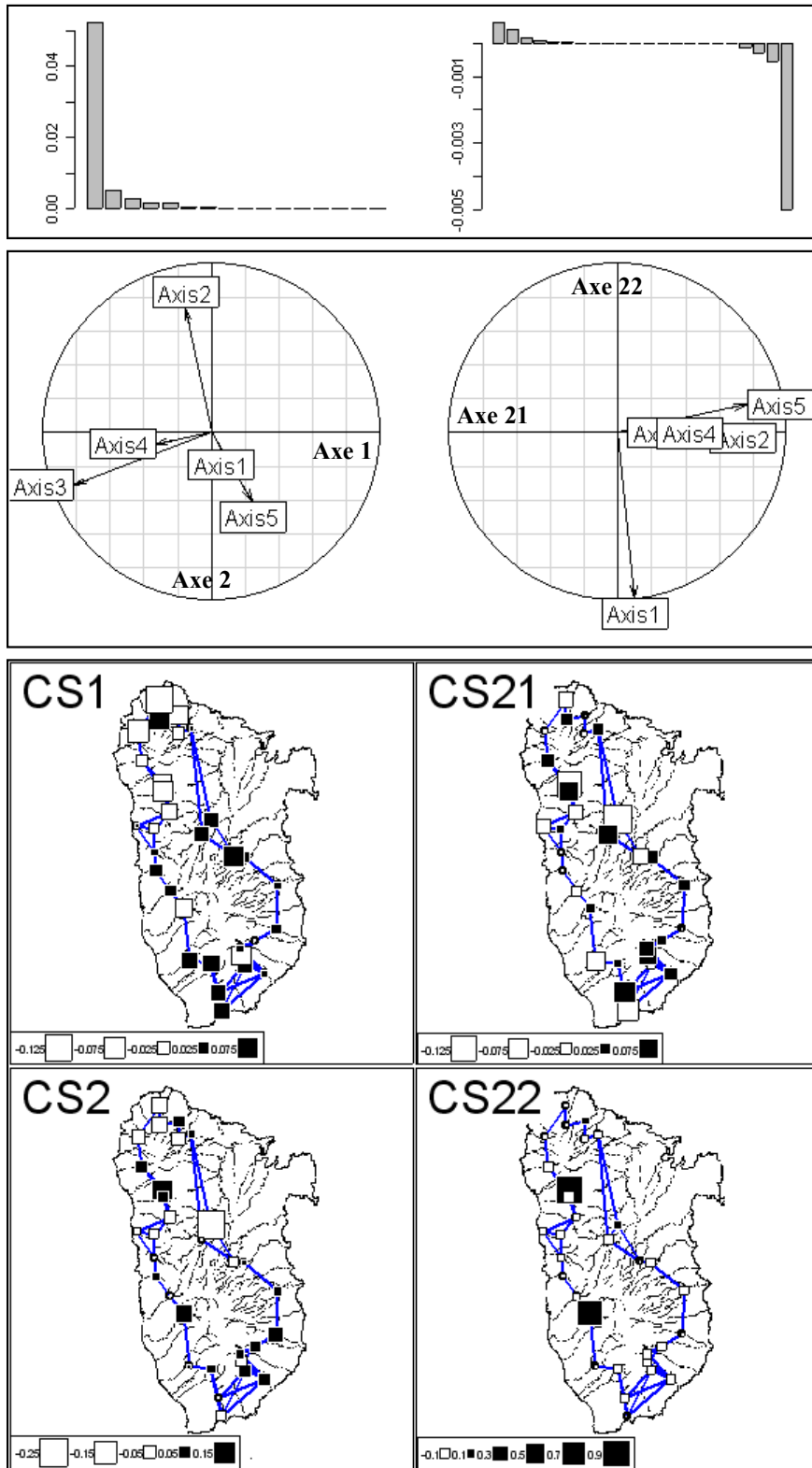


Figure 1.23 : Valeurs propres de l'ACP du jeu de données atya\$gen (Annexe 1.2) (à gauche) et de l'analyse sous contrainte (à droite) (en haut). Projection des axes principaux de l'ACP sur les axes de l'analyse sous contrainte (au milieu). Cartographie des scores de l'analyse sous contrainte (en bas).

7. DISCUSSION ET PERSPECTIVES

L'analyse sous contrainte simplifie la lecture des résultats lorsque l'information spatiale est mélangée à une information de nature différente. De plus, elle assure la séparation entre les composantes cartographiables et les composantes locales. L'analyse sous contrainte spatiale est donc une analyse d'inertie fortement orientée dans l'interprétation vers la lecture de l'autocorrélation. Cette logique rejoint assez fortement la logique des méthodes 'varimax' (Kaiser, 1958) qui recherchent une rotation des axes principaux qui facilite l'interprétation et la lecture des cercles de corrélations (Figure 1.24). Parfois, cette transformation est suffisante pour retrouver explicitement les composantes cartographiables (Goovaerts, 1992).

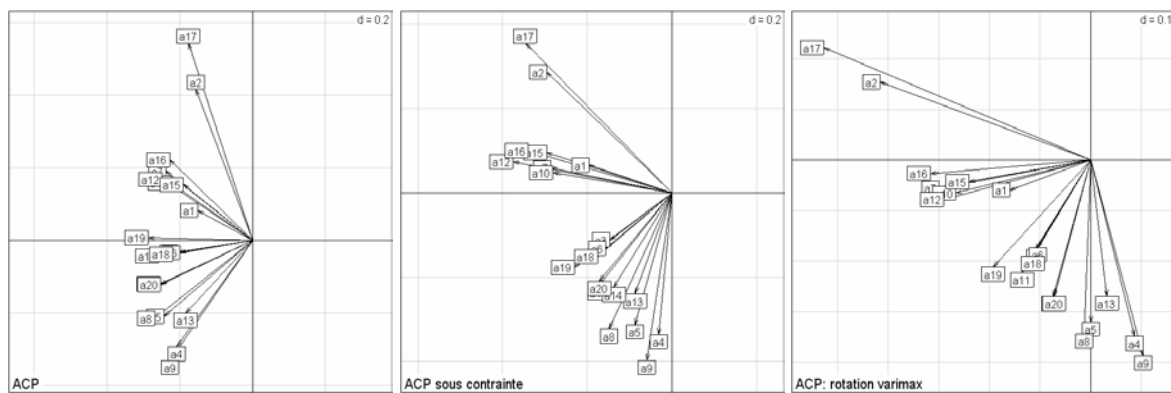


Figure 1.24 : Projections des variables sur les axes principaux de l'ACP (à gauche), sur les axes principaux de l'ACP sous contrainte (au milieu), sur les axes principaux de l'ACP après rotation varimax.

On a vu par ailleurs qu'il existait diverses options pour définir le voisinage, pondérer les relations de voisinage, intégrer le point de vue du voisinage dans le contexte des analyses multivariées. La multiplicité des options possibles implique une multiplicité des analyses sous contraintes envisageables. On en a fixé une, tout en laissant libre cours quant à l'analyse multivariée envisagée. Ce choix n'est pas arbitraire et il a été justifié au cours de la présentation. La version 'Moran' a été préférée à la version 'Geary' car on a considéré qu'elle répondait aux besoins les plus répandus des expérimentateurs, à savoir rechercher les composantes les plus cartographiables. On rentre alors au cœur des problèmes posés par la biométrie, en particulier des relations entre modèles et données. A quoi servira un modèle élégant du point de vue mathématique s'il ne répond pas directement au besoin de ses utilisateurs ? La solution la plus réjouissante du point de vue mathématique n'est pas forcément la plus réjouissante du point de vue des objectifs auxquels elle est sensée répondre. Pierre Delattre, dans sa réflexion sur la mathématique en tant que langage interdisciplinaire pose parfaitement bien le problème. « Dans la mesure où la mathématique constitue

effectivement le langage rationnel le plus précis qui soit à notre disposition, le but ultime de toute science est bien de parvenir à s'exprimer sous cette forme. Mais ce qui est proprement mathématique ne constitue en fait qu'une syntaxe. Lorsque la mathématique opère sur des équations, elle fait abstraction de la signification particulière des variables et des paramètres, sauf en ce qui concerne leur appartenance à certaines grandes catégories (grandeurs scalaires, vectorielles, tensorielles, opérateurs divers, etc.), au même titre que la syntaxe d'une langue ne tient compte que des catégories auxquelles appartiennent les mots (substantif, verbe, adjectif, etc.). La sémantique du langage interdisciplinaire se situe dans la justification de la mise en équations. La connaissance de la syntaxe d'une langue ne suffit pas pour exprimer dans cette langue des choses intelligibles ou intéressantes. Si l'on néglige la sémantique, on risque d'aboutir au calembour logique, comme cela peut arriver lorsqu'on se livre à une mathématisation prématurée ou abusive, c'est-à-dire insuffisamment justifiée au niveau épistémologique ». Toutefois, il ne faut pas tomber dans le travers inverse de la statistique *ad hoc* en mettant à la disposition des utilisateurs des outils qui paraissent les mieux adaptés à leurs besoins sans qu'ils soient justifiés du point de vue mathématique. Une bonne pratique de la biométrie nécessite donc une bonne maîtrise du dialogue qui s'instaure entre théorie mathématique et données biologiques. Encore faut-il, pour que ce dialogue se crée, que le biométricien soit confronté aux problèmes que se posent les biologistes. Il peut l'être soit directement, dans le cadre d'une consultation statistique ou indirectement en essayant de répondre aux préoccupations générales des biologistes qu'il peut appréhender au travers de son expérience, de la littérature et des données disponibles servant d'illustrations.

Ce travail est une illustration de l'importance et de l'efficacité du développement coopératif pour le biométricien. La fonction `multispati(...)` (Annexe 2.12) intègre en effet à la fois des outils développés dans le champ de la statistique spatiale et des outils développés dans celui de l'analyse de données. Elle permet la généralisation des concepts développés dans le champ de la statistique spatiale (par exemple, le scatterplot de Moran multivarié et une généralisation de l'usage du scatterplot de Moran univarié), et réciproquement de ceux développés dans le champ de l'analyse de données (le scatterplot de Moran multivarié est également une généralisation de l'usage de la carte factorielle). Ce développement combiné n'est possible que dans la mesure où les deux champs coexistent dans le même environnement. En plaçant ces deux bibliothèques *a priori* indépendantes dans le même environnement, cela assure l'émergence de nouveaux outils à l'interface entre les deux disciplines. En effet, cela contribue à oublier les logiques internes très différentes selon les disciplines (logique

géométrique et algébrique de l'analyse de données, logique probabiliste de , logique combinatoire de ...). Cette fois-ci, c'est le troisième élément du dialogue qui intervient comme catalyseur des pratiques de la biométrie. Le fait d'avoir un environnement gratuit, 'open source', réunissant sous le contrôle d'une autorité compétente les contributions de différents domaines de la statistique est à coup sûr une source immense de progrès pour la biométrie et la statistique de manière générale.

L'analyse sous contrainte est manifestement un outil appréciable pour l'analyse des données écologiques. Toutefois, son apport reste limité dans la mesure où chaque analyse sous contrainte n'intègre qu'une, voire deux échelles (globale, locale). On reprendra cette partie de la discussion de manière plus approfondie en discussion générale.

8. BIBLIOGRAPHIE

Abramovich, F., Bailey, T.C., & Sapatinas, T. (2003). Wavelet analysis and its applications.

Allain, C. & Cloitre, M. (1991) Characterizing the lacunarity of random and deterministic fractal sets. *Physical Review A*, 44, 3552-3557.

Anselin, L. (1996). The Moran scatterplot as an ESDA tool to assess local instability in spatial association. In *Spatial analytical perspectives on GIS* (eds M.M. Fischer, H.J. Scholten & D. Unwin), pp. 111-125. Taylor and Francis, London.

Anselin, L. & Hudak, S. (1992) Spatial econometrics in practice: A review of software options. *Regional Science and Urban Economics*, 22, 509-536.

Anselin, L., Syabri, I., & Smirnov, O. (2002) Visualizing multivariate spatial correlation with dynamically linked windows. In *CSISS Specialist Meeting on New Tools in Spatial Data Analysis* (eds L. Anselin & S.J. Rey), Santa Barbara, CA.

Aubry, P. (2000) Le traitement des variables régionalisées en écologie. Apports de la géomatique et de la géostatistique. Thèse de doctorat, Université Claude Bernard.

Aufaure, M.A., Yeh, L., & Zeitouni, K. (2000). Fouille de données spatiales. Ecole Thématique "Nouveaux défis en Sciences de l'Information : Documents & Evolution", Faculté des Sciences de Saint-Jérôme, Marseille.

Banet, T.A. & Lebart, L. (1984). Local and Partial Principal Component Analysis (PCA) and Correspondence Analysis (CA). In *COMPSTAT 84* (ed I.A.f.S. Computing.), pp. 113-123. Physica-Verlag, Vienna.

Bavaud, F. (1998) Models for spatial weights: a systematic look. *Geographical Analysis*, 50, 155-171.

Benali, H. & Escofier, B. (1990) Analyse factorielle lissée et analyse factorielle des différences locales. *Revue de Statistique Appliquée*, 38, 55-76.

Besse, P. (1979) Etude descriptive d'un processus ; approximation, interpolation. Thèse de 3^{ème} cycle, Université Paul Sabatier, Toulouse.

Blondel, J. (1985) *Biogéographie évolutive* Masson, Paris.

Borcard, D. & Legendre, P. (1994) Environmental control and spatial structure in ecological communities: an example using oribatid mites (Acari, Oribatei). *Environmental and Ecological Statistics*, 1, 37-61.

Borcard, D., Legendre, P., & Drapeau, P. (1992) Partialling out the spatial component of ecological variation. *Ecology*, 73, 1045-1055.

Chessel, D. & Mercier, P. (1993). Couplage de triplets statistiques et liaisons espèces-environnement. In *Biométrie et Environnement* (eds J.D. Lebreton & B. Asselain), pp. 15-44. Masson, Paris.

Chevenet, F., Dolédec, S., & Chessel, D. (1994) A fuzzy coding approach for the analysis of long-term ecological data. *Freshwater Biology*, 31, 295-309.

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

Conradsen, K., Nielsen, B.K., & Thyrssted, T.A. (1985) Comparison of min/max autocorrelation factor analysis and ordinary factor analysis. In *Proceedings from Symposium in Applied Statistics*, Vol. 47-56. Technical University of Denmark, Lyngby, Denmark.

Cornillon, P.-A., Amenta, P., & Sabatier, R. (1999). Three-way data arrays with double neighbourhood relations as a tool to analyze a contiguity structure. In *Classification and data analysis. Theory and Application* (eds M. Vichi & O. Opitz), pp. 263-270. Springer-Verlag, Berlin.

Cornillon, P.-A. & Sabatier, D. (1998). Local multivariate analysis. In *Advances in data science and classification* (eds A. Rizzi, M. Vichi & H.H. Bock). Springer.

Couteron, P. & Ollier, S. (sous presse) A generalized variogram-based framework for multiscale ordination. *Ecology*.

Cox, D.R. & Lewis, P.A.W. (1969) *L'analyse statistique des séries d'évènements* Traduction de Larrieu (J.) Dunod, Paris.

de Belair, G. (1981) *Biogéographie et aménagement : la plaine de La Mafragh* (Annaba, Algérie). Thèse de 3^o cycle. Université Paul Valéry, Montpellier.

Delattre, P. (1995) *Interdisciplinaires (recherches)*. Encyclopaedia Universalis, 12 (version CD-ROM 5.0 1999).

Dessier, A. & Laurec, A. (1978) *Le cycle annuel du zooplancton à Pointe-Noire (RP Congo)*. Description mathématique. *Oceanologica acta*, 1, 285-304.

Di Bella, G. & Jona-Lasinio, G. (1996) Including spatial contiguity information in the analysis of multispecific patterns. *Environmental and Ecological Statistics*, 3, 269-280.

Dolédec, S., Chessel, D., & Olivier, J.M. (1995) *L'analyse des correspondances décentrée: application aux peuplements ichtyologiques du haut-Rhône*. *Bulletin Français de la Pêche et de la Pisciculture*, 336, 29-40.

Durand, J.-D., Guinand, B., & Bouvet, Y. (1999) Local and global multivariate analysis of geographical mitochondrial DNA variation in *Leuciscus cephalus* L. 1758 (Pisces: Cyprinidae) in the Balkan Peninsula. *Biological Journal of the Linnean Society*, 67, 19-42.

Ersbll, B.K. (1989) *Transformations and classifications of remotely sensed data*. Ph.D. thesis, University of Denmark, Lyngby.

Escoufier, Y. (1987). The duality diagramm : a means of better practical applications. In *Development in numerical ecology* (eds P. Legendre & L. Legendre), pp. 139-156. NATO advanced Institute , Serie G .Springer Verlag, Berlin.

Estève, J. (1978). Les méthodes d'ordination : éléments pour une discussion. In *Biométrie et Ecologie* (eds J.M. Legay & R. Tomassone), pp. 223-250. Société Française de Biométrie, Paris.

Faraj, A. & Cailly, F. (2001) Spatial contiguity analysis: a method for describing spatial structures of seismic data. *Journal of Petroleum Science and Engineering*, 31, 93–111.

Fievet, E., Eppe, F., & Dolédec, S. (2001). Etude de la variabilité morphométrique et génétique des populations de *Cacadors* (*Atya innocous* et *Atya scabra*) de l'île de Basse-Terre. Direction Régionale de L'Environnement Guadeloupe, Laboratoire des hydrosystèmes fluviaux, Université Lyon 1, 43 Bd du 11 Novembre 1918, 69622, Villeurbanne cedex, France.

Gabriel, K.R. & Sokal, R.R. (1969) A new statistical approach to geographic variation analysis. *Systematic Zoology*, 18, 259-278.

Geary, R.C. (1954) The contiguity ratio and statistical mapping. *The incorporated Statistician*, 5, 115-145.

Ghertsos, K., Luczak, C., & Dauvin, J.-C. (2001) Identification of global and local components of spatial structure of marine benthic communities: example from the Bay of Seine (Eastern English Channel). *Journal of Sea Research*, 45, 63-77.

Gittleman, J.L. & Kot, M. (1990) Adaptation: statistics and a null model for estimating phylogenetic effects. *Systematic Zoology*, 39, 227-241.

Goodall, D.W. (1954) Objective methods for the classification of vegetation III. An essay in the use of factor analysis. *Australian Journal of Botany*, 2, 304-324.

Goovaerts, P. (1992) Multivariate geostatistical tools for studying scale-dependent correlation structures and describing space-time variation, Thèse de doctorat, Université Catholique de Louvain , Louvain la Neuve.

Goulard, M., Voltz, M., & Monestiez, P. (1987) Comparaison d'approches multivariées pour l'étude de la variabilité spatiale des sols. *Agronomie*, 7, 657-665.

Greenacre, M.J. (1984) Theory and applications of correspondence analysis Academic Press, London.

Grunsky, E.C. (2002) R: a data analysis and statistical programming environment—an emerging tool for the geosciences. *Computers & Geosciences*, 28, 1219-1222.

Grunsky, E.C. & Agterberg, F.P. (1988) Spatial and multivariate analysis of geochemical data from metavolcanic rocks in the Ben Nevis area, Ontario. *Mathematical Geology*, 20, 825-861.

Grunsky, E.C. & Agterberg, F.P. (1989) The application of spatial factor analysis to unconditional simulations with implications for mineral exploration. In Proceedings, 21st International Symposium on Computers in the Mineral Industry, pp. 194-208. Society of Mining Engineers of AIME, Littleton, Colorado, Las Vegas, Nevada, March 1989.

Grunsky, E.C. & Agterberg, F.P. (1991) SPFA: a FORTRAN-77 program for spatial factor analysis of multivariate data. *Computers & Geosciences*, 17, 133-160.

Grunsky, E.C., Chen, Q., & Agterberg, F.P. (1996). Applications of spatial factor analysis to multivariate data. In *Geologic Modeling and Mapping* (eds A. Foerster & D.F. Merriams), pp. 229-261. Plenum, New York.

Hatheway, W.H. (1971). Contingency table analysis of rain forest vegetation. In *Statistical Ecology. III Many species populations ecosystems and systems analysis* (eds G.P. Patil, E.C. Pielou & W.E. Waters), pp. 271-314. Pennsylvania State University Press.

Hill, M.O. (1974) Correspondence analysis : A neglected multivariate method. *Journal of the Royal Statistical Society, C*, 23, 340-354.

Hill, M.O. & Smith, A.J.E. (1976) Principal component analysis of taxonomic data with multi-state discrete characters. *Taxon*, 25, 249-255.

Ihaka, R. & Gentleman, R. (1996) R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5, 299-314.

Jayet, H. (1999) *Analyse spatiale quantitative, une introduction* Hermes.

Kaiser, H.F. (1958) The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23, 187-200.

Kiers, H.A.L. (1994) Simple structure in component analysis techniques for mixtures of qualitative and quantitative variables. *Psychometrika*, 56, 197-212.

Kroonenberg, P.M. & Lombardo, R. (1999) Nonsymmetric correspondence analysis: a tool for analysing contingency tables with a dependence structure. *Multivariate Behavioral Research*, 34, 367-396.

Le Foll, Y. (1982) Pondération des distances en analyse factorielle. *Statistique et Analyse des données*, 7, 13-31.

Lebart, L. (1969) *Analyse statistique de la contiguïté*. Publication de l'Institut de Statistiques de l'Université de Paris, 28, 81-112.

Light, R.J. & Margolin, B.H. (1971) An analysis of variance for categorical data. *Journal of the American Statistical Association*, 66, 534-544.

Méot, A., Chessel, D., & Sabatier, R. (1993). Opérateurs de voisinage et analyse des données spatio-temporelles. In *Biométrie et Environnement* (eds J.D. Lebreton & B. Asselain), pp. 45-72. Masson, Paris.

Mom, A. (1998) Eigenstructure of distance matrices with an equal distance subset. *Linear Algebra and its Applications*, 280, 245-251.

Monestiez, P. (1978). Méthodes de classification automatique sous contraintes spatiales. In *Biométrie et Ecologie* (eds J.M. Legay & R. Tomassone), pp. 367-379. Société Française de Biométrie, Paris.

Monestiez, P., Goulard, M., & Charmet, G. (1994) Geostatistics for spatial genetic structures: study of wild populations of perennial ryegrass. *Theoretical and applied genetics*, 88, 33-41.

Moran, P.A.P. (1948) The interpretation of statistical maps. *Journal of the Royal Statistical Society, B*, 10, 243-251.

Moran, P.A.P. (1950) Notes on continuous stochastic phenomena. *Biometrika*, 37, 17-23.

Nielsen, A.A. (1995a) Change detection in multi-spectral, bi-temporal spatial data using orthogonal transformations. In <http://citeseer.nj.nec.com/63505.html>.

Nielsen, A.A. (1995b) Multi-channel remote sensing data and orthogonal transformations for change detection. In <http://citeseer.nj.nec.com/56095.html>.

Nielsen, A.A. (1999) C04351 Statistical Image Analysis, Spring 1999 Orthogonal Transformations. In <http://citeseer.nj.nec.com/428248.html>.

Nielsen, A.A. & Conradsen, K. (1997) Multivariate alteration detection (MAD) in multispectral, bi-temporal image data: a new approach to change detection studies. In <http://www.imm.dtu.dk/~aa/tech-rep-1997-11/>. Tech. rep. 199711, Department of Mathematical Modelling, Technical University of Denmark.

Nielsen, A.A., Conradsen, K., Pedersen, J.L., & Steinfeld, A. (1997) Spatial factor analysis of stream sediment geochemistry data from South Greenland. In *Proceedings of the Third Annual Conference of the International Association for Mathematical Geology* (ed V. Pawlowsky-Glahn), pp. 955-960, Barcelona, Spain.

Nielsen, A.A., Conradsen, K., & Simpson, J.J. (1998) Multivariate alteration detection (MAD) and MAF post-processing in multispectral, bi-temporal image data: new approaches to change detection studies. *Remote Sensing of Environment*, 64, 1-19.

Nielsen, A.A. & Larsen, R. (1994) Restoration of Geris data using the maximum noise fractions transform. In *First International Airborne Remote Sensing Conference and Exhibition*, Strasbourg, France, 11–15 September 1994.

Noy-Meir, I. & Anderson, D.J. (1971). Multivariate pattern analysis, or multiscale ordination: towards a vegetation hologram ? In *Statistical Ecology, III Many species populations ecosystems and systems analysis* (eds G.P. Patil, E.C. Pielou & W.E. Waters), pp. 208-231. Pennsylvania State University Press.

Ollier, S., Dray, S., & Chessel, D. (soumis) Taking into account spatial dependence in multivariate analysis: a generalization of Wartenberg's multivariate spatial correlation. *Geographical Analysis*.

Pace, R.K. & Barry, R. (1997) Sparse spatial autoregressions. *Statistics and Probability Letters*, 33, 291-297.

Pace, R.K. & LeSage, J.P. (2002) Semiparametric maximum likelihood estimates of spatial dependence. *Geographical Analysis*, 34, 76-90.

Pace, R.K. & LeSage, J.P. (2003) Conditional autoregressions with doubly stochastic weight matrices.

Pace, R.K. & Zou, D. (2000) Closed-form maximum likelihood estimates of nearest neighbor spatial dependence. *Geographical Analysis*, 32, 154-172.

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

Royer, J.J. (1984) Proximity analysis: a method for multivariate geodata processing. Application to geochemical processing. *Sciences de la Terre, Série informatique* 20, 585-591.

Sandjivy, L. & Galli, A. (1984) Analyse krigeante et analyse spectrale. *Science de la Terre, Série Informatique*, 21, 115-124.

Smouse, P. & Peakall, R. (1999) Spatial autocorrelation analysis of individual multiallele and multilocus genetic structure. *Heredity*, 82, 561-573.

Solow, A.R. (1994) Detecting change in the composition of a multispecies community. *Biometrics*, 50, 556-565.

Switzer, P. & Green, A.A. (1984). Min/max autocorrelation factors for multivariate spatial imagery. Tech. rep. 6, Stanford University.

Tenenhaus, M. & Young, F.W. (1985) An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical multivariate data. *Psychometrika*, 50, 91-119.

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

Tiefelsdorf, M., Griffith, D.A., & Boots, B. (1999) A variance-stabilizing coding scheme for spatial link matrices. *Environment and Planning A*, 31, 165-180.

Tisné-Agostini, D. (1988) Description par analyse en composantes principales de l'évolution de la production du clémentinier en association avec 12 types de porte-greffe. Rapport technique, DEA Analyse et modélisation des systèmes biologiques, Université Lyon 1.

Upton, G. & Fingleton, B. (1985) Spatial data analysis by example. Vol. 1: Point pattern and quantitative data John Wiley & Sons, Chichester.

Ver Hoef, J.M. & Glenn-Lewin, C.G. (1989) Multiscale ordination: a method for detecting pattern at several scales. *Vegetatio*, 82, 59-67.

Wackernagel, H. (2003) Multivariate geostatistics. An introduction with applications, Third edition edn. Springer.

Wagner, H.H. (2003) Spatial covariance in plant communities: integrating ordination, geostatistics, and variance testing. *Ecology*, 84, 1045-1057.

Wagner, H.H. (2004) Direct multi-scale ordination with canonical correspondence analysis. *Ecology*, 85, 342-351.

Wartenberg, D.E. (1985) Multivariate spatial correlations: a method for exploratory geographical analysis. *Geographical Analysis*, 17, 263-283.

TYPOLOGIE DE STRUCTURES MULTIÉCHELLES UNIVARIÉES

Développement méthodologique à partir d'une consultation statistique

1.	INTRODUCTION.....	79
2.	DONNÉES D'ALTIMÉTRIE LASER	80
2.1.	Contexte	80
2.2.	Description de l'expérience.....	81
2.3.	Les données.....	84
3.	STRUCTURE D'UNE VARIABLE QUANTITATIVE	85
4.	FAMILLES DE K FORMES BILINÉAIRES SYMÉTRIQUES.....	89
4.1.	Définitions.....	89
4.2.	La classe d'objets 'kfbs'	90
4.3.	Formes de Geary/Lebart : le variogramme	91
4.4.	Formes de Moran/Smouse : le corrélogramme	94
4.5.	Formes de Greig-Smith/Noy-Meir : les msbs	96
4.6.	Formes de Hill : les ttlv	101
4.7.	Typologie d'un ensemble de formes bilinéaires	105
5.	BASES ORTHONORMÉES ET FAMILLES DE K PROJECTEURS	107
5.1.	Définitions.....	107
5.2.	La classe d'objets 'orthobasis'	109
5.3.	Les bases associées à la diagonalisation des matrices symétriques	110
5.4.	Expression analytique des vecteurs propres de l'opérateur de Méot	115
5.5.	La base associée à l'analyse spectrale à une dimension.....	122
5.6.	Les bases d'ondelettes à une dimension.....	124
6.	NORMALISATION DES FORMES BILINÉAIRES	129
6.1.	Introduction.....	129
6.2.	Définitions.....	130
6.3.	Typologie de structures	137
7.	APPLICATIONS AUX DONNÉES D'ALTIMÉTRIE LASER	139
8.	DISCUSSION ET PERSPECTIVES	139
9.	BIBLIOGRAPHIE	140

1. INTRODUCTION

La motivation initiale de ce chapitre réside dans l'analyse des données proposées par Raphaël Pélissier (UMR AMAP, Montpellier) lors de mon DEA. Les données, dont nous présenterons les caractéristiques, correspondent à des transects d'altimétrie laser dont nous voulions déterminer la structure spatiale pour en faire une typologie (Figure 2.4). Bien que la caractérisation de la structure spatiale et temporelle d'une variable mesurée le long d'un transect ait largement été étudiée (Dale et al., 2002), le problème posé par la typologie de plusieurs variables sur la base de leur structure interne n'avait pratiquement jamais été envisagée sur une base systématique. A en croire les questions posées sur le forum d'ai-geostats (<http://www.ai-geostats.org>), dédié à l'analyse des données spatiales, ce n'est pourtant pas la première fois que ce problème se pose dans le champ expérimental.

L'expérience conduite par Ann Zumwalt et les problèmes qu'elle pose constituent une bonne illustration: « *I am a graduate student studying the functional morphology of bones. Part of my thesis entails **characterizing the shape of a relatively complex 3D bone surface.** » On est bien dans un problème de caractérisation d'une structure. « *I am testing to see whether exercise affects the morphology of this surface, so I am looking for a way to test for differences between shapes/specimens.* » On comprend alors que la caractérisation de la structure n'est pas une fin en soi mais une étape préliminaire à l'étude des variations de la structure en fonction des conditions expérimentales. Ann Zumwalt précise ensuite la nature de ces données : « *I have 3D grid data (x,y,z) that represents the surfaces (I am scanning the bones with a 3D laser scanner to obtain this data). Can any of you suggest methods to analyze this data that will allow me to differentiate surfaces that are morphologically dissimilar?* ». En réponse, Wilner River propose de coupler une analyse de rugosité à une analyse discriminante. Le problème est bien un problème de typologie de structures spatiales. C'est un problème de nature spatiale et multivarié mais l'on comprend bien qu'il soit radicalement différent de l'analyse sous contrainte définie au chapitre précédent.*

Kalle Kronholm expose un problème fort semblable : « *I am studying the spatial variability of penetration resistance (a proxy for strength) in snow layers in an Alpine snow cover. Are weak layers that are responsible for snow avalanche release, less spatially variable than layers that are not critical for snow stability?* ». Ici encore il s'agit de caractériser les structures spatiales d'unités statistiques dans le but de les comparer entre elles. Il précise ensuite la nature de ses données ainsi que le traitement qu'il envisage : « *At 113 locations, measurements of penetration resistance for each layer were made in a nested grid.*

I have data from approximately 100 layers. I fit a spherical model semivariogram for each layer to the experimental semivariograms. Is it possible to compare directly the range, the sill and the nugget of the spherical model semivariograms fitted for each layer? Are there any pit-falls that I should be aware of? ». Il pose explicitement le problème de comparaison multiple de variogrammes, mais ce problème n'a visiblement jamais été abordé d'un point de vue méthodologique.

Enfin, Jennifer Dickie donne un dernier exemple: « *I'm looking at the spatial distribution of soil properties across different vegetation types. I've sampled a total of 11 plots at three different nested spatial scales. Is there a way I can characterise the variability at each spatial scale that would allow me to compare both between scales for each plot and between the plots? ».*

Quelques exemples existent également dans la littérature : Bohte (1980) définit une classification de séries temporelles à partir de leurs corrélogrammes respectifs. Qu et al. (2003) proposent une analyse discriminante des coefficients d'ondelettes pour discriminer des spectres de masse de protéines, et Coutron (2002) envisage la typologie de photographies aériennes sur la base de leur analyse spectrale par transformée de Fourier. Dans chacune de ces publications, le problème de la typologie de structures est abordé de manière très spécifique, avec pour objectif de traiter un type de données avec un type de méthode.

Dans ce chapitre, on aborde le problème d'une manière plus générale. Dans un premier paragraphe, on présente le contexte, les données et la problématique en jeu qui m'ont conduit à étudier ces problèmes méthodologiques. Ensuite, on propose une revue des différentes méthodes d'analyse de la structure d'une variable à plusieurs échelles, en se limitant exclusivement, ou presque, à l'étude des variables mesurées le long d'un transect. Chaque méthode est réécrite sous forme matricielle, l'objectif étant d'une part de pouvoir les comparer les unes aux autres sur une base mathématique, et d'autre part de pouvoir connecter ces méthodes aux analyses multivariées. Enfin, le retour aux données permet d'évaluer la pertinence des méthodes proposées et de répondre aux questions thématiques posées.

2. DONNÉES D'ALTIMÉTRIE LASER

2.1. Contexte

La structure spatiale des peuplements forestiers, au sens d'organisation fonctionnelle des éléments constitutifs (Barbault, 1992), détermine l'environnement local et conditionne en partie les processus naturels de croissance, régénération, mortalité. Réciproquement, elle est

l'expression intégrée de ces différents processus, dans le temps (facteurs historiques) et dans l'espace (Barbault, 1992). De fait, l'analyse des structures spatiales est à la fois un problème difficile et central de la phyto-écologie (Greig-Smith, 1952, 1961; Watt, 1947), qui soulève de multiples questions méthodologiques. Pour des raisons liées au coût d'acquisition des données en forêt tropicale humide, l'étude des structures spatiales a jusqu'à présent privilégié les échelles locales (parcelles, transects) et phytogéographiques (notamment à partir de la compilation de données d'herbier, voir par exemple Gimaret-Carpentier (1999)). L'exploration des échelles intermédiaires, comme les relations entre les différents niveaux d'observation, n'en sont qu'à leurs balbutiements (Brown & Maurer, 1989). C'est pourquoi il est aujourd'hui nécessaire de développer des approches spatialisées qui permettent d'appréhender la structure spatiale des écosystèmes sur de grandes surfaces. L'analyse des structures spatiales à partir de données de télédétection à haute résolution spatiale telle que l'altimétrie laser, en fait partie. En effet, avec le développement plus ou moins récent des outils de la télédétection (Legay & Barbault, 1995), il est désormais possible d'appréhender les écosystèmes forestiers par l'étude de leur canopée, sur de grandes surfaces d'observation, avec une intensité d'échantillonnage suffisante (Weishampel et al., 1996). De nombreuses études ont démontré la pertinence de l'altimétrie laser pour estimer la hauteur et le volume ligneux d'un peuplement (Nelson, 1988; Ritchie et al., 1993; St-Onge, 1999), caractériser son architecture tridimensionnelle (Drake & Weishampel, 2000), et ainsi cartographier des types de peuplements forestiers (St-Onge et al., 1998).

2.2. Description de l'expérience

2.2.1. Principes

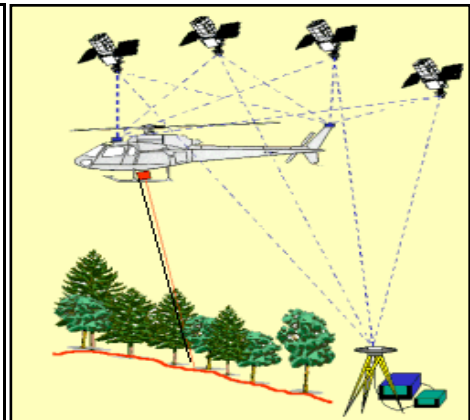
L'expérience proposée s'inscrit dans ce contexte. Il s'agit en fait d'une expérience 'rapportée' dans la mesure où l'on utilise des données collectées pour répondre à des objectifs complètement différents de ceux que l'on recherche. Le risque est grand car les objectifs sont censés définir *a priori* les contraintes du plan d'expérience mais les données sont alléchantes étant donné le prix à payer pour récolter des données de terrain en milieu tropical humide. Elle consiste à utiliser les données d'altimétrie laser, enregistrées par le BRGM afin d'évaluer un modèle numérique de terrain en Guyane française au travers d'une campagne aéroportée. Ces données sont susceptibles de contenir une information sur la structure de la canopée. Comme elles couvrent les trois quarts nord de la Guyane française, elles englobent potentiellement des échelles de variations, liées notamment à des variations floristiques et

structurales découlant de variations géomorphologiques, climatiques ou historiques, qui sont très difficiles à étudier aux travers de relevés de terrain, forcément peu étendus. Chercher à savoir si ce signal laser peut rendre compte de la variabilité des structures de canopée est donc une question pertinente.

L'ensemble des conditions expérimentales et les caractéristiques techniques de la campagne de géophysique aéroportée en Guyane française (caractéristiques de navigation, positionnement GPS, caractéristiques des capteurs laser et radar...) sont détaillées dans Delor et al. (1998). On rappelle brièvement les principes généraux de l'altimétrie laser (Figure 2.1).

L'altimétrie laser est construite sur le principe de la télédétection dite "active" (Weishampel et al., 1996). Un signal électromagnétique dont les caractéristiques (longueur d'onde, puissance) conditionnent les propriétés de la mesure est émis depuis un hélicoptère selon une fréquence et une taille de faisceau donnés. La réception de la portion réfléchi par tout objet intercepté constitue la deuxième phase nécessaire à l'acquisition de la mesure. La mesure du temps écoulé entre l'émission et la réception du signal donne une estimation de la distance séparant l'hélicoptère du premier objet sur lequel s'est réfléchi le signal. Par ailleurs, la position (coordonnées spatiales et altitude) de l'hélicoptère sont déterminées par triangulation. La combinaison des ces quatre informations permet d'estimer les profil des variations d'une surface au sol (géomorphologie, canopée,...).

Figure 2.1 : Principes de l'altimétrie laser



2.2.2. La base de données du BRGM

La base de données du BRGM donnant l'ensemble des valeurs du signal laser est structurée par lignes de vol parallèles orientées N30° et espacées entre elles de 500 m (Figure 2.2). Le laser utilisé par le BRGM avait une fréquence de 10 Hz, soit, étant donnée la vitesse de vol, un point de mesure tous les 7 m le long de chaque ligne de vol. L'altimètre laser réalise automatiquement un contrôle et une sélection des données en fonction de la qualité du signal réfléchi (intensité, temps de retour). La base de données renferme par conséquent des données manquantes mais elles sont peu nombreuses, généralement isolées et réparties aléatoirement.

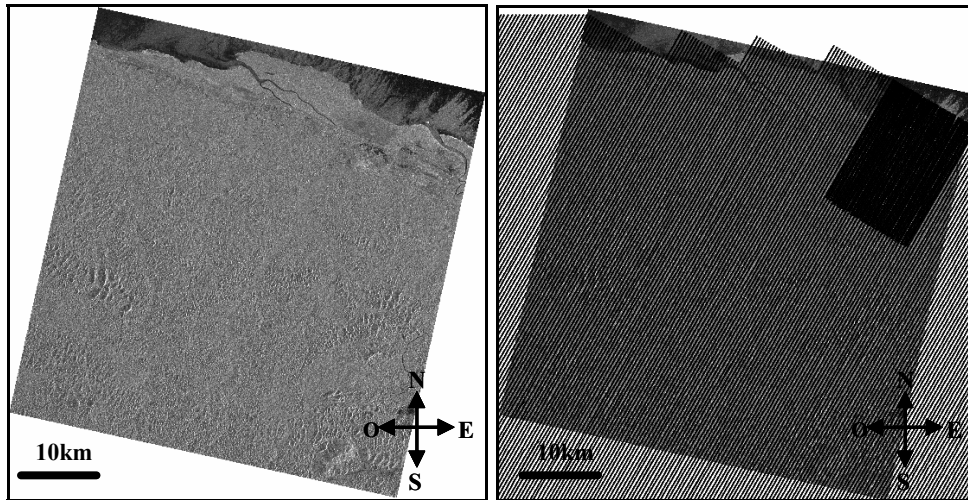


Figure 2.2 : Extrait d'une image radar d'origine inconnue recouvrant la partie nord du territoire guyanais. On reconnaît au nord la façade atlantique avec les embouchures des fleuves Counamama et Iracoubo. A droite, on a superposé les lignes de vol extraites de la base de données du BRGM. Elles sont orientées $N30^\circ$ et espacées de 500 mètres environ.

2.2.3. Plan d'expérience et échantillonnage

L'objectif poursuivi par le BRGM était d'estimer les variations de la topographie afin de réaliser un Model Numérique de Terrain de la Guyane française. Notre objectif est différent : il est d'évaluer si les variations du signal laser rendent compte de la variabilité environnementale à différentes échelles spatiales. A priori, il est certain que les variations topographiques que l'on aperçoit sur l'image radar vont ressortir puisque les données ont été enregistrées dans ce but précis. Ce qui nous intéresse, c'est de savoir si la variabilité du signal laser ne renferme pas d'autres signatures environnementales. Afin de pouvoir comparer la variabilité spatiale du signal laser aux caractéristiques environnementales, nous avons limité l'étude de la base de données à la forêt de Counami, pour laquelle les caractéristiques floristiques et géomorphologiques sont bien connues (Boyé et al., 1979; Couteron et al., 2002; Hutter, 2001; Milési et al., 1995). On s'est attaché en particulier aux variations géomorphologiques pour lesquelles il a été défini trois principales unités qui diffèrent de part leur altitude et leur complexité géomorphologique. On a extrait des 30 lignes qui recoupent la zone d'étude 264 transects de 64 mesures chacun (Figure 2.3). L'objectif était d'une part de caractériser la variabilité du signal laser de chaque transect à différentes échelles, d'autre part de vérifier si la variabilité de cette variabilité était corrélée à la variabilité environnementale.

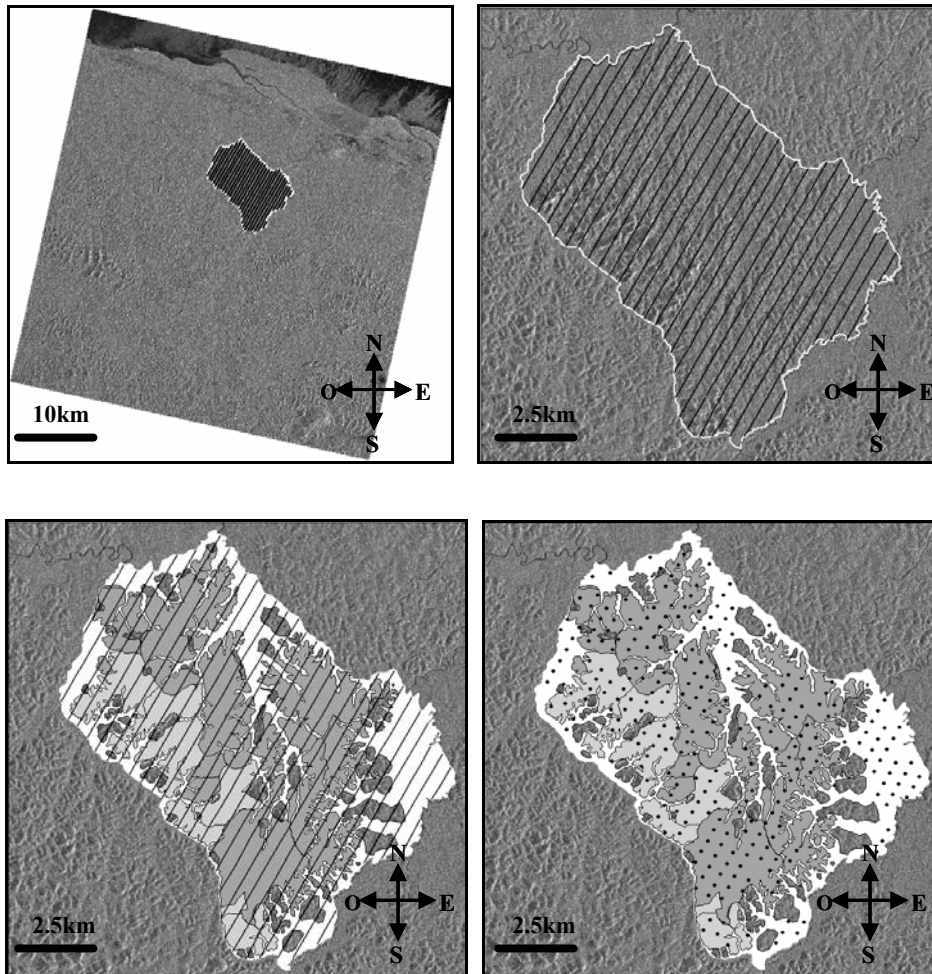


Figure 2.3 : Carte du site de COUNAMI avec les 30 lignes de vol recoupant la zone d'étude. Carte du site de COUNAMI avec les trois principales unités géomorphologiques. La première unité (en blanc) recouvre les plaines alluviales des trois principaux fleuves du site d'étude (COUNAMAMA, COUNAMI and IRACUOBO). Ces plaines alluviales sont marquées par l'absence de relief et un taux d'hydromorphie élevé. La seconde unité (en gris foncé) correspond aux reliefs peu marqués dont l'altitude s'élève au maximum à 60 mètres au-dessus du niveau de la mer. La dernière unité (en gris clair) est caractérisée par des reliefs beaucoup plus marqués dont les pentes sont plus importantes. Leur altitude s'élève généralement au-dessus de 60 mètres au-dessus du niveau de la mer. Carte du site de COUNAMI avec les 264 transects représentés sur la carte par un point correspondant à l'une de leurs extrémités.

2.3. Les données

Les données sont regroupées dans un tableau à 264 lignes et 64 colonnes. Chaque ligne correspond aux 64 mesures d'un transect (Figure 2.4). A la marge du tableau est associé un facteur à trois modalités définissant l'unité géomorphologique de chacun des transects (Annexe 1.12). Cette figure pose clairement le problème méthodologique de la comparaison multiple de structures. De manière plus générale, est posée la question de l'étude de variabilité de la variabilité dans un plan d'expérience donné : on cherche à savoir si la structure du signal laser varie en fonction des types géomorphologiques ? Afin de répondre à cette question, on s'est d'abord intéressé aux différentes approches permettant de décrire et

tester l'existence d'une structure à différentes échelles pour une variable mesurée le long d'un transect.

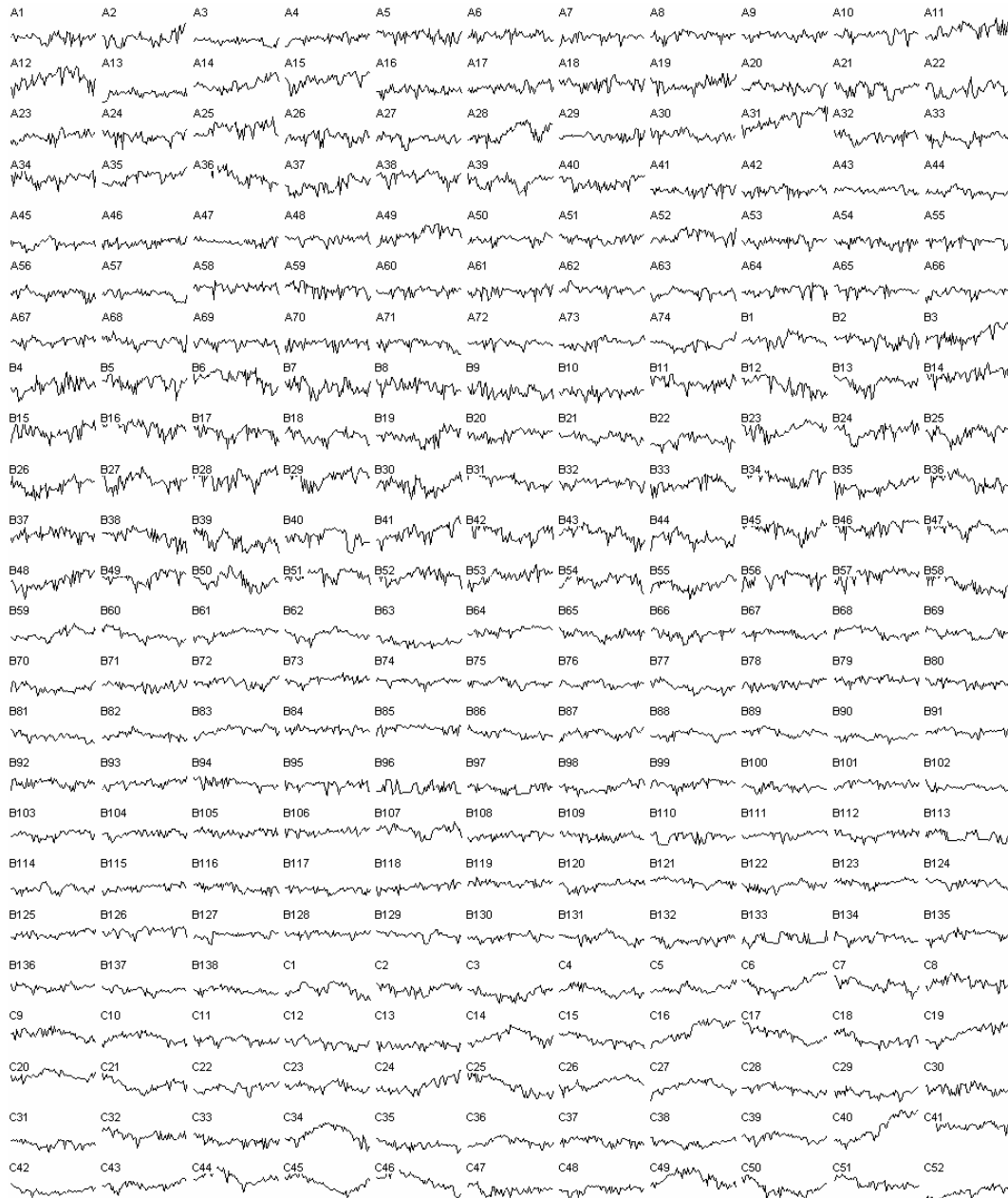


Figure 2.4 : Représentation des 264 transects après normalisation. Chaque transect est situé exclusivement sur une des trois unités géomorphologiques (A, B, C)

3. STRUCTURE D'UNE VARIABLE QUANTITATIVE

L'analyse des structures spatiales et temporelles d'une variable à plusieurs échelles a donné lieu au développement de nombreuses méthodes statistiques (pour une introduction, voir par exemple les articles de (Brillinger et al., 2002; Fortin et al., 2002; Guttorp et al., 2002; Percival, 2003; et plus généralement l'ensemble des articles traitant de ce sujet dans l'Encyclopedia of Environmetrics). Cette diversité s'explique par la grande variété des

supports de mesures ((Perry et al., 2002), Figure 2.5), la multiplicité des manières de définir et d'introduire la notion d'échelle ((Dungan et al., 2002), Figure 2.58) et la diversité des processus étudiés (Hill, 1973).

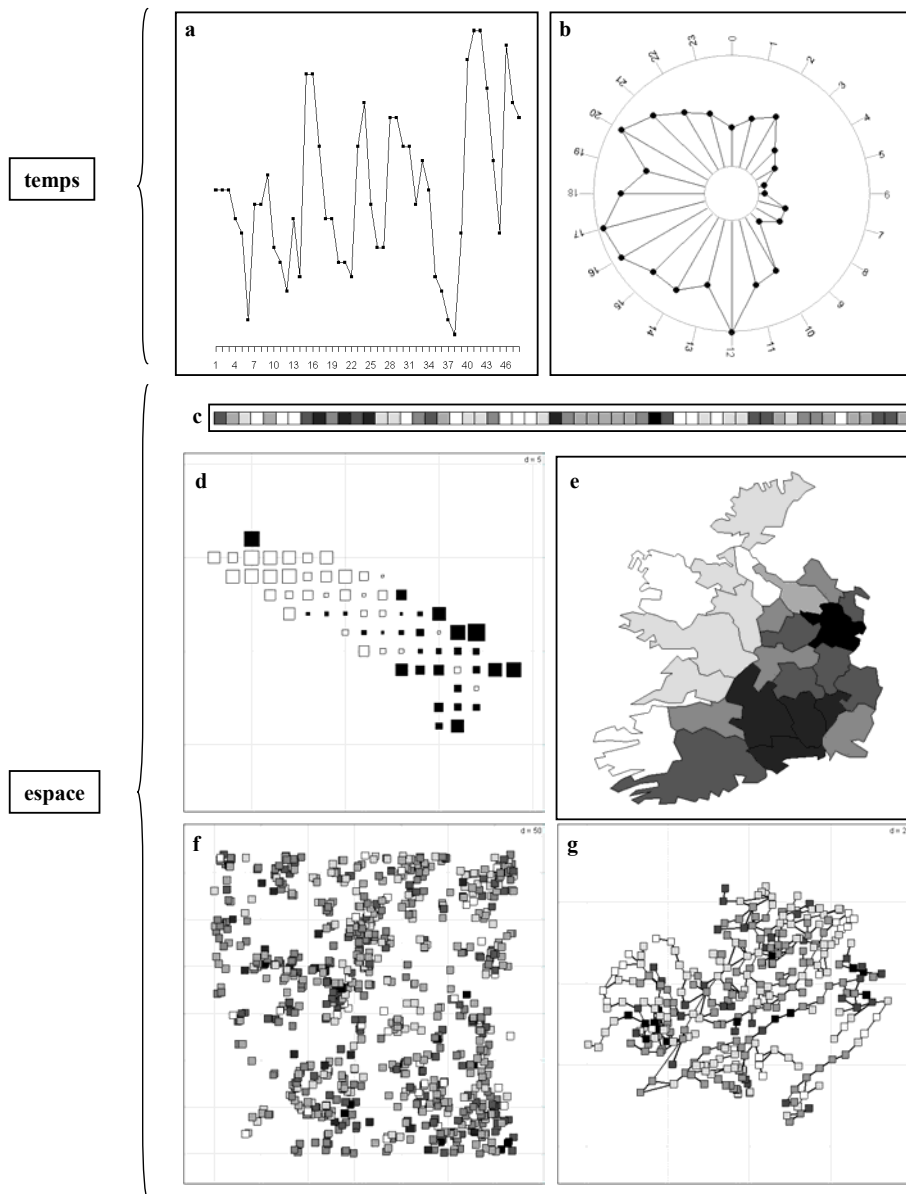


Figure 2.5 : Diversité des supports de mesures. **a :** exemple de série temporelle régulière : évolution du taux d'hormone lutéinisante dans le sang mesuré toutes les 10 minutes pendant 8 heures (Diggle, 1990). **b :** exemple de série temporelle circulaire : évolution du nombre d'entrants au service des urgences mesuré chaque heure pendant une journée (Fisher, 1993). **c :** exemple de mesures spatialisées à une dimension et régulièrement espacées : abondance de l'espèce *Gardenië Sokotensis* mesurée le long d'un transect de 56 placettes (Couteron et al., 1996). **d :** exemple de mesures spatialisées à deux dimensions sur une grille régulière : taux de chlorophylle A mesuré pour 63 sites de l'étang de Thau sur une grille régulière dont la maille est de 1 kilomètre (Borcard et al., 2004). **e :** exemple de mesures spatialisées à deux dimensions sur des unités surfaciques : nombre de voitures pour mille habitants mesuré pour les 25 comtés d'Irlande (Geary, 1954). **f :** exemple de processus ponctuel marqué (Renshaw, 2002) : la marque correspond à la hauteur des arbres de l'espèce *Combretum micranthum* en savanne arborée (Couteron, 2001). **g :** exemple de mesures spatialisées à deux dimensions sur un graphe : indice de la qualité de l'eau mesuré sur 295 tronçons du bassin de la Haute Saône (Hérissé, 2001).

De plus, ces méthodes ont généralement été développées dans des champs scientifiques variés, en intégrant les préoccupations, les contraintes et le langage spécifiques à chaque discipline. Leur utilisation dans un champ scientifique différent tel que l'écologie statistique nécessite donc une réécriture en fonction des contraintes propres à cette discipline. C'est ce que fait remarquer Hill (1973) au sujet de l'analyse spectrale : « *spectral analysis has been obscured by its use in electronic engineering and control systems. There is, for example, a blaffing concern with smoothing techniques which has been deplored by Bartlett (1967)... In ecological contexts, spectral analysis bears no particularly important relation to any presumed underlying structure of the data and must be regarded as merely another method of pattern analysis* ». Le transfert d'une discipline vers une autre, en l'occurrence l'écologie statistique, a déjà été maintes fois proposée (voir par exemple l'article de Dale et Mah (1998), sur l'introduction des ondelettes en écologie statistique ; voir également l'article de Renshaw (1997)) sur l'introduction de l'analyse spectrale bidimensionnelle en biométrie forestière). On assiste alors à la multiplication des choix possibles pour l'utilisateur, ce qui lui rend la tâche plus difficile car il ne dispose pas des moyens nécessaires pour ordonner ces différentes pratiques. Afin de pouvoir comparer les différentes approches et justifier du choix de la pratique la mieux adaptée au traitement des données, il faut rattacher les différentes pratiques à une référence théorique commune, c'est-à-dire à un modèle commun. Pour paraphraser Daniel Chessel (1992), « *le modèle gère les pratiques qui gèrent les données. Les données interrogent les pratiques qui interrogent le modèle. Cela engendre de multiples débats parallèles qui se croisent de manière erratique* ».

Jusqu'à présent, les biométriciens n'ont fait qu'interroger les pratiques multiéchelles à partir des données. Par exemple, Hill (1973) compare l'analyse spectrale, l'analyse de variance hiérarchique et l'analyse qu'il propose (Two Term Local Quadrat Variance) par le biais des données de Greig-Smith et Chadwick (1965) sur l'abondance de l'espèce *Acacia ehrenbergiana*. Ripley (1978), Leps (1990) puis Dale (1999) reproduisent la même expérience avec des données simulées et des données de terrain. Chaque fois les conclusions sont intéressantes mais limitées par la structure des données utilisées. La seule tentative un peu sérieuse d'unification de certaines pratiques sur la base d'un modèle statistique commun est proposée par Ver-Hoef et al. (1993). Les auteurs rapportent l'ensemble des pratiques au formalisme des variables régionalisées. Ils évaluent chaque approche en terme d'estimation des propriétés d'un processus défini comme combinaison linéaire d'effets fixes et aléatoires. Le formalisme utilisé est propre à la géostatistique et reste peu propice à l'intégration des méthodes d'analyse des structures en analyse des données.

C'est pourquoi on a cherché à intégrer l'ensemble des approches dans un cadre théorique cohérent en utilisant le formalisme matriciel propre à l'analyse des données. On montre en particulier que la plupart des analyses de la structure d'une variable à différentes échelles k se résument à une famille de matrices symétriques réelles $(\mathbf{A}_k)_{1 \leq k \leq K}$, où K représente le nombre d'échelles considérées (Figure 2.6).

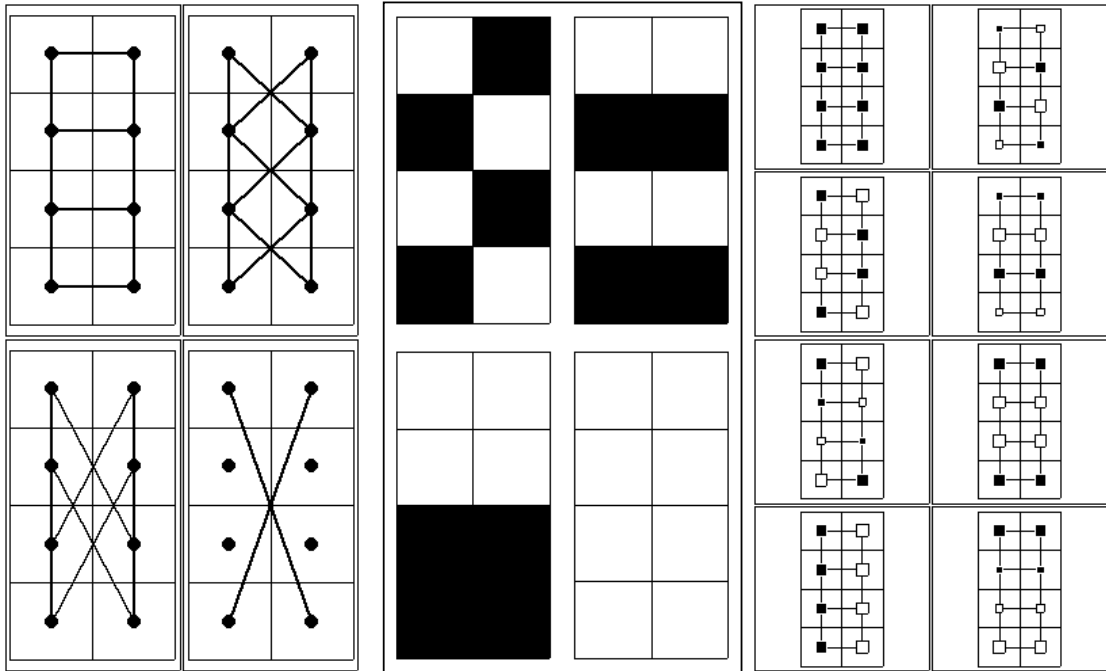


Figure 2.6 : Introduction de la notion d'échelle par le biais de la puissance d'un graphe (**à gauche**, $K = 4$) ((Smouse & Peakall, 1999)), par le biais de partitions emboîtées (**au centre**, $3 = 4$) ((Noy-Meir & Anderson, 1971)) et par le biais des vecteurs propres d'un opérateur de voisinage (**à droite**, $K = 8$) ((Méot et al., 1993)).

L'intensité de la structure d'une variable \mathbf{x} à une échelle k donnée est alors définie par la forme quadratique associée $\mathbf{x}'\mathbf{A}_k\mathbf{x}$. Cette remarque fondamentale est déjà dans l'article de Hill (Hill, 1973). Il souligne que si la forme bilinéaire associée à la matrice \mathbf{A}_k est positive et si $\text{sum}(\mathbf{A}_k) = 0$, la normalisation :

$$\frac{\mathbf{x}'\mathbf{A}_k\mathbf{x}}{\text{Trace}(\mathbf{A}_k)}$$

s'impose car

$$\text{Trace}\left(\frac{\mathbf{A}_k}{\text{Trace}(\mathbf{A}_k)}\right) = 1$$

donne une statistique de la variance.

En généralisant ce point de vue à l'ensemble des méthodes d'analyse des structures, on montre alors, que les principales méthodes se répartissent en deux grandes familles : les familles de formes bilinéaires positives parmi lesquelles on retrouve les familles de projecteurs, et les familles de formes bilinéaires non positives.

4. FAMILLES DE K FORMES BILINÉAIRES SYMÉTRIQUES

4.1. Définitions

Les matrices symétriques réelles à n lignes et n colonnes définissent les formes bilinéaires symétriques sur \mathbb{R}^n (Harville, 1997) :

$$f_A(\mathbf{x}, \mathbf{y}) = \mathbf{x}^t \mathbf{A} \mathbf{y} = \mathbf{y}^t \mathbf{A} \mathbf{x}$$

et les formes quadratiques :

$$q_A(\mathbf{x}) = \mathbf{x}^t \mathbf{A} \mathbf{x}$$

Par ailleurs, un couple arbitraire d'une matrice symétrique \mathbf{A} et d'une pondération \mathbf{D} définit une (ou plusieurs) base \mathbf{D} -orthonormale de vecteurs \mathbf{A} -orthogonaux en colonnes dans une matrice \mathbf{B} qui vérifie :

$$\begin{cases} \mathbf{B}^t \mathbf{D} \mathbf{B} = \mathbf{I}_n \\ \mathbf{B}^t \mathbf{A} \mathbf{B} = \text{diag}(\lambda_1, \dots, \lambda_n) \end{cases}$$

Le graphe des valeurs propres et l'ensemble des vecteurs propres de \mathbf{A} caractérisent la valeur de la forme et explicite sa fonction. Pour les avoir il suffit de diagonaliser

$$\mathbf{D}^{-\frac{1}{2}} \mathbf{A} \mathbf{D}^{-\frac{1}{2}} = \mathbf{V} \mathbf{\Lambda} \mathbf{V}^t \Rightarrow \mathbf{B} = \mathbf{D}^{-\frac{1}{2}} \mathbf{V}.$$

Une forme est dite positive si et seulement si

$$\forall \mathbf{x} \in \mathbb{R}^n \quad \mathbf{x}^t \mathbf{A} \mathbf{x} \geq 0.$$

Ceci n'est vraie que si et seulement si toutes ses valeurs propres sont positives ou nulles.

Le rang de \mathbf{A} est r et l'on peut éliminer les valeurs propres nulles et utiliser la famille des r vecteurs orthogonaux pour caractériser la forme bilinéaire par rapport à la pondération \mathbf{D} :

$$\begin{cases} \mathbf{B}_r^t \mathbf{D} \mathbf{B}_r = \mathbf{I}_r \\ \mathbf{B}_r^t \mathbf{A} \mathbf{B}_r = \text{diag}(\lambda_1, \dots, \lambda_r) \end{cases}$$

La matrice \mathbf{A} s'écrit alors sous la forme :

$$\mathbf{A} = \mathbf{D}\mathbf{B}\mathbf{A}\mathbf{B}\mathbf{D} \Rightarrow \mathbf{A} = \mathbf{D}\mathbf{B}_r\mathbf{A}_r\mathbf{B}'_r\mathbf{D}$$

On ajoute également à cette décomposition canonique la notion de **D**-centrage de la forme bilinéaire. En effet, en statistique, ce qui se passe sur une variable constante est généralement sans intérêt d'un point de vue typologique (elle a une variance nulle, une autocovariance nulle, une variance locale nulle, une corrélation nulle avec n'importe quoi, elle n'a pas de composante spatiale, ...). On ne fera donc un calcul de forme bilinéaire que sur des variables **D**-centrées. On ne considérera par la suite que la pondération uniforme $\mathbf{D} = \text{diag}(1/n, \dots, 1/n)$. D'autres pondérations pourraient également être utilisées. Par exemple, le graphe de voisinage donne la matrice **M** et la pondération de voisinage qui lui est associée. De même, dans une AFC sous contrainte spatiale on peut vouloir conserver la pondération issue de l'AFC tout en intégrant la matrice de la structure spatiale (Couteron & Ollier, sous presse).

4.2. La classe d'objets 'kfbs'

On va définir différentes familles de formes bilinéaires symétriques. On a besoin d'une structure pour manipuler ces objets. On définit alors une nouvelle classe d'objets dans l'environnement du logiciel R que l'on appelle 'kfbs' pour K Formes Bilinéaires Symétriques (Annexe 2.5). On introduit également un ensemble de fonctions qui vont permettre leur manipulation. Les K matrices symétriques $(\mathbf{A}_k)_{1 \leq k \leq K}$ associées aux formes bilinéaires sont rangées en K colonnes dans une matrice de stockage de dimension $n(n+1)/2 \times K$ qui contient les K demi matrices inférieures. Les attributs d'un objet de la classe 'kfbs' sont :

- 'dim' pour la dimension de la matrice de stockage $n(n+1)/2 \times K$
- 'npoints' pour le nombre de points n
- 'nforms' pour le nombre de formes bilinéaires K
- 'scalprod' pour la nature des formes bilinéaires (positives ou non)
- 'trace1' pour la trace des matrices \mathbf{A}_k
- 'trace2' pour la trace des matrices \mathbf{A}_k^2
- 'sum' pour la somme des matrices \mathbf{A}_k
- 'rang' pour le rang des matrices \mathbf{A}_k
- 'norm' pour la norme spectrale des matrices \mathbf{A}_k

- ‘labels’ pour le nom des matrices A_k
- ‘call’ rappelle la ligne de commande qui a permis la création de l’objet
- ‘class’ pour la classe de l’objet

4.3. Formes de Geary/Lebart : le variogramme

4.3.1. Définition et propriétés

La forme de Geary/Lebart est définie à partir d’un graphe de voisinage. Elle est introduite dans Lebart (1969), précisée dans Banet et Lebart (1984) et reprise dans Di-Bella et Jona-Lasinio (1996). La forme de Geary/Lebart s’étend à une famille de formes dès que l’on considère non plus un graphe de voisinage mais une famille de graphes de voisinage. Il existe plusieurs manières de générer une famille de graphes (objet de la classe ‘knb’, voir Annexe 2.8). On choisit, à titre d’illustration, d’introduire les échelles par le biais des puissances du graphe (Smouse & Peakall, 1999). Dans un graphe connexe, deux points sont reliés par un chemin. Parmi ces chemins il en existe un au moins de longueur minimum. Cette longueur définit la distance entre les deux points. Deux points sont alors voisins à la puissance k si la distance qui les sépare vaut exactement k (Figure 2.7).

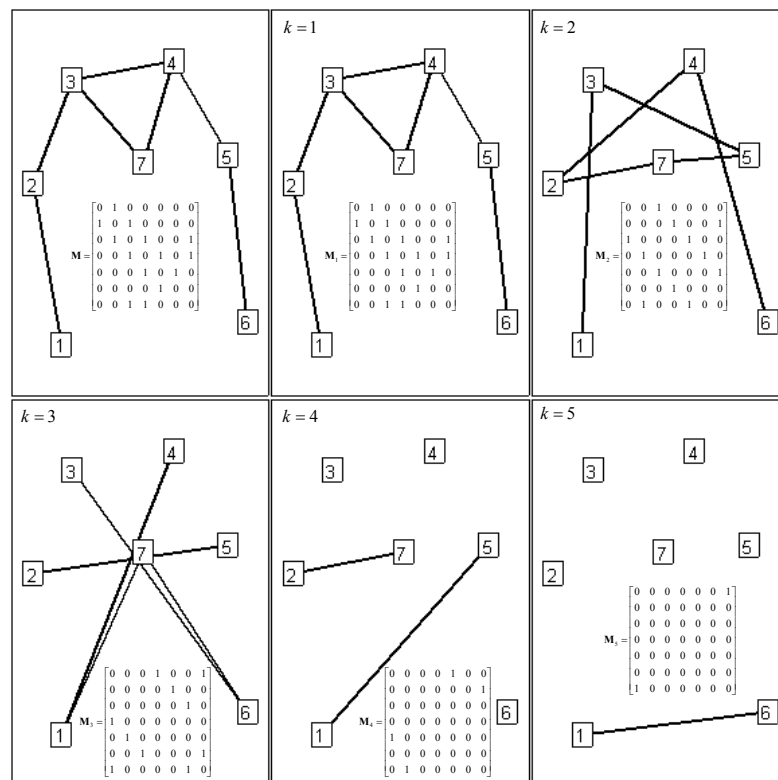


Figure 2.7 : graphe et puissances d’un graphe.

Les matrices $(\mathbf{M}_k)_{1 \leq k \leq K}$ définies par les puissances d'un graphe connexe vérifient $\sum_{k=1}^K \mathbf{M}_k = \mathbf{U} - \mathbf{Id}_n$ où $\mathbf{U} = \mathbf{1}_n \mathbf{1}_n^t$ et \mathbf{Id}_n est la matrice identité (Lebart, 1969). Les matrices $(\mathbf{M}_k)_{1 \leq k \leq K}$ sont les matrices de voisinage à la puissance k . Les matrices $\mathbf{N}_k = \text{Diag}(\mathbf{M}_k \mathbf{1}_n)$ sont les matrices diagonales des degrés du graphe à la puissance k . Les valeurs $\mathbf{1}_n^t \mathbf{N}_k \mathbf{1}_n = 2m_k$ correspondent aux nombres de couples de voisins (2 fois le nombre de paires). Il vient immédiatement que :

$$\sum_{k=1}^K \mathbf{N}_k = (n-1) \mathbf{Id}_n$$

car chaque point possède $n-1$ voisins, d'où :

$$\sum_{k=1}^K (\mathbf{N}_k - \mathbf{M}_k) = n \mathbf{Id}_n - \mathbf{U}.$$

On peut alors considérer deux familles d'opérateurs.

Les premiers sont les opérateurs de variance locale associés au calcul de l'indice de Geary et de la métrique de Lebart. Ils sont définis de $\mathbb{R}^n \times \mathbb{R}^n$ dans \mathbb{R} par :

$$(\mathbf{x}, \mathbf{y}) \mapsto \mathbf{x}^t (\mathbf{N}_k - \mathbf{M}_k) \mathbf{y} = \sum_{i \sim_k j} (x_i - x_j)(y_i - y_j)$$

Les matrices $(A_k)_{1 \leq k \leq K} = (\mathbf{N}_k - \mathbf{M}_k)_{1 \leq k \leq K}$ sont **symétriques** et **positives** et définissent les formes bilinéaires de Geary/Lebart. Elles sont très générales car elles sont définies pour toute famille de K graphes de voisinage. Leurs propriétés dépendent des propriétés des familles de graphes. Par exemple, lorsque l'on définit la famille de graphes à partir des puissances d'un graphe linéaire, on retrouve les métriques PQV (Paired Quadrat Variance) définies par Goodall (1974) qui sont des estimateurs du variogramme de Matheron ((Ver Hoef et al., 1993)) :

$$PQV_k(\mathbf{x}) = \frac{1}{n-k} \sum_{i=1}^{n-k} (x_i - x_{i+k})^2$$

Dans une certaine mesure, on peut dire que les familles de formes de Geary/Lebart généralisent la notion de variogramme à l'ensemble des familles de graphes.

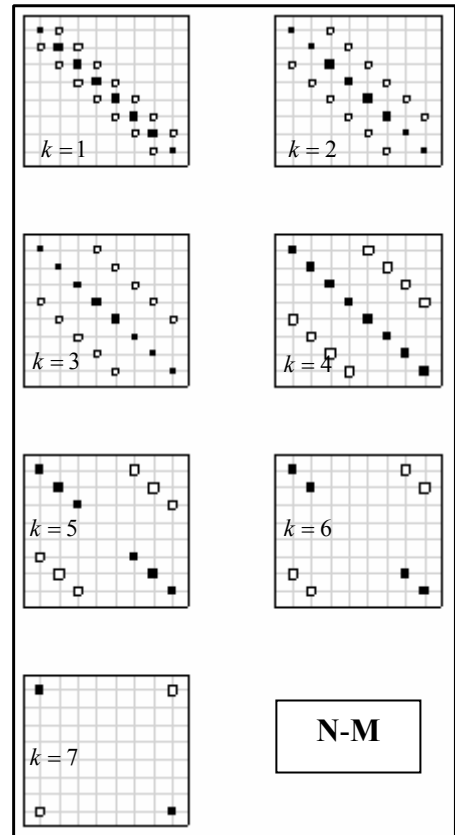
4.3.2. La fonction **knb2kfb**s (...)

La fonction `knb2kfbs(...)` (Annexe 2.5) définit des objets de la classe 'kfbs' à partir d'un objet de la classe 'knb' (Annexe 2.8) : elle calcule les formes de Geary/Lebart associées à une famille de graphes de voisinage.

Exemple :

```
ng <- neig(n.line = 8) # graphe linéaire
knb <- neig2knb(ng) # puissances du graphes
geary.kfbs <- knb2kfbs(knb, method = "Geary") # formes de Geary/Lebart
print(geary.kfbs) # matrice de stockage
```

```
      [,1] [,2] [,3] [,4] [,5] [,6] [,7]
[1,]  1    1    1    1    1    1    1
[2,] -1    0    0    0    0    0    0
[3,]  0   -1    0    0    0    0    0
[4,]  0    0   -1    0    0    0    0
[5,]  0    0    0   -1    0    0    0
[6,]  0    0    0    0   -1    0    0
[7,]  0    0    0    0    0   -1    0
[8,]  0    0    0    0    0    0   -1
[9,]  2    1    1    1    1    1    0
[10,] -1    0    0    0    0    0    0
[11,]  0   -1    0    0    0    0    0
[12,]  0    0   -1    0    0    0    0
[13,]  0    0    0   -1    0    0    0
[14,]  0    0    0    0   -1    0    0
[15,]  0    0    0    0    0   -1    0
[16,]  2    2    1    1    1    0    0
[17,] -1    0    0    0    0    0    0
[18,]  0   -1    0    0    0    0    0
[19,]  0    0   -1    0    0    0    0
[20,]  0    0    0   -1    0    0    0
[21,]  0    0    0    0   -1    0    0
[22,]  2    2    2    1    0    0    0
[23,] -1    0    0    0    0    0    0
[24,]  0   -1    0    0    0    0    0
[25,]  0    0   -1    0    0    0    0
[26,]  0    0    0   -1    0    0    0
[27,]  2    2    2    1    0    0    0
[28,] -1    0    0    0    0    0    0
[29,]  0   -1    0    0    0    0    0
[30,]  0    0   -1    0    0    0    0
[31,]  2    2    1    1    1    0    0
[32,] -1    0    0    0    0    0    0
[33,]  0   -1    0    0    0    0    0
[34,]  2    1    1    1    1    1    0
[35,] -1    0    0    0    0    0    0
[36,]  1    1    1    1    1    1    1
attr(,"npoints") # attributs
[1] 8
attr(,"nforms")
[1] 7
attr(,"scalprod") # les formes sont positives
[1] TRUE
attr(,"tracel") # leurs traces sont non nulles
[1] 14 12 10 8 6 4 2
attr(,"trace2")
[1] 40 32 24 16 12 8 4
attr(,"sum") # leurs sommes sont nulles
[1] 0 0 0 0 0 0 0
attr(,"rank")
[1] 7 6 5 4 3 2 1
attr(,"norm")
[1] 3.848 3.414 3.000 2.000 2.000 2.000 2.000
attr(,"labels")
[1] "G1" "G2" "G3" "G4" "G5" "G6" "G7"
attr(,"call")
```



```

knb2kfbs(knb = knb, method = "Geary")
attr(,"class")
[1] "kfbs"

summary(geary.kfbs)
K bilinear symmetric forms
Points : 8      Forms : 7
All positive forms : TRUE
Call : knb2kfbs(knb = knb, method = "Geary")
  tr(f) tr(f^2) sum rank norm
G1   14    40    0     7 3.848
G2   12    32    0     6 3.414
G3   10    24    0     5 3.000
G4    8    16    0     4 2.000
G5    6    12    0     3 2.000
G6    4     8    0     2 2.000
G7    2     4    0     1 2.000
    
```

4.4. Formes de Moran/Smouse : le corrélogramme

4.4.1. Définition et propriétés

Les seconds opérateurs sont les opérateurs d'autocovariance locale associés au calcul de l'indice de Moran et de la forme bilinéaire introduite par Smouse et Peakall (1999). Ils sont définis de $\mathbb{R}^n \times \mathbb{R}^n$ dans \mathbb{R} par :

$$(\mathbf{x}, \mathbf{y}) \mapsto \mathbf{x}' \mathbf{M}_k \mathbf{y} = \sum_{i \mathbf{M}_k \text{-voisin } j}^n x_i y_j$$

Les matrices $(A_k)_{1 \leq k \leq K} = (\mathbf{M}_k)_{1 \leq k \leq K}$ sont les matrices **symétriques** mais **non positives** associées aux formes bilinéaires de Moran/Smouse. De plus, comme un couple de points donné ne peut être constitué de voisins qu'à un seul niveau, on a :

$$j \neq k \Rightarrow \text{Trace}(\mathbf{M}_j \mathbf{M}_k) = 0$$

et comme la matrice est symétrique et ne contient que des 0 et des 1, on a :

$$\mathbf{N}_k = \text{Diag}(\mathbf{M}_k^2) = \mathbf{N}_k \Rightarrow \text{Trace}(\mathbf{M}_k^2) = 2m_k$$

Les formes de Moran/Smouse constituent donc une **famille orthogonale** pour le produit scalaire euclidien car :

$$\text{Trace}(\mathbf{M}_j \mathbf{M}_k) = 2m_k \delta_{jk} \text{ où } \delta_{jk} \text{ représente le symbole de Kronecker}$$

Enfin, par analogie avec les formes de Geary/Lebart, on peut dire que les formes de Moran/Smouse généralisent la notion de correlogramme à l'ensemble des familles de graphes.

4.4.2. La fonction `knb2kfbs (...)`

La fonction `knb2kfbs(...)` (Annexe 2.5) définit des objets de la classe 'kfbs' à partir d'un objet de la classe 'knb' (Annexe 2.8) : elle calcule les formes de Moran/Smouse associées à une famille de graphes de voisinage.

Exemple :

```

ng <- neig(n.line = 8)      # graphe linéaire
knb <- neig2knb(ng)        # puissances du graphes
moran.kfbs <- knb2kfbs(knb, method = "Moran") # formes de Moran/Smouse
print(moran.kfbs)         # matrice de stockage

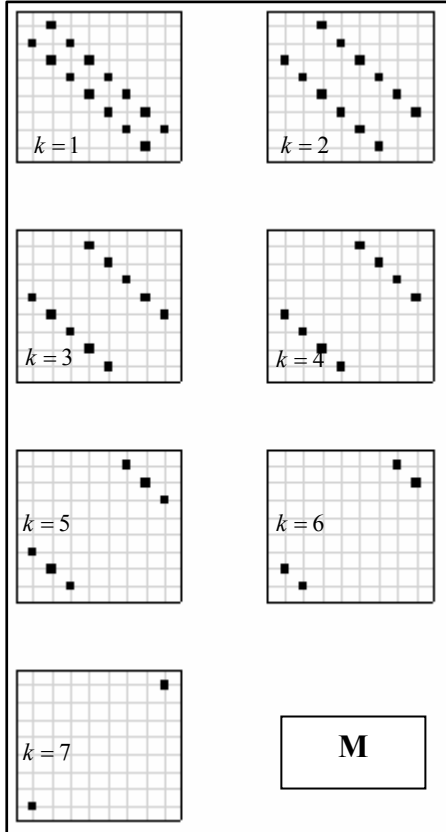
```

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]
[1,]	0	0	0	0	0	0	0
[2,]	1	0	0	0	0	0	0
[3,]	0	1	0	0	0	0	0
[4,]	0	0	1	0	0	0	0
[5,]	0	0	0	1	0	0	0
[6,]	0	0	0	0	1	0	0
[7,]	0	0	0	0	0	1	0
[8,]	0	0	0	0	0	0	1
[9,]	0	0	0	0	0	0	0
[10,]	1	0	0	0	0	0	0
[11,]	0	1	0	0	0	0	0
[12,]	0	0	1	0	0	0	0
[13,]	0	0	0	1	0	0	0
[14,]	0	0	0	0	1	0	0
[15,]	0	0	0	0	0	1	0
[16,]	0	0	0	0	0	0	0
[17,]	1	0	0	0	0	0	0
[18,]	0	1	0	0	0	0	0
[19,]	0	0	1	0	0	0	0
[20,]	0	0	0	1	0	0	0
[21,]	0	0	0	0	1	0	0
[22,]	0	0	0	0	0	0	0
[23,]	1	0	0	0	0	0	0
[24,]	0	1	0	0	0	0	0
[25,]	0	0	1	0	0	0	0
[26,]	0	0	0	1	0	0	0
[27,]	0	0	0	0	0	0	0
[28,]	1	0	0	0	0	0	0
[29,]	0	1	0	0	0	0	0
[30,]	0	0	1	0	0	0	0
[31,]	0	0	0	0	0	0	0
[32,]	1	0	0	0	0	0	0
[33,]	0	1	0	0	0	0	0
[34,]	0	0	0	0	0	0	0
[35,]	1	0	0	0	0	0	0
[36,]	0	0	0	0	0	0	0

```

attr(,"npoints")
[1] 8
attr(,"nforms")
[1] 7
attr(,"scalprod")
[1] FALSE
attr(,"trace1")
[1] 0 0 0 0 0 0 0
attr(,"trace2")
[1] 14 12 10 8 6 4 2
attr(,"sum")
[1] 14 12 10 8 6 4 2
attr(,"rank")
[1] 8 8 6 8 6 4 2
attr(,"norm")
[1] 1.879 1.618 1.414 1.000 1.000 1.000 1.000
attr(,"labels")
[1] "M1" "M2" "M3" "M4" "M5" "M6" "M7"
attr(,"call")

```



```

# attributs

# les formes sont non positives

# leurs traces sont nulles

# la trace de leurs carrés = sommes

```

```

knb2kfbs(knb = knb, method = "Moran")
attr(,"class")
[1] "kfbs"

summary(moran.kfbs)
K bilinear symetric forms
Points : 8      Forms : 7
All positive forms : FALSE
Call : knb2kfbs(knb = knb, method = "Moran")
      tr(f) tr(f^2) sum rank norm
M1      0      14  14    8 1.879
M2      0      12  12    8 1.618
M3      0      10  10    6 1.414
M4      0       8   8    8 1.000
M5      0       6   6    6 1.000
M6      0       4   4    4 1.000
M7      0       2   2    2 1.000

```

4.5. Formes de Greig-Smith/Noy-Meir : les msbs

4.5.1. Définition et propriétés

Elles sont introduites par Noy-Meir et Anderson (1971). Explicitement les auteurs utilisent p espèces et n sites et x_{ij} est l'abondance de l'espèce j dans le site i . Le tableau de départ est \mathbf{X} (n lignes-relevés et p colonnes-espèces). Par la suite, on ne considérera qu'une seule espèce ($p = 1$) dont l'abondance est représentée par le vecteur colonne \mathbf{x} . Pour les tailles de blocs $(b_k)_{1 \leq k \leq K+1}$, les sites sont regroupés en $n_k = n/b_k$ blocs de b_k sites élémentaires. L'objectif de ce paragraphe est de traduire l'écriture des auteurs en termes matriciels (Figure 2.8).

- Noy-Meir et Anderson commencent par calculer, pour chaque taille de bloc b_k , les abondances sommées par bloc $(\mathbf{x}_k)_{1 \leq k \leq K+1}$ (étape 1, Figure 2.8). Soit la matrice \mathbf{H}_k de dimension $n \times n_k$ des indicatrices d'appartenance d'un site élémentaire aux différents blocs (le tableau disjonctif complet associé à la partition de l'ensemble des sites en blocs) :

$$\mathbf{x}_k = \mathbf{H}'_k \mathbf{x} .$$

- Ils calculent ensuite, pour chaque site, les différences $(\mathbf{y}_k)_{1 \leq k \leq K}$ entre l'abondance moyenne à la taille de blocs b_k et l'abondance moyenne à la taille de blocs b_{k+1} (étape 2, Figure 2.8):

$$\mathbf{y}_k = \frac{1}{b_k} \mathbf{H}_k \mathbf{H}'_k \mathbf{x} - \frac{1}{b_{k+1}} \mathbf{H}_{k+1} \mathbf{H}'_{k+1} \mathbf{x}$$

Par convention, $k=1$ correspond à la plus petite taille de blocs $b_1=1$ d'où $\mathbf{x}_1 = \mathbf{x}$ et $\mathbf{H}_1 = \mathbf{I}_n$; $k=K$ correspond à la plus grande taille de blocs utile b_K ; et $k=K+1$ correspond à la plus grande taille de blocs possible $b_{K+1} = n$ soit $\mathbf{x}_{K+1} = \mathbf{1}'_n \mathbf{x}$ et

$$\frac{1}{b_{h+1}} \mathbf{H}_{h+1} \mathbf{H}'_{h+1} \mathbf{x} = \frac{1}{n} \mathbf{1}_n \mathbf{1}'_n \mathbf{x} = \bar{\mathbf{x}}.$$

D'où :

$$\mathbf{y}_1 + \mathbf{y}_2 + \dots + \mathbf{y}_k = \mathbf{x} - \frac{1}{n} \mathbf{1}_n \mathbf{1}'_n \mathbf{x} = \mathbf{x} - \bar{\mathbf{x}} = \mathbf{y}$$

$\frac{1}{b_k} \mathbf{H}_k \mathbf{H}'_k = \mathbf{P}_k$ est le projecteur orthogonal sur le sous-espace des indicatrices des blocs de taille b_k , sous-espace de dimension $n_k = n/b_k$ donc

$$\mathbf{y}_k = \frac{1}{b_k} \mathbf{H}_k \mathbf{H}'_k \mathbf{x} - \frac{1}{b_{k+1}} \mathbf{H}_{k+1} \mathbf{H}'_{k+1} \mathbf{x} = (\mathbf{P}_k - \mathbf{P}_{k+1}) \mathbf{x}$$

- Ils calculent, pour finir, les indices de dispersion $(g_k)_{1 \leq k \leq K}$ (étape 3, Figure 2.8):

$$g_k = \mathbf{y}'_k \mathbf{y}_k = \mathbf{x}' (\mathbf{P}_k - \mathbf{P}_{k+1}) (\mathbf{P}_k - \mathbf{P}_{k+1}) \mathbf{x}$$

Or :

$$(\mathbf{P}_k - \mathbf{P}_{k+1})(\mathbf{P}_k - \mathbf{P}_{k+1}) = \mathbf{P}_k^2 + \mathbf{P}_{k+1}^2 - \mathbf{P}_{k+1} \mathbf{P}_k - \mathbf{P}_k \mathbf{P}_{k+1} = \mathbf{P}_k + \mathbf{P}_{k+1} - \mathbf{P}_{k+1} \mathbf{P}_k - \mathbf{P}_k \mathbf{P}_{k+1}$$

Si les blocs sont emboîtés, et seulement dans ce cas, le sous-espace des indicatrices des blocs de taille b_k contient le sous-espace des indicatrices des blocs de taille b_{k+1} et $\mathbf{P}_{k+1} \mathbf{P}_k = \mathbf{P}_{k+1}$ (théorème des trois perpendiculaires), $\mathbf{P}_k \mathbf{P}_{k+1} = \mathbf{P}_{k+1}$ (à cause de l'inclusion) d'où :

$$g_k = \mathbf{y}'_k \mathbf{y}_k = \mathbf{x}' (\mathbf{P}_k - \mathbf{P}_{k+1}) \mathbf{x} = \mathbf{x}' \mathbf{A}_k \mathbf{x}$$

$\mathbf{A}_k = \mathbf{P}_k - \mathbf{P}_{k+1}$ est alors le projecteur sur le sous espace complémentaire orthogonal du sous-espace des indicatrices des blocs de taille b_{k+1} dans le sous-espace des indicatrices des blocs de taille b_k . La famille des projecteurs $(\mathbf{A}_k)_{1 \leq k \leq K}$ définit la famille des formes bilinéaires de Noy-Meir et Anderson. Les indices de dispersion $(g_k)_{1 \leq k \leq K}$ sont les formes quadratiques associées à ces matrices.

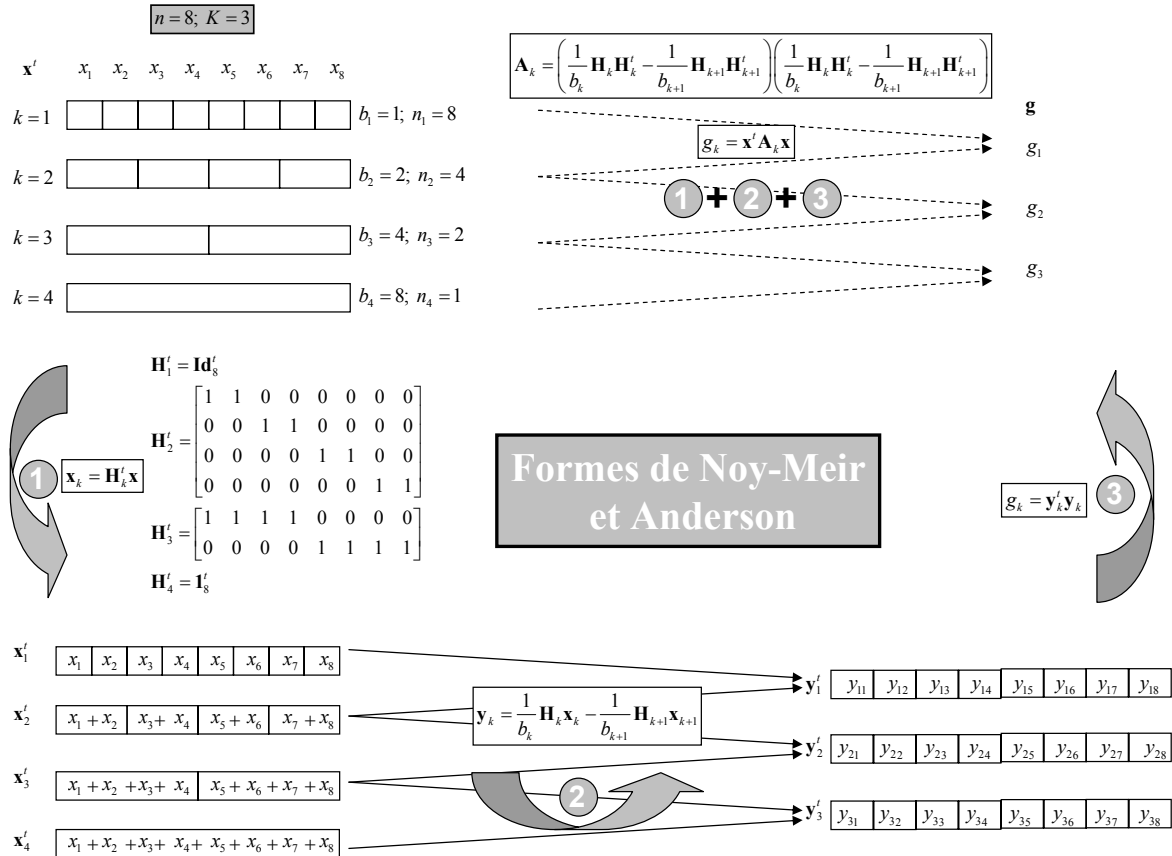


Figure 2.8 : Formes de Noy-Meir et Anderson pour un transect constitués par 8 cadrats

Les métriques de Noy-Meir et Anderson sont donc un cas particulier de la situation générale où $\mathbb{R}^n = V_1 \overset{\perp}{\otimes} V_2 \overset{\perp}{\otimes} \dots \overset{\perp}{\otimes} V_k \overset{\perp}{\otimes} [1_n]$ se décompose en sous-espaces orthogonaux pour la métrique canonique. L'identité se décompose en somme de projecteurs orthogonaux,

$$\mathbf{Id}_n = \mathbf{A}_1 + \mathbf{A}_2 + \dots + \mathbf{A}_k + \mathbf{A}_{1_n},$$

qui définissent une famille de k formes bilinéaires symétriques positives. De plus, les formes bilinéaires étant des projecteurs orthogonaux, elles vérifient

$$i \neq k \Rightarrow \mathbf{A}_j \mathbf{A}_k = 0 \Rightarrow \text{Trace}(\mathbf{A}_j \mathbf{A}_k) = 0$$

$$\mathbf{A}_j \mathbf{A}_j = \mathbf{A}_j \Rightarrow \text{Trace}(\mathbf{A}_j^2) = \text{Trace}(\mathbf{A}_j) = \dim(V_j)$$

On peut généraliser l'approche de Noy-Meir et Anderson. En particulier, l'emboîtement nécessaire des blocs de taille successive qui veut qu'une partition à un niveau regroupe un nombre égal de blocs au niveau précédent pour assurer que les plans soient toujours orthogonaux n'est nécessaire que pour avoir des formules explicites. Au départ, dans l'analyse de variance hiérarchique introduite par Greig-Smith (1952), il utilise en effet des lignes de placettes découpées en blocs de taille 1, 2, 4, 8, ..., 128, 512, ... voire 2048 placettes et le plan d'échantillonnage a été conçu pour cela. C'est évidemment extrêmement contraignant. Avec une ligne de 127 placettes, on ne peut strictement rien faire.

On peut s'affranchir de la contrainte des blocs emboîtés en acceptant des calculs impossibles sous forme explicite. En effet la suite des sous espaces $V_1, V_2, \dots, V_K, [\mathbf{1}_n]$ de la décomposition est engendrée simplement. On considère $E_1, E_2, \dots, E_K, [\mathbf{1}_n]$ la suite de sous-espaces de \mathbb{R}^n engendrés par les indicatrices des blocs de taille donnée, par exemple 1, 2, ..., 2^K , avec $n = 2^{K+1}$. L'indicatrice unique du seul bloc de taille n est évidemment $\mathbf{1}_n$. Le premier est de dimension n , le second de dimension $n/2$, ... Les sous espaces de projections sont alors le complémentaire de chacun d'entre eux dans le précédent, ce qu'on appelle en analyse de variance l'espace E_k / E_{k+1} partie orthogonale de E_{k+1} dans $E_k + E_{k+1} = E_k$ à cause de l'inclusion. Pour une suite de partitions non emboîtées et/ou non régulières (par exemple quand le dernier bloc est incomplet) il suffira de prendre les sous-espaces :

$$\begin{aligned} V_1 &= E_1 \cap E_2^\perp \\ V_2 &= E_2 \cap E_3^\perp \\ &\vdots \\ E_K &= E_K \cap [\mathbf{1}_n]^\perp \end{aligned}$$

Chacun d'entre eux est orthogonal au précédent par construction. L'existence de ces sous-espaces a été introduite par Afriat (1957). Pour les calculer, en partant de générateurs quelconques, on passe par les bases orthonormées du premier et de l'orthogonal du second par décomposition QR complète puis l'intersection des deux par l'analyse canonique en conservant les valeurs propres égales à 1. On obtient directement des bases orthonormées et les formes bilinéaires symétriques et positives associées par $\mathbf{A}_k = \mathbf{B}_k \mathbf{B}_k'$.

En suivant la notation de la remarquable synthèse de Ver-Hoef et al. (1993), nous appellerons cette famille de formes bilinéaires les msbs pour ‘Mean Square Block Size’. Pour le moment, elles ne s'appliquent qu'aux structures de données alignées. Ce sont des cas particuliers de matrices \mathbf{A}_k définies comme des projecteurs orthogonaux (dans \mathbb{R}^n pour la métrique canonique). On a dans ce cas :

$$\mathbf{A}_k^2 = \mathbf{A}_k \text{ donc } Trace(\mathbf{A}_k^2) = Trace(\mathbf{A}_k)$$

Si $\mathbf{1}_n$ est dans l'orthogonal de l'image du projecteur, leur somme est nulle. Elles portent un nom du type k_k+1 qui signifie que les n points ont été répartis en blocs de taille k et en blocs de taille $k+1$. Les indicatrices des blocs définissent les sous-espaces vectoriels de \mathbb{R}^n E_k et E_{k+1} . Le projecteur est le projecteur orthogonal sur $E_k \cap E_{k+1}^\perp$.

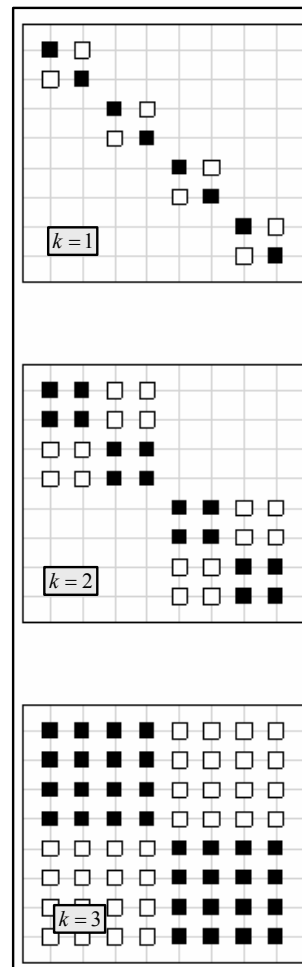
4.5.2. La fonction `msbs.kfbs(...)`

La fonction `msbs.kfbs(...)` (Annexe 2.5) définit des objets de la classe ‘kfbs’ pour un transect et plusieurs tailles de bloc : elle calcule les formes de Noy-Meir/Greig-Smith.

Exemple :

```
noy.kfbs <- msbs.kfbs(n = 8, tbloc = c(1,2,4,8)) # formes de Noy-Meir/Anderson
print(noy.kfbs) # matrice de stockage
```

```
      [,1]      [,2]      [,3]
[1,]  5.000e-01  2.500e-01  0.125
[2,] -5.000e-01  2.500e-01  0.125
[3,]  1.247e-16 -2.500e-01  0.125
[4,]  2.553e-17 -2.500e-01  0.125
[5,] -3.925e-17 -2.082e-17 -0.125
[6,] -1.225e-16 -2.776e-17 -0.125
[7,] -8.233e-17  2.776e-17 -0.125
[8,] -5.758e-17  2.082e-17 -0.125
[9,]  5.000e-01  2.500e-01  0.125
[10,] -1.224e-16 -2.500e-01  0.125
[11,] -2.791e-17 -2.500e-01  0.125
[12,]  3.925e-17  3.469e-17 -0.125
[13,]  1.503e-16  2.776e-17 -0.125
[14,]  6.608e-17 -2.776e-17 -0.125
[15,]  7.383e-17 -3.469e-17 -0.125
[16,]  5.000e-01  2.500e-01  0.125
[17,] -5.000e-01  2.500e-01  0.125
[18,] -1.225e-16 -1.388e-17 -0.125
[19,] -6.700e-17 -6.939e-18 -0.125
[20,]  7.641e-17  6.939e-18 -0.125
[21,]  6.482e-17  1.388e-17 -0.125
[22,]  5.000e-01  2.500e-01  0.125
[23,]  9.476e-17 -3.081e-33 -0.125
[24,]  6.700e-17  6.939e-18 -0.125
[25,] -7.977e-17 -6.939e-18 -0.125
[26,] -6.145e-17  1.541e-33 -0.125
[27,]  5.000e-01  2.500e-01  0.125
[28,] -5.000e-01  2.500e-01  0.125
[29,]  4.623e-19 -2.500e-01  0.125
[30,]  4.689e-19 -2.500e-01  0.125
[31,]  5.000e-01  2.500e-01  0.125
```



```

[32,] -4.623e-19 -2.500e-01  0.125
[33,] -4.689e-19 -2.500e-01  0.125
[34,]  5.000e-01  2.500e-01  0.125
[35,] -5.000e-01  2.500e-01  0.125
[36,]  5.000e-01  2.500e-01  0.125
attr(,"npoints")
[1] 8
attr(,"nforms")
[1] 3
attr(,"scalprod") # leurs formes sont positives
[1] TRUE
attr(,"tracel") # leurs traces = traces de leurs carrés
[1] 4 2 1
attr(,"trace2")
[1] 4 2 1
attr(,"sum") # leurs sommes sont nulles
[1] -5.421e-20  0.000e+00 -5.551e-17
attr(,"rank")
[1] 4 2 1
attr(,"norm")
[1] 1 1 1
attr(,"labels")
[1] "1_2" "2_4" "4_8"
attr(,"class")
[1] "kfbs"
attr(,"call")
msbs.kfbs(n = 8, tbloc = c(1, 2, 4, 8))

```

```

summary(noy.kfbs)
K bilinear symetric forms
Points : 8      Forms : 3
All positive forms : TRUE
Call : msbs.kfbs(n = 8, tbloc = c(1, 2, 4, 8))
      tr(f) tr(f^2)      sum rank norm
1_2      4      4 -5.421e-20    4    1
2_4      2      2  0.000e+00    2    1
4_8      1      1 -5.551e-17    1    1

```

4.6. Formes de Hill : les ttlv

4.6.1. Définition et propriétés

Elles ont été introduites par Hill (1973) sur la base d'un défaut assez sévère des précédentes qui tient au point de départ. Si le nombre de placettes n'est pas divisible par le nombre de blocs, le dernier bloc est incomplet. On pourrait évidemment commencer par le bloc incomplet ou mettre ce bloc n'importe où. En ce sens le calcul dépend de la position du premier point du premier bloc. L'auteur propose alors pour une variable $\mathbf{x}' = (x_1, x_2, \dots, x_n)$, de prendre comme métrique pour la taille de bloc $b = 1$ la quantité

$$\text{average of } \left(\frac{1}{2}(x_1 - x_2)^2, \frac{1}{2}(x_2 - x_3)^2, \text{etc} \right),$$

pour la taille de bloc $b = 3$ la quantité

$$\text{average of } \left(\frac{1}{6}(x_1 + x_2 + x_3 - x_4 - x_5 - x_6)^2, \frac{1}{6}(x_2 + x_3 + x_4 - x_5 - x_6 - x_7)^2, \dots \right),$$

et ainsi de suite pour l'ensemble des tailles de bloc $(b_k)_{1 \leq k \leq K}$.

On propose de réécrire les métriques de Hill sous forme matricielle. Les matrices associées aux métriques de Hill sont facilement calculables car directement reliées aux matrices associées aux métriques de Geary/Lebart. Il suffit d'observer l'exemple ci-dessous avec $n = 7$ et $b_k = 3$:

$$\mathbf{A}_k = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & -1 & -1 & -1 & 0 \\ 1 & 2 & 2 & 0 & -2 & -2 & -1 \\ 1 & 2 & 2 & 0 & -2 & -2 & -1 \\ -1 & 0 & 0 & 2 & 0 & 0 & -1 \\ -1 & -2 & -2 & 0 & 2 & 2 & 1 \\ -1 & -2 & -2 & 0 & 2 & 2 & 1 \\ 0 & -1 & -1 & -1 & 1 & 1 & 1 \end{bmatrix}$$

Au centre du calcul de \mathbf{A}_k on reconnaît le graphe de voisinage des blocs sous sa forme $\mathbf{N} - \mathbf{M}$. On peut y substituer \mathbf{M} pour obtenir des corrélogrammes par blocs. Les matrices \mathbf{A}_k ont une somme d'éléments nulle et les coefficients des termes $\frac{1}{2}, \frac{1}{6}, \dots, \frac{1}{2m}$ ainsi que ceux associés à la moyenne des termes seront directement intégrés dans $Trace(\mathbf{A}_k)$ pour obtenir les estimateurs de la variance définis par Hill (1973).

On obtient des formes bilinéaires **symétriques** et **positives** $(\mathbf{A}_k)_{1 \leq k \leq K}$ que l'on nomme **t1lv** pour 'Two Terms Local Variances'. Pour le moment, elles ne s'appliquent qu'aux structures de données alignées. Elles sont une combinaison des précédentes (on somme les données par blocs de taille données k (qui donne son nom à la forme) et on utilise la relation de voisinage entre blocs contigus). Tous les couples de blocs voisins sont utilisés. Leurs matrices sont du type :

$$\mathbf{A}_k = \mathbf{H}_k (\mathbf{N}_{\mathbf{H}_k} - \mathbf{M}_{\mathbf{H}_k}) \mathbf{H}_k^t$$

4.6.2. La fonction `t1lv.kfbs(...)`

La fonction `t1lv.kfbs(...)` (Annexe 2.5) définit des objets de la classe 'kfbs' pour un transect et plusieurs tailles de bloc : elle calcule les formes de Hill.

Exemple :

```
hill.geary.kfbs <- t1lv.kfbs(n = 8,
tbloc = c(1,2,4), method = "Geary") # variogramme
print(hill.geary.kfbs)                # matrice de stockage
```



```

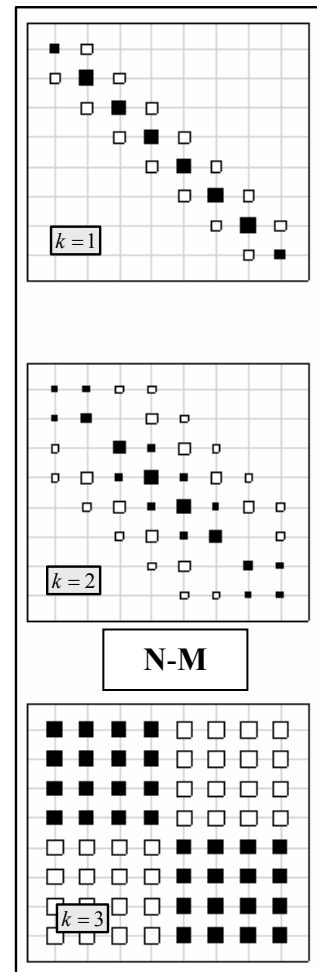
      [,1] [,2] [,3]
 [1,]    1    1    1
 [2,]   -1    1    1
 [3,]    0   -1    1
 [4,]    0   -1    1
 [5,]    0    0   -1
 [6,]    0    0   -1
 [7,]    0    0   -1
 [8,]    0    0   -1
 [9,]    2    2    1
[10,]   -1    0    1
[11,]    0   -2    1
[12,]    0   -1   -1
[13,]    0    0   -1
[14,]    0    0   -1
[15,]    0    0   -1
[16,]    2    3    1
[17,]   -1    1    1
[18,]    0   -2   -1
[19,]    0   -1   -1
[20,]    0    0   -1
[21,]    0    0   -1
[22,]    2    4    1
[23,]   -1    1   -1
[24,]    0   -2   -1
[25,]    0   -1   -1
[26,]    0    0   -1
[27,]    2    4    1
[28,]   -1    1    1
[29,]    0   -2    1
[30,]    0   -1    1
[31,]    2    3    1
[32,]   -1    0    1
[33,]    0   -1    1
[34,]    2    2    1
[35,]   -1    1    1
[36,]    1    1    1
attr(,"npoints")
 [1] 8
attr(,"nforms")
 [1] 3
attr(,"scalprod")
 [1] TRUE
attr(,"tracel")
 [1] 14 20 8
attr(,"trace2")
 [1] 40 116 64
attr(,"sum")
 [1] 0 0 0
attr(,"rank")
 [1] 7 5 1
attr(,"norm")
 [1] 3.848 7.236 8.000
attr(,"call")
ttl.v.kfbs(n = 8, tbloc = c(1, 2, 4), method = "Geary")
attr(,"labels")
 [1] "ttg_1" "ttg_2" "ttg_4"
attr(,"class")
 [1] "kfbs"

```

```

summary(hill.geary.kfbs)
K bilinear symmetric forms
Points : 8    Forms : 3
All positive forms : TRUE
Call : ttl.v.kfbs(n = 8, tbloc = c(1, 2, 4), method = "Geary")
      tr(f) tr(f^2) sum rank norm
ttg_1   14     40  0    7 3.848
ttg_2   20    116  0    5 7.236

```



```

# leurs formes sont positives
# leurs traces sont non nulles
# leurs sommes sont nulles

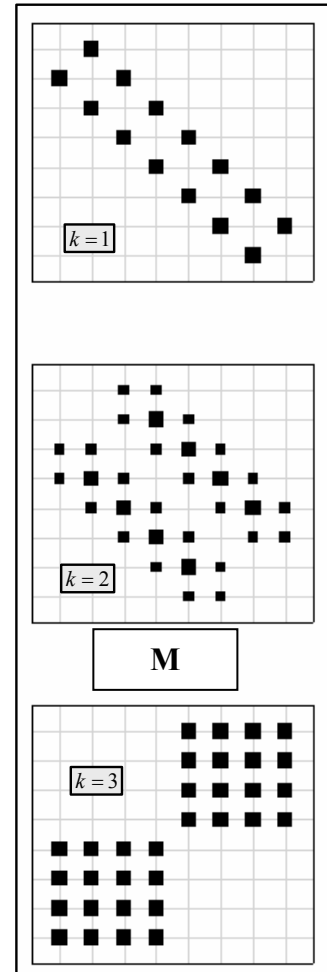
```

```
ttg_4      8      64  0      1 8.000
```

```
hill.moran.kfbs <- ttlv.kfbs(n = 8,
  tbloc = c(1,2,4), method = "Moran")
print(hill.moran.kfbs)
```

```
# correlogramme
# matrice de stockage
```

	[,1]	[,2]	[,3]
[1,]	0	0	0
[2,]	1	0	0
[3,]	0	1	0
[4,]	0	1	0
[5,]	0	0	1
[6,]	0	0	1
[7,]	0	0	1
[8,]	0	0	1
[9,]	0	0	0
[10,]	1	1	0
[11,]	0	2	0
[12,]	0	1	1
[13,]	0	0	1
[14,]	0	0	1
[15,]	0	0	1
[16,]	0	0	0
[17,]	1	1	0
[18,]	0	2	1
[19,]	0	1	1
[20,]	0	0	1
[21,]	0	0	1
[22,]	0	0	0
[23,]	1	1	1
[24,]	0	2	1
[25,]	0	1	1
[26,]	0	0	1
[27,]	0	0	0
[28,]	1	1	0
[29,]	0	2	0
[30,]	0	1	0
[31,]	0	0	0
[32,]	1	1	0
[33,]	0	1	0
[34,]	0	0	0
[35,]	1	0	0
[36,]	0	0	0



M

```
attr("npoints")
```

```
[1] 8
```

```
attr("nforms")
```

```
[1] 3
```

```
attr("scalprod")
```

```
[1] FALSE
```

```
attr("trace1")
```

```
[1] 0 0 0
```

```
attr("trace2")
```

```
[1] 14 56 32
```

```
attr("sum")
```

```
[1] 14 40 32
```

```
attr("rank")
```

```
[1] 8 6 2
```

```
attr("norm")
```

```
[1] 1.879 5.791 4.000
```

```
attr("call")
```

```
ttl.v.kfbs(n = 8, tbloc = c(1, 2, 4), method = "Moran")
```

```
attr("labels")
```

```
[1] "ttm_1" "ttm_2" "ttm_4"
```

```
attr("class")
```

```
[1] "kfbs"
```

```
# les formes sont non positives
```

```
# leurs traces sont nulles
```

```
# la trace de leurs carrés = sommes
```

```
summary(hill.moran.kfbs)
```

```
K bilinear symetric forms
```

```
Points : 8      Forms : 3
```

```
All positive forms : FALSE
Call : ttm.kfbs(n = 8, tbloc = c(1, 2, 4), method = "Moran")
      tr(f) tr(f^2) sum rank norm
ttm_1    0     14  14     8 1.879
ttm_2    0     56  40     6 5.791
ttm_4    0     32  32     2 4.000
```

4.7. Typologie d'un ensemble de formes bilinéaires

On a abordé quelques propriétés des formes bilinéaires symétriques. On peut s'intéresser également à la redondance implicite des points de vue développés dans une famille de K formes bilinéaires symétriques, voire à la redondance des points de vue développés par plusieurs familles. On s'intéresse alors à la matrice des produits scalaires entre formes bilinéaires (Lavit, 1988). Dans l'ensemble des formes bilinéaires symétriques de \mathbb{R}^n la fonction :

$$(\mathbf{A}_k, \mathbf{A}_{k'}) \rightarrow \text{Trace}(\mathbf{A}_k \mathbf{A}_{k'}) = \text{Trace}(\mathbf{A}_{k'} \mathbf{A}_k) = \sum_{i=1}^n \sum_{j=1}^n \mathbf{A}_{kij} \mathbf{A}_{k'ij}$$

définie, pour tout couple de matrices symétriques $(\mathbf{A}_k, \mathbf{A}_{k'})$, un produit scalaire euclidien. Les normes des matrices sont quelconques et pour comparer deux matrices on prendra leur cosinus :

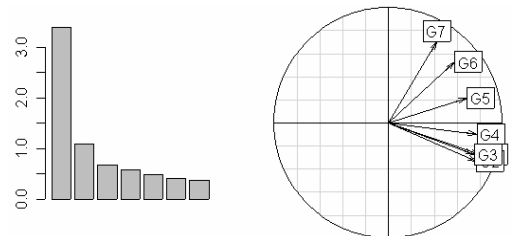
$$r_{kk'} = \frac{\text{Trace}(\mathbf{A}_k \mathbf{A}_{k'})}{\sqrt{\text{Trace}(\mathbf{A}_k^2) \text{Trace}(\mathbf{A}_{k'}^2)}}$$

La matrice R, de terme général $(r_{kk'})_{\substack{1 \leq k' \leq K \\ 1 \leq k \leq K}}$, se calcule avec la fonction `statis.kfbs(...)` (Annexe 2.21). La diagonalisation de cette matrice donne une image euclidienne des relations entre formes bilinéaires. On constate que :

- les formes de Geary/Lebart sont fortement redondantes

```
geary.statis <- statis.kfbs(geary.kfbs)
geary.statis$RV
```

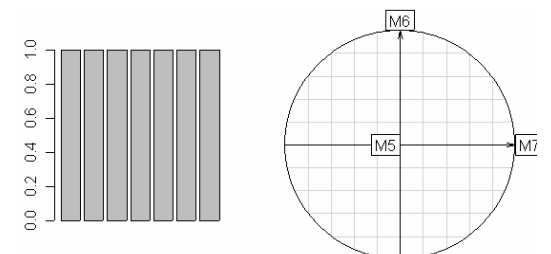
	G1	G2	G3	G4	G5	G6	G7
G1	1.0000	0.6149	0.5809	0.5534	0.4564	0.3354	0.1581
G2	0.6149	1.0000	0.5774	0.5303	0.4082	0.2500	0.1768
G3	0.5809	0.5774	1.0000	0.5103	0.3536	0.2887	0.2041
G4	0.5534	0.5303	0.5103	1.0000	0.4330	0.3536	0.2500
G5	0.4564	0.4082	0.3536	0.4330	1.0000	0.4082	0.2887
G6	0.3354	0.2500	0.2887	0.3536	0.4082	1.0000	0.3536
G7	0.1581	0.1768	0.2041	0.2500	0.2887	0.3536	1.0000



- les formes de Moran/Smouse sont, par définition, orthogonales

```
moran.statis <- statis.kfbs(moran.kfbs)
moran.statis$RV
```

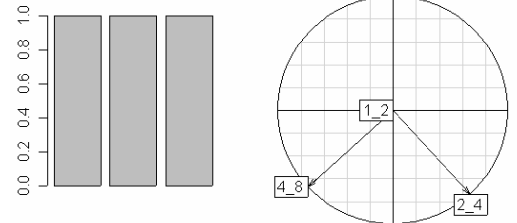
	M1	M2	M3	M4	M5	M6	M7
M1	1	0	0	0	0	0	0
M2	0	1	0	0	0	0	0



M3	0	0	1	0	0	0	0
M4	0	0	0	1	0	0	0
M5	0	0	0	0	1	0	0
M6	0	0	0	0	0	1	0
M7	0	0	0	0	0	0	1

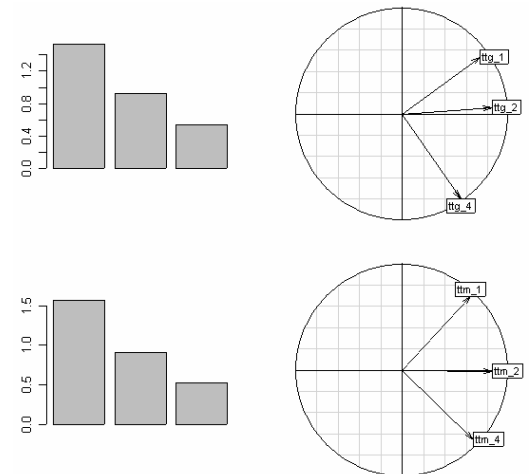
- les formes de Noy-Meir/Anderson constituent également une famille orthogonale car les espaces de projection sont, par définition, orthogonaux

```
noy.statis <- statis.kfbs(noy.kfbs)
round(noy.statis$RV, 4)
      1_2  2_4  4_8
1_2   1    0    0
2_4   0    1    0
4_8   0    0    1
```



- les formes de Hill constituent une famille redondante

```
hill.geary.statis <- statis.kfbs(hill.geary.kfbs)
hill.geary.statis$RV
      ttg_1  ttg_2  ttg_4
ttg_1  1.00000  0.41111  0.07906
ttg_2  0.41105  1.00000  0.27854
ttg_4  0.07906  0.27854  1.00000
hill.moran.statis <- statis.kfbs(hill.moran.kfbs)
hill.moran.statis$RV
      ttm_1  ttm_2  ttm_4
ttm_1  1.0000  0.3571  0.0945
ttm_2  0.3571  1.0000  0.3780
ttm_4  0.0945  0.3780  1.0000
```



Les propriétés des familles de K formes bilinéaires symétriques sont donc très variables. On peut s'intéresser à la redondance des points de vue définis par chaque famille. On réalise alors une typologie sur l'ensemble des formes bilinéaires que l'on vient de définir indépendamment de la famille de départ. L'image euclidienne que l'on obtient est très parlante (Figure 2.9).

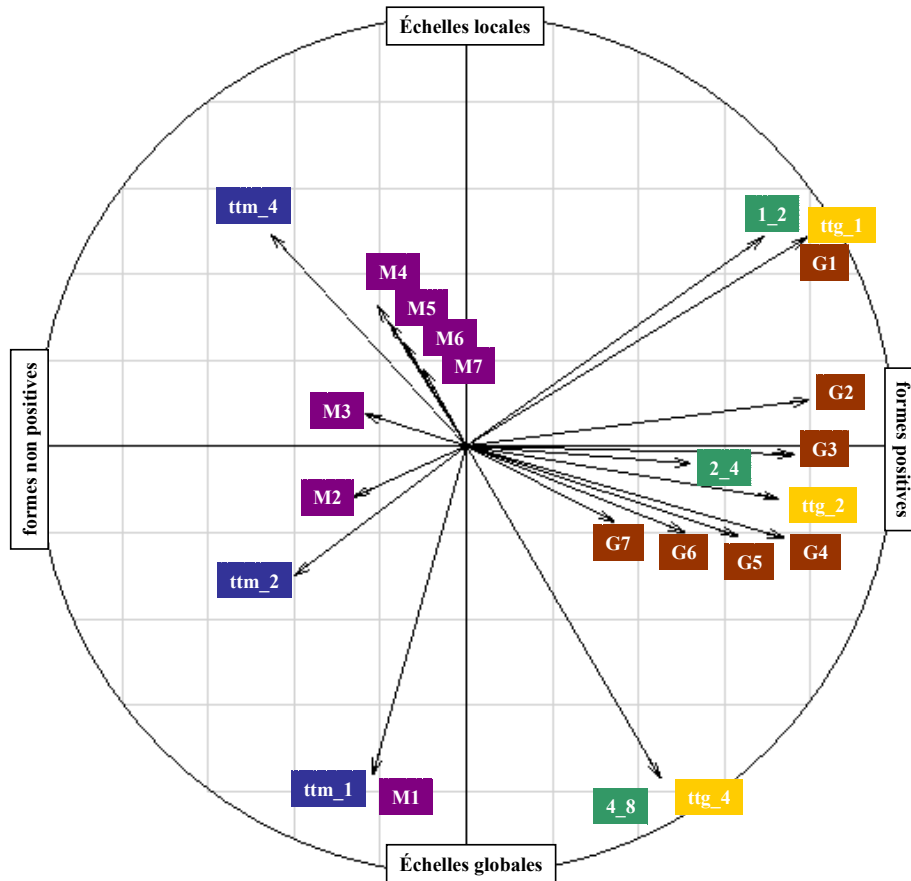


Figure 2.9 : Cercle des corrélations entre formes bilinéaires symétriques : les étiquettes bleues correspondent aux formes de Hill version Moran, les jaunes à celles de Hill version Geary, les marron à celles de Geary, les violettes à celles de Moran, et les vertes à celles de Noy-Meir.

L'axe 1 oppose les formes bilinéaires positives aux formes bilinéaires non positives. L'axe 2 sépare les formes selon une logique d'échelles. La logique définie par la famille de Noy-Meir/Anderson est la même que celle associée à la famille de Hill. Les formes de Geary/Moran ont un fonctionnement différent : elles sont particulièrement redondantes. On peut se demander quelle famille sera la plus pertinente pour décrire la structure d'une variable à différentes échelles. *A priori*, les meilleures sont les familles orthogonales.

5. BASES ORTHONORMÉES ET FAMILLES DE K PROJECTEURS

5.1. Définitions

Soit \mathbf{B} une base \mathbf{D} -orthonormée de \mathbb{R}^n . Elle vérifie $\mathbf{B}'\mathbf{D}\mathbf{B} = \mathbf{Id}_n$. On s'assure dans un premier temps que tous les vecteurs de la base \mathbf{B} sont \mathbf{D} -orthogonaux au vecteur $\mathbf{1}_n$, c'est-à-dire centrés pour la pondération \mathbf{D} . En effet, en statistique, ce qui se passe sur une variable

constante est généralement sans intérêt du point de vue de l'étude de sa structure. On récupère alors simplement les vecteurs propres \mathbf{D} -orthogonaux au vecteur $\mathbf{1}_n$ par orthonormalisation de Gram-Schmidt. On élimine ensuite le vecteur $\mathbf{1}_n$ de la base orthonormée et l'on travaille exclusivement sur des variables centrées pour la pondération \mathbf{D} . La base \mathbf{B} , obtenue après orthonormalisation, n'a donc plus que $n-1$ vecteurs.

Soit \mathbf{x} un vecteur de \mathbb{R}^n . On appellera \mathbf{x}_0 sa version centrée pour la pondération \mathbf{D} :

$$\mathbf{x}_0 = \mathbf{x} - \mathbf{1}_n m(\mathbf{x}) = (\mathbf{Id}_n - \mathbf{U}_{nn} \mathbf{D}) \mathbf{x} \text{ avec } m(\mathbf{x}) = \mathbf{1}_n^t \mathbf{D} \mathbf{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

et $\tilde{\mathbf{x}}$ sa version standardisée :

$$\tilde{\mathbf{x}} = \frac{\mathbf{x}_0}{\sqrt{v(\mathbf{x})}} \text{ avec } v(\mathbf{x}) = \mathbf{x}_0^t \mathbf{D} \mathbf{x}_0 = \frac{1}{n} \sum_{i=1}^n (x_i - m(\mathbf{x}))^2$$

Chaque vecteur $\tilde{\mathbf{x}}$ admet une décomposition unique sur les vecteurs de la base \mathbf{B} :

$$\tilde{\mathbf{x}} = \sum_{i=1}^{n-1} \langle \tilde{\mathbf{x}} | \mathbf{b}_i \rangle_D \mathbf{b}_i = \sum_{i=1}^{n-1} r_i \mathbf{b}_i$$

Les corrélations r_i sont définies par le vecteur

$$\mathbf{r} = (r_i)_{1 \leq i \leq n-1} = \mathbf{B}' \mathbf{D} \tilde{\mathbf{x}}.$$

Elles correspondent aux coefficients de la transformée orthonormale de $\tilde{\mathbf{x}}$ par \mathbf{B} (Percival & Walden, 2000) qui donne une décomposition canonique de la variance :

$$\|\tilde{\mathbf{x}}\|_{\mathbf{D}}^2 = 1 = \tilde{\mathbf{x}}^t \mathbf{D} \tilde{\mathbf{x}} = (\mathbf{B} \mathbf{r})^t \mathbf{D} \mathbf{B} \mathbf{r} = \mathbf{r} \mathbf{B}' \mathbf{D} \mathbf{B} \mathbf{r} = \mathbf{r}^t \mathbf{r} = \sum_{i=1}^n r_i^2$$

La variance d'une variable se décompose donc en somme de carrés de corrélation. Ces carrés de corrélation déterminent dans une certaine mesure la structure d'une variable par rapport à un ensemble de figures de références. Cette décomposition de la variance par les vecteurs d'une base orthonormée est un cas particulier de la décomposition de la variance par une famille de K formes bilinéaires symétriques. En effet, on peut facilement définir une famille de K projecteurs à partir d'une base orthonormée, la plus simple d'entre elles étant formée par les $n-1$ projecteurs $\left(\mathbf{\Pi}_k = \frac{1}{n} \mathbf{b}_k \mathbf{b}_k^t \right)_{1 \leq k \leq n-1}$. De manière générale, toute partition de l'ensemble

des $n-1$ vecteurs de la base définit une famille de projecteurs. Si l'on note \mathbf{B}_E , la matrice constituée par l'ensemble E des $\text{card}(E)$ vecteurs de \mathbf{B} , toute partition

$E_1 \cup E_2 \cup \dots \cup E_K = \{1, 2, \dots, n-1\}$ définit une famille de projecteurs $\left(\mathbf{\Pi}_k = \frac{1}{n} \mathbf{B}_{E_k} \mathbf{B}_{E_k}^t \right)_{1 \leq k \leq K}$. Il existe donc un lien étroit entre les méthodes utilisant des familles de formes bilinéaires symétriques et celles qui introduisent des bases orthonormées. En particulier, $\frac{1}{n} \mathbf{B} \mathbf{B}^t$ est le projecteur (pour la métrique $\mathbf{D} = \frac{1}{n} \mathbf{I}_n$) sur l'orthogonal de $\mathbf{1}_n$.

Inversement, on a vu dans le paragraphe précédent que les matrices symétriques \mathbf{A}_k donnent une ou plusieurs bases \mathbf{D} -orthonormées de vecteurs \mathbf{A}_k -orthogonaux. Les vecteurs propres de \mathbf{A}_k forment alors une famille de figures de références (« *templates* ») ordonnées depuis celles dont les valeurs pour la forme considérée sont les plus grandes possibles (les premières valeurs propres) vers celles dont les valeurs sont les plus petites possibles (les dernières valeurs propres). Il existe bien d'autres manières de définir des figures de références. L'analyse spectrale, au même titre que l'analyse en ondelettes, introduit des familles de figures de références (Percival, 1993) permettant la décomposition canonique des variables et de leurs variances.

Dans la suite de l'exposé, on présente les différentes familles de figures de références utilisées pour décomposer la variance d'une variable de \mathbb{R}^n dont on veut caractériser la structure interne. Seule la pondération uniforme $\mathbf{D} = \text{diag}(1/n, \dots, 1/n)$ sera envisagée.

5.2. La classe d'objets 'orthobasis'

On va définir différentes familles de figures de références. On a besoin d'une structure pour manipuler ces objets. On définit alors une nouvelle classe d'objets dans l'environnement du logiciel R que l'on appelle 'orthobasis' (Annexe 2.15). On introduit également un ensemble de fonctions qui vont permettre leur manipulation. Les figures de références sont rangées dans une matrice à n lignes et $n-1$ colonnes. Elles correspondent aux $n-1$ vecteurs \mathbf{D} -orthonormés de la base \mathbf{B} considérée. Les attributs d'un objet de la classe 'orthobasis' sont :

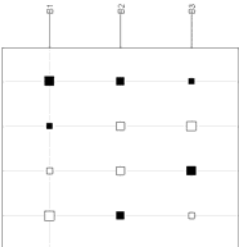
- 'names' pour les noms des vecteurs de la base
- 'row.names' pour les noms des unités statistiques
- 'class' pour la classe de l'objet

- ‘values’ pour les valeurs propres associées aux vecteurs de la base, lorsque ces derniers sont obtenus par diagonalisation d’un opérateur de voisinage. Les vecteurs sont alors rangés par ordre décroissant de leurs valeurs propres.
- ‘weights’ pour la pondération **D** considérée
- ‘call’ rappelle la ligne de commande qui a permis la création de l’objet

Les objets de la classe ‘orthobasis’ sont liés aux objets de la classe ‘kfbs’. On passe facilement d’une base orthonormée munie d’une partition à une famille de K formes bilinéaires symétriques par la fonction `orthobasis2kfbs(...)` (Annexe 2.5 et Annexe 2.9) :

```

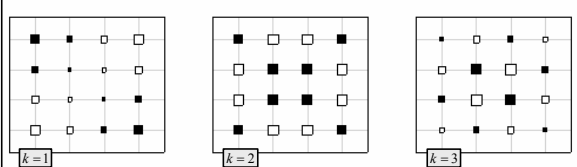
orthobas <- orthobasis.line(4)
orthobas
Orthonormal basis: data.frame with 4 rows and 3 columns
-----
Columns are an orthonormal basis of ln-orthogonal for
the inner product defined by the weights attribute
-----
names = B1 ... B3
row.names = u1 ... u4
weights = 0.25 ... 0.25
values = 0.3738 ... -1.040
class = orthobasis data.frame
call =orthobasis.line(n = 4)
orthobas*1
      B1 B2      B3
u1  1.3066  1  0.5412
u2  0.5412 -1 -1.3066
u3 -0.5412 -1  1.3066
u4 -1.3066  1 -0.5412
    
```



```

level <- as.factor(1:3)          # partition {1,2,3}={1}∪{2}∪{3}=E1∪E2∪E3
kfbs <- orthobasis2kfbs(orthobas, level)
summary(kfbs)
summary(kfbs)
K bilinear symetric forms
Points : 4      Forms : 3
All positive forms : TRUE
    
```

	tr(f)	tr(f^2)	sum	rank	norm
1	1	1	0.000e+00	1	1
2	1	1	-5.551e-17	1	1
3	1	1	-1.110e-16	1	1



5.3. Les bases associées à la diagonalisation des matrices symétriques

Toute matrice symétrique **A** donne une ou plusieurs bases **D**-orthonormées de vecteurs **A**-orthogonaux. Les vecteurs propres de **A** forment alors une famille de figures de références que l’on ordonne, par convention, depuis celles dont l’autocorrelation est maximale vers celles dont l’autocorrelation est minimale. Cette pratique est bien connue pour les

matrices binaires associés aux graphes de voisinage (Méot et al., 1993), pour les matrices de pondération de voisinage (Griffith, 2000), pour les matrices de proximité et pour les carrés des matrices de distances euclidiennes doublement centrée (Borcard et Legendre, 2002). On a implémenté dans R plusieurs fonctions permettant d'obtenir ces figures de références.

5.3.1. La fonction `orthobasis.mat(...)`

Elle définit une base orthonormée à partir d'une matrice de proximité symétrique \mathbf{A} . Une matrice de proximité \mathbf{A} , de terme général a_{ij} , est le centre d'une curieuse contradiction. La valeur du terme général est, par définition, d'autant plus grande que les points sont plus proches. Mais on décrète qu'un point n'est jamais proche de lui-même ($a_{ii} = 0$) ou au contraire que le maximum de la proximité se fait entre un point et lui-même ($a_{ii} = a_{\max}$). De toute manière, pour la forme de Geary, cela n'a aucune importance. Mais la différence apparaît dans celle de Moran. Le lien entre les deux se retrouve dans leur somme qui est le produit scalaire associé à la pondération marginale de proximité. Les valeurs a_{ii} jouent un rôle dans cette pondération. Nous avons déjà évoqué, que, pour pratiquer des tests d'hypothèse, il est très intéressant que cette pondération soit uniforme car c'est le seul cas qui laisse invariant la moyenne et la variance des observations par permutation des données. Nous dirons donc qu'une matrice symétrique \mathbf{A} a la propriété de pondération uniforme 'cnw' (constant neighboring weights) si :

$$\mathbf{A}\mathbf{1}_n = \rho\mathbf{1}_n$$

Pour obtenir une forme 'cnw' à partir d'une forme quelconque, il suffit de poser :

$$1) \quad w_i = \sum_{\substack{j=1 \\ j \neq i}}^n a_{ij} \quad 2) \quad w_{\max} = \max(w_i) \quad 3) \quad a_{ii} = w_{\max} - w_i \quad ()$$

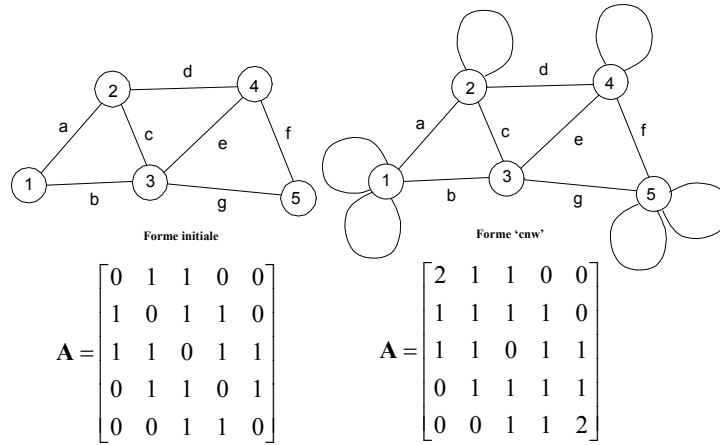


Figure 2.10 : Matrice de proximité binaire associée au graphe sous sa forme initiale (**à gauche**). Matrice de proximité associée au même graphe complété sous sa forme 'cnw' (**à droite**).

Pour éliminer la question des unités on posera alors que $\rho = 1/n$ ce qui est obtenu par :

$$4) \quad \mathbf{P} = \frac{\mathbf{A}}{\mathbf{1}'_n \mathbf{A} \mathbf{1}_n}$$

$\mathbf{P} = [p_{ij}]$ est alors une distribution de fréquence bivariée symétrique dont les pondérations marginales sont uniformes. On intègre alors le double centrage, afin d'obtenir des vecteurs propres orthonormés pour la pondération uniforme :

$$5) \quad \mathbf{A}_{cnw} = \left(\mathbf{I}_n - \frac{1}{n} \mathbf{U}_n \right) \mathbf{P} \left(\mathbf{I}_n - \frac{1}{n} \mathbf{U}_n \right) = \mathbf{P} - \frac{1}{n^2} \mathbf{U}_n$$

Alors, la forme quadratique $\mathbf{x}' \mathbf{A}_{cnw} \mathbf{x}$ est celle d'une autocorrélation de Moran et mesure $\mathbf{x}' \mathbf{A}_{cnw} \mathbf{x}$ mesure l'écart entre la variance ordinaire :

$$\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \frac{1}{n^2} (x_i - x_j)^2$$

et la variance de proximité :

$$\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n p_{ij} (x_i - x_j)^2$$

On retrouve alors au travers de la forme quadratique $\mathbf{x}' \mathbf{A}_{cnw} \mathbf{x}$, la statistique de Moran utilisée par la procédure `gearymoran(...)` (Annexe 2.3) associée à \mathbf{A} . Lorsque \mathbf{x} est centré et réduit, $\tilde{\mathbf{x}}' \mathbf{A}_{cnw} \tilde{\mathbf{x}} = \tilde{\mathbf{x}}' \mathbf{P} \tilde{\mathbf{x}}$: la valeur de la forme quadratique est un coefficient de corrélation compris entre -1 et 1. La valeur de cette forme a un sens naturel pour les vecteurs centrés et normés pour la pondération uniforme. Cette forme 'cnw' admet au moins une base de vecteurs propres orthonormés pour la pondération uniforme :

$$\mathbf{A}_{cnw} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^t = \mathbf{U}_r\mathbf{\Lambda}_r\mathbf{U}_r^t = \frac{1}{n}(\sqrt{n}\mathbf{U}_r)(n\mathbf{\Lambda}_r)(\sqrt{n}\mathbf{U}_r^t)\frac{1}{n} = \frac{1}{n}\mathbf{S}_r\mathbf{\Omega}_r\mathbf{S}_r^t\frac{1}{n}$$

En écrivant :

$$\mathbf{A}_{cnw} = \frac{1}{n}\mathbf{S}_r\mathbf{\Omega}_r\mathbf{S}_r^t\frac{1}{n}$$

nous avons dans les colonnes de \mathbf{S}_r des scores numériques centrés (car $\mathbf{1}_n$ est toujours vecteur propre pour 0, les autres étant orthogonaux à $\mathbf{1}_n$ donc centrés) et réduits (leur variance vaut 1). En outre :

$$\mathbf{S}_r^t\mathbf{A}_{cnw}\mathbf{S}_r = \mathbf{S}_r^t\frac{1}{n}\mathbf{S}_r\mathbf{\Omega}_r\mathbf{S}_r^t\frac{1}{n}\mathbf{S}_r = \mathbf{I}_r\mathbf{\Omega}_r\mathbf{I}_r = \mathbf{\Omega}_r$$

Les éléments ω_k de la matrice diagonale $\mathbf{\Omega}_r$ sont les valeurs de la forme pour les scores et sont donc des coefficients de corrélation rangés par ordre décroissant. Les colonnes de \mathbf{S}_r sont donc des variables centrées et réduites maximisant successivement, sous contrainte d'orthogonalité, la forme de Moran. Elles définissent les figures de référence de la matrice de proximité \mathbf{A} et les ω_k les valeurs de référence correspondantes. L'examen de ces éléments permet de comprendre avec précision ce que mesure la matrice \mathbf{A} d'origine.

Exemple :

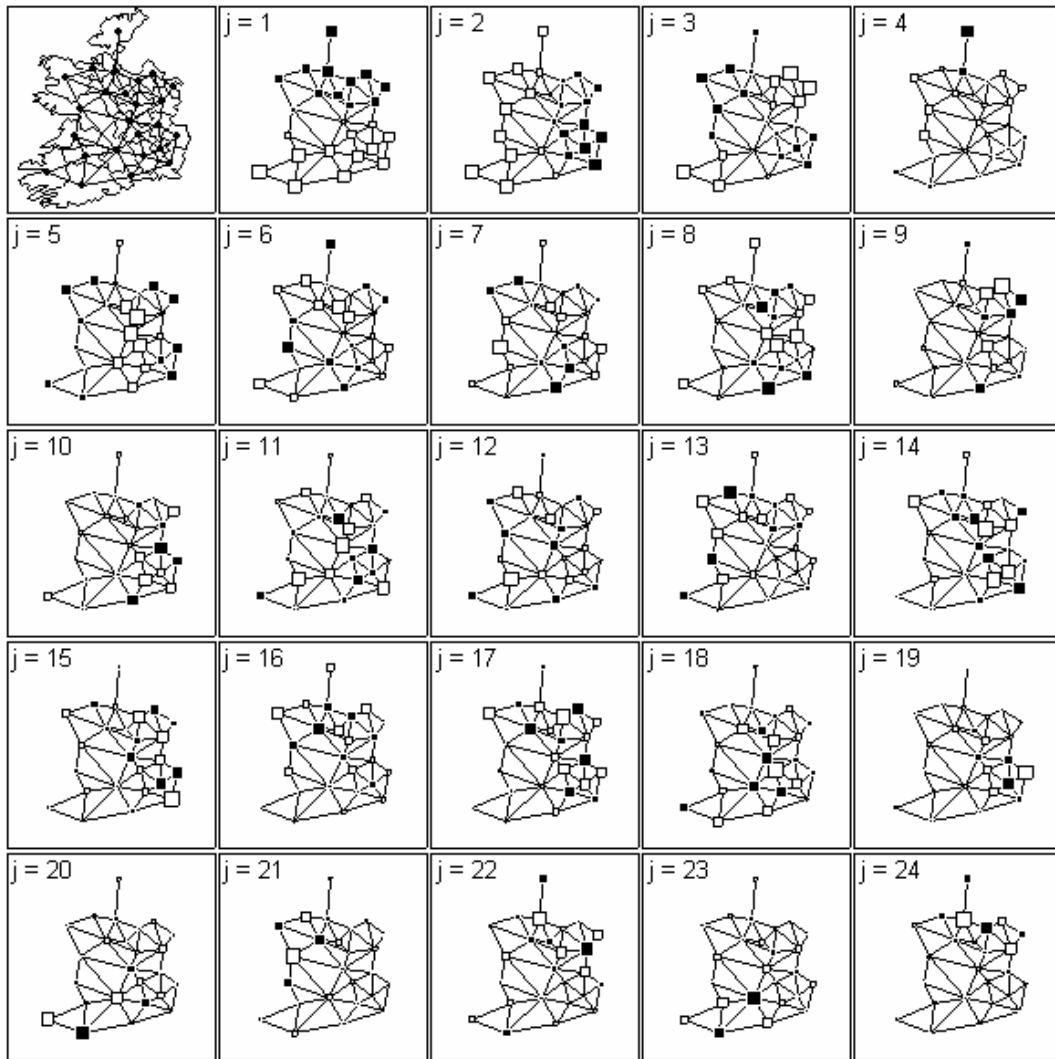


Figure 2.11 : Base orthonormée associée à la matrice des pondérations de voisinage (normalisée par lignes), définie à partir de la longueur de la frontière entre deux comtés voisins. Plus la frontière qu'ils partagent est grande, plus le lien entre voisins est important. Les premiers vecteurs maximisent l'indice de Moran. Les derniers vecteurs maximisent l'indice de Geary.

5.3.2. La fonction `orthobasis.neig(...)`

Elle définit une base orthonormée à partir d'une matrice de voisinage binaire \mathbf{A} . On diagonalise l'opérateur de voisinage $(\mathbf{P}-\mathbf{F})\mathbf{D}$ introduit par Méot et al. (1993), avec $\mathbf{F} = \mathbf{A}/\mathbf{1}'_n \mathbf{A} \mathbf{1}_n$ et $\mathbf{P} = \text{diag}(\mathbf{F}\mathbf{1}_n)$. On calcule ensuite les valeurs propres de l'opérateur de lissage conjugué et l'on ordonne les vecteurs propres par variance locale croissante.

Exemple :

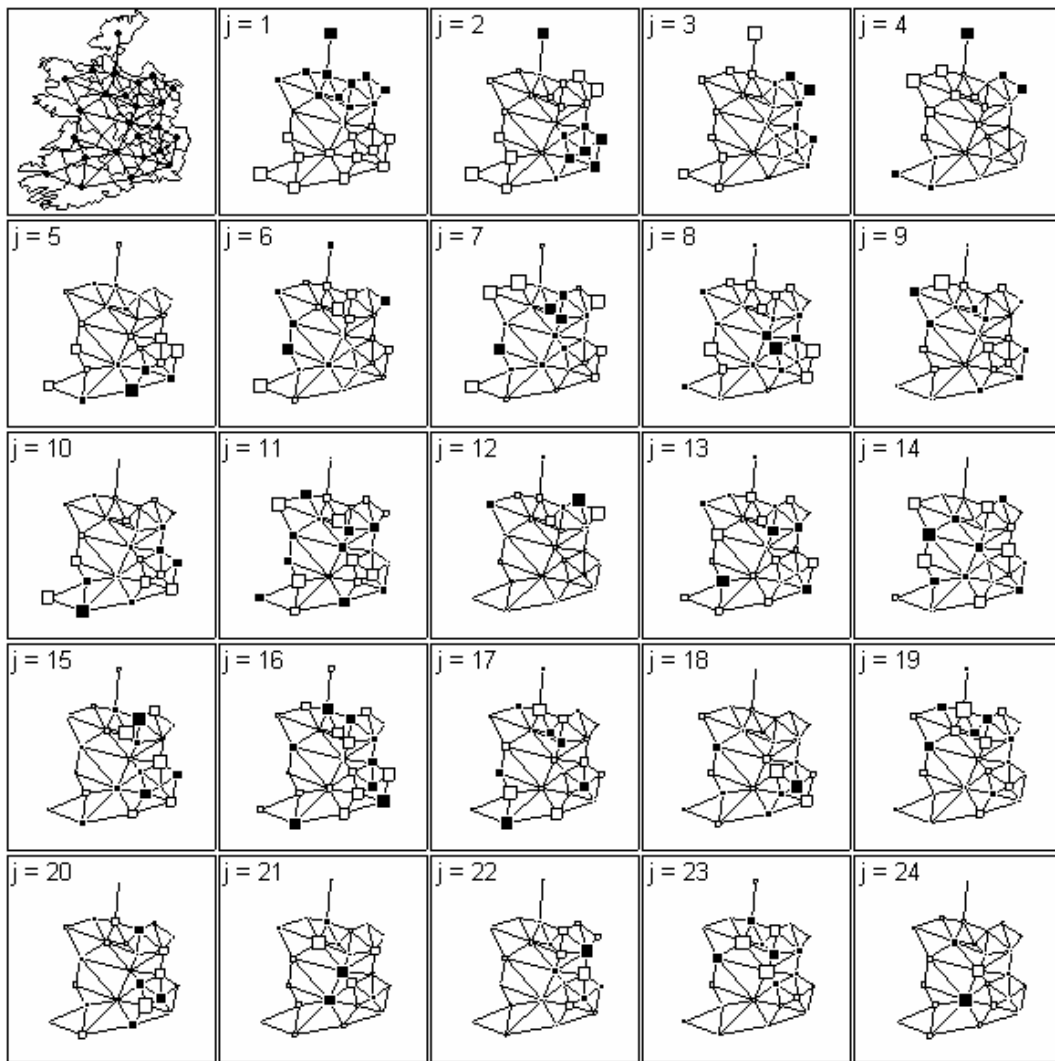


Figure 2.12 : Base orthonormée associée à la matrice de voisinage binaire.

5.4. Expression analytique des vecteurs propres de l'opérateur de Méot

Pour certains graphes particuliers, il est possible d'obtenir l'expression littérale des valeurs propres et des vecteurs propres des opérateurs de voisinage ((Griffith, 2000)). Le principal intérêt de cette opération est de pouvoir obtenir les valeurs propres et les vecteurs propres des opérateurs sans avoir à diagonaliser la matrice qui leur est associée. En particulier, lorsque l'on travaille sur des données avec beaucoup d'unités statistiques telles que les images satellites, la diagonalisation de cette matrice est impossible numériquement.

Les expressions analytiques sont connues pour trois types de graphes et l'opérateur de voisinage $\mathbf{E} = (\mathbf{P} - \mathbf{F})\mathbf{D}$ (Méot et al., 1993) qui leur est associé. Elles sont données dans Méot et al. (1993) puis démontrées dans la thèse de Cornillon (1998). Toutefois, dans les deux publications, des erreurs ont été commises dans l'écriture des expressions analytiques, bien

que la démonstration reste juste. On a donc repris ces résultats en y apportant les corrections nécessaires. La pondération \mathbf{D} est à nouveau uniforme.

5.4.1. Valeurs propres et vecteurs propres dans le cadre d'un graphe linéaire

Soit un graphe linéaire à l'ordre 1. Les points sont répartis sur une droite ; un point est seulement voisin du suivant et du précédent. Les vecteurs propres et les valeurs propres de l'opérateur de voisinage \mathbf{E} associé au graphe linéaire vérifient

$$\mathbf{E}\mathbf{y}^j = \lambda^j \mathbf{y}^j \text{ avec } \begin{cases} (\mathbf{y}^j)^t = (y_1^j, \dots, y_i^j, \dots, y_n^j) \\ 1 \leq j \leq n-1 \end{cases}$$

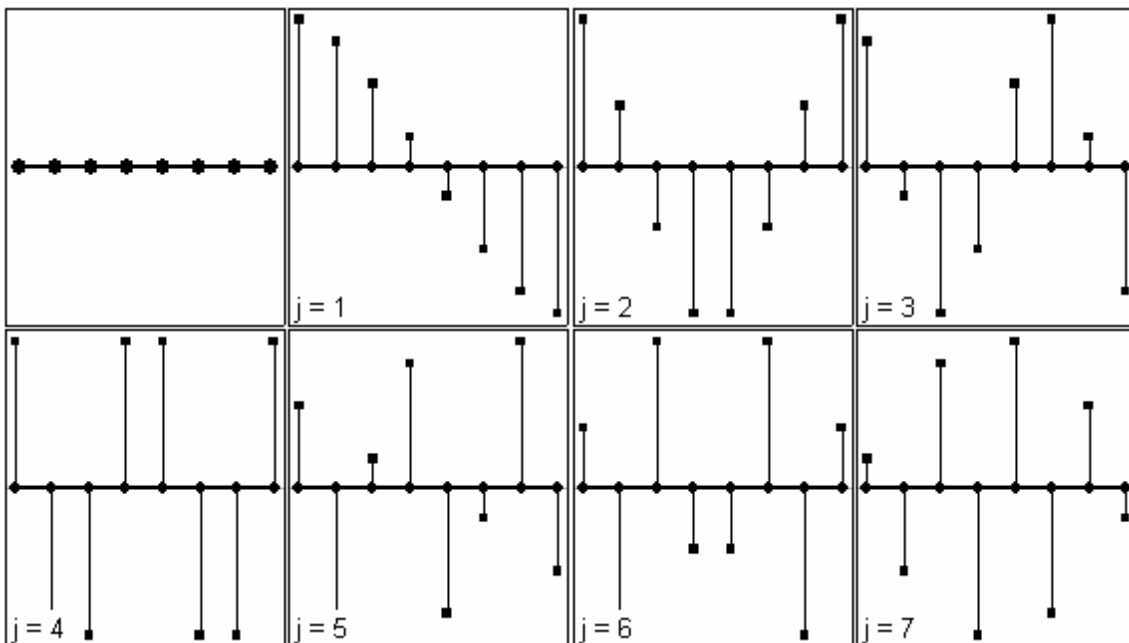
Ils s'écrivent

$$\lambda^j = \frac{4}{n} \sin^2 \left(\frac{j\pi}{2n} \right)$$

où A est une constante de normalisation

$$y_i^j = A \cos \left(\frac{j\pi}{2n} (2i-1) \right)$$

Exemple :



Paramètres

$n = 8$
 $1 \leq j \leq 7$

neig(n.line = ...)

1 2 3 4 5 6 7 8

Opérateur de voisinage nE

	1	2	3	4	5	6	7	8
1	1	-1	0	0	0	0	0	0
2	-1	2	-1	0	0	0	0	0
3	0	-1	2	-1	0	0	0	0
4	0	0	-1	2	-1	0	0	0
5	0	0	0	-1	2	-1	0	0
6	0	0	0	0	-1	2	-1	0
7	0	0	0	0	0	-1	2	-1
8	0	0	0	0	0	0	-1	1

Vecteurs propres

$y_i^j = A \cos\left(\frac{j\pi}{2n}(2i-1)\right)$

		0	1	2	3	j	4	5	6	7
1	1									
2	2									
3	3									
4	4									
i							(y_i^j)			
5	5									
6	6									
7	7									
8	8									

Valeurs propres

$\lambda^j = \frac{4}{n} \sin^2\left(\frac{j\pi}{2n}\right)$

			(λ^j)			
--	--	--	---------------	--	--	--

5.4.2. Valeurs propres et vecteurs propres dans le cadre d'un graphe circulaire

Soit un graphe circulaire à l'ordre 1. Les points sont répartis sur un cercle ; un point est seulement voisin du suivant et du précédent. Les vecteurs propres et les valeurs propres de l'opérateur de voisinage \mathbf{E} associé au graphe circulaire vérifient :

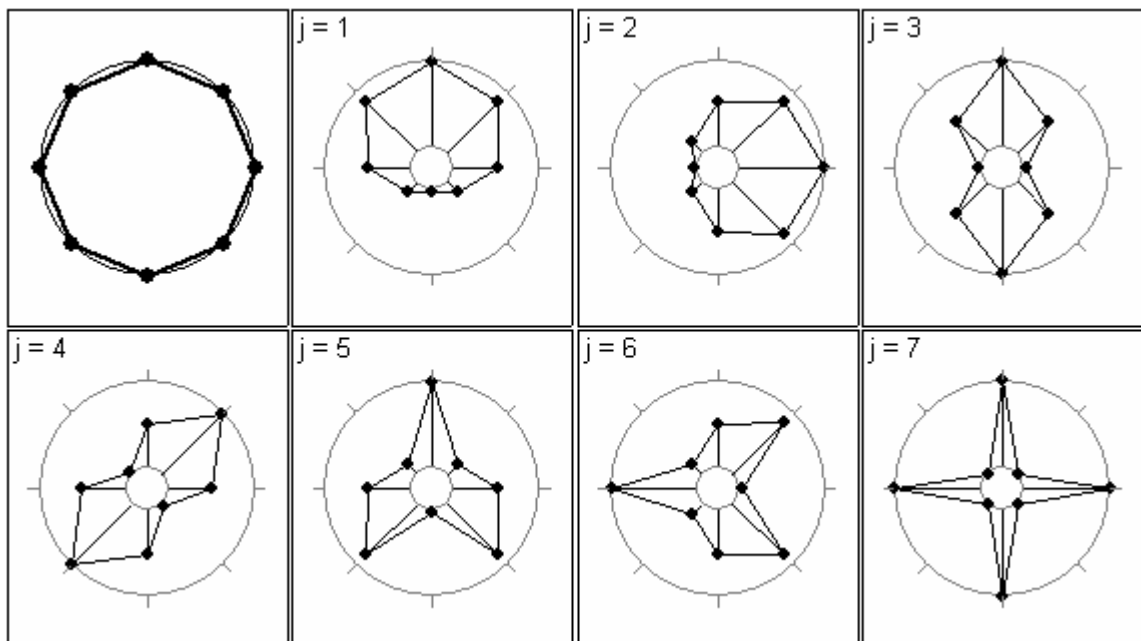
$$\mathbf{E}\mathbf{y}^j = \lambda^j \mathbf{y}^j \text{ avec } \begin{cases} (\mathbf{y}^j)^t = (y_1^j, \dots, y_i^j, \dots, y_n^j) \\ 1 \leq j \leq n-1 \end{cases}$$

Ils s'écrivent

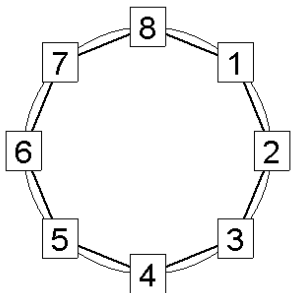
$$\lambda^j = \frac{4}{n} \sin^2 \left(\left[\frac{j+1}{2} \right] \pi / n \right) \text{ et } \begin{cases} y_i^j = A \cos \left(\left[\frac{j+1}{2} \right] 2\pi i / n \right) \text{ si } j \text{ est impair} \\ y_i^j = A \sin \left(\left[\frac{j+1}{2} \right] 2\pi i / n \right) \text{ si } j \text{ est pair} \end{cases}$$

où A est une constante de normalisation et $[\]$ est l'opérateur partie entière.

Exemple :



neig(n.circle = ...)



Paramètres
 $n = 8$
 $1 \leq j \leq 7$

Opérateur de voisinage nE

	1	2	3	4	5	6	7	8
1	2	-1	-1	0	0	0	0	-1
2	-1	2	-1	0	0	0	0	0
3	0	-1	2	-1	0	0	0	0
4	0	0	-1	2	-1	0	0	0
5	0	0	0	-1	2	-1	0	0
6	0	0	0	0	-1	2	-1	0
7	0	0	0	0	0	-1	2	-1
8	-1	0	0	0	0	0	-1	2

Vecteurs propres

$y_i^j = A \cos\left(\left[\frac{j+1}{2}\right] 2\pi i / n\right)$ si j est impair

$y_i^j = A \sin\left(\left[\frac{j+1}{2}\right] 2\pi i / n\right)$ si j est pair

		0	1	2	3	j	4	5	6	7
1	1									
2	2									
3	3									
4	4									
	i						(y_i^j)			
5	5									
6	6									
7	7									
8	8									

Valeurs propres

			(λ^j)			
--	--	--	---------------	--	--	--

$\lambda^j = \frac{4}{n} \sin^2\left(\left[\frac{j+1}{2}\right] \pi / n\right)$

5.4.3. Valeurs propres et vecteurs propres dans le cas d'un graphe de type grille complète, pour une relation de la tour à l'ordre 1

Soit un graphe défini par une grille avec la relation de la tour à l'ordre 1. r représente le nombre de lignes, c le nombre de colonnes et $n = l \times m$ le nombre de nœuds de la grille.

Les vecteurs propres et les valeurs propres de l'opérateur de voisinage E associé au graphe vérifient

$$\mathbf{E}\mathbf{y}^j = \lambda^j \mathbf{y}^j = \lambda^{(p,q)} \mathbf{y}^{(p,q)} \text{ avec } \begin{cases} (\mathbf{y}^j)^t = (y_1^j, \dots, y_i^j, \dots, y_n^j) \\ 1 \leq j \leq n-1 \\ 0 \leq p \leq r-1 \\ 0 \leq q \leq c-1 \end{cases}$$

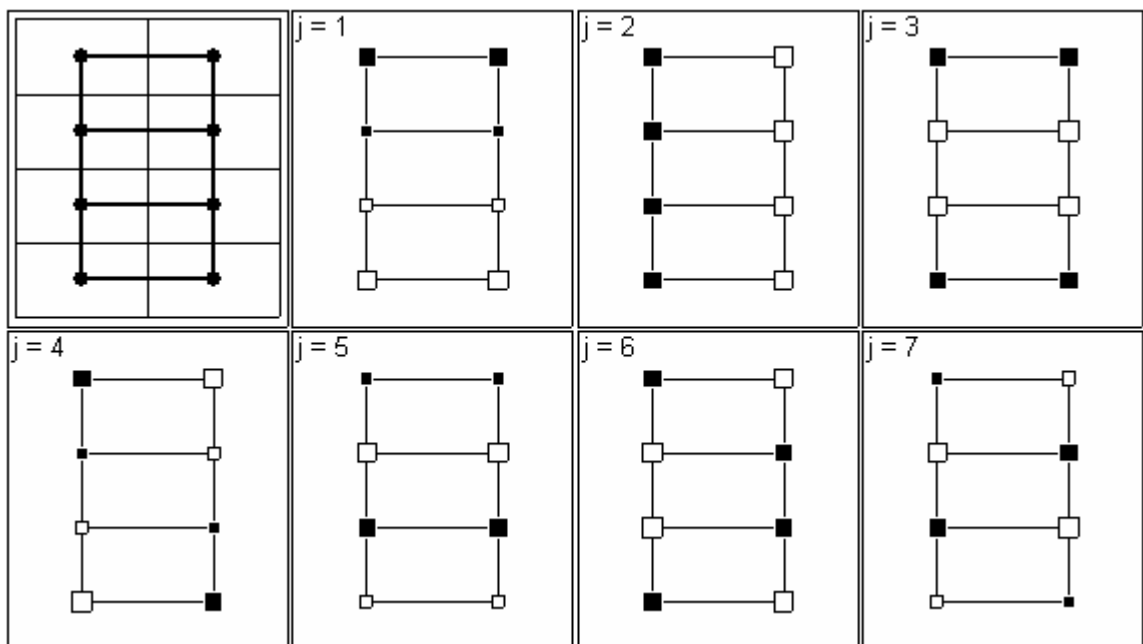
Ils s'écrivent

$$\lambda^{(p,q)} = \frac{4}{n} \left(\sin^2 \left(\frac{p\pi}{2r} \right) + \sin^2 \left(\frac{q\pi}{2c} \right) \right)$$

$$\mathbf{y}_{(p,q)}^{(p,q)} = A \left(\cos \left(\frac{p\pi}{r} (p+1-0.5) \right) \cos \left(\frac{q\pi}{c} (q+1-0.5) \right) \right)$$

où A est une constante de normalisation.

Exemple :



gridrowcol (...)

Paramètres

$r = 4$
 $c = 2$
 $n = 8$
 $1 \leq i \leq 8$
 $1 \leq j \leq 7$
 $0 \leq p \leq r-1$
 $0 \leq q \leq c-1$

Opérateur de voisinage nE

	<small>R1C2</small>	<small>R2C2</small>	<small>R3C2</small>	<small>R4C2</small>				
<small>R1C1</small>	2	-1	-1	0	0	0	0	0
<small>R1C2</small>	-1	2	0	-1	0	0	0	0
<small>R2C1</small>	-1	0	3	-1	-1	0	0	0
<small>R2C2</small>	0	-1	-1	3	0	-1	0	0
<small>R3C1</small>	0	0	-1	0	3	-1	-1	0
<small>R3C2</small>	0	0	0	-1	-1	3	0	-1
<small>R4C1</small>	0	0	0	0	-1	0	2	-1
<small>R4C2</small>	0	0	0	0	0	-1	-1	2

Vecteurs propres

$$y_{(p,q)}^{(p,q)} = A \cos\left(\frac{p\pi}{r}(p+1-0.5)\right) \cos\left(\frac{q\pi}{c}(q+1-0.5)\right)$$

	0	1	2	3	<small>j</small>	4	5	6	7
	(0,0)	(0,1)	(1,0)	(1,1)	<small>(p,q)</small>	(2,0)	(2,1)	(3,0)	(3,1)

--	--	--	--	--	--	--	--	--	--

Valeurs propres

$$\lambda^{(p,q)} = \frac{4}{n} \left(\sin^2\left(\frac{p\pi}{2r}\right) + \sin^2\left(\frac{q\pi}{2c}\right) \right)$$

5.5. La base associée à l'analyse spectrale à une dimension

L'analyse spectrale est une des techniques d'analyse du signal la plus ancienne et la plus utilisée en sciences. Les concepts associés à l'analyse spectrale sont relativement anciens et sa première mise en œuvre remonte aux études de Stokes (1879) puis Schuster (1898) sur la recherche de périodicités en météorologie. Depuis, elle a été utilisée dans presque tous les champs scientifiques y compris l'écologie statistique (Couteron, 2002; Muggleston & Renshaw, 1996; Renshaw, 1997; Ripley, 1978) suite à un regain d'intérêt dans les années 50, lié à la redécouverte de l'algorithme FFT, d'une part, et au développement du calcul informatique d'autre part (Percival, 1993). Contrairement à d'autres domaines, l'usage par des thématiciens relevant de l'écologie ou de l'observation de la terre est néanmoins resté modeste.

Ce paragraphe a pour objectif de présenter brièvement l'analyse spectrale à une dimension en insistant sur l'idée centrale plutôt que sur les détails. Pour une présentation complète, le lecteur pourra entre autres consulter les ouvrages de Jenkins et Watt (1968), Priestley (1981) ou Percival et Walden (2000). Selon Percival (1993), « *the basic idea behind spectral analysis is to decompose the variance of a series into a number of components, each one of which can be associated with a particular frequency* ». La transformation qui assure la décomposition d'une série en un ensemble de composantes est appelée transformée de Fourier. Elle définit les $n-1$ vecteurs de la base \mathbf{B} associée à l'analyse spectrale comme étant des fonctions trigonométriques de la forme :

$$\left. \begin{aligned} c_i^k &= \sqrt{2} \cos(\omega_k i) = \sqrt{2} \cos\left(\frac{2k\pi}{n} i\right) \\ s_i^k &= \sqrt{2} \sin(\omega_k i) = \sqrt{2} \sin\left(\frac{2k\pi}{n} i\right) \end{aligned} \right\} \text{ avec } 1 \leq k \leq \left\lfloor \frac{n-1}{2} \right\rfloor$$

$$c_i^{\left\lfloor \frac{n+1}{2} \right\rfloor} = \cos(\pi i) \quad \text{si il existe, c'est à dire si } n \text{ est pair}$$

Chaque couple de vecteurs $(\mathbf{c}_k, \mathbf{s}_k)$ est associé à une fréquence particulière ω_k appelée fréquence de Fourier.

Un des outils central de l'analyse spectrale est le périodogramme. Ce dernier est lié à la décomposition de la variance associée aux différentes fréquences de Fourier. Il définit également des estimateurs de la densité spectrale du processus. Par la suite, on ne s'intéressera qu'à la dimension descriptive du périodogramme. Par définition, les valeurs du périodogramme sont définies pour chaque fréquence de Fourier par :

$$I(\omega_k) = \frac{1}{n} \left(\left(\sum_{i=1}^n \tilde{x}_i \cos(i\omega_k) \right)^2 + \left(\sum_{i=1}^n \tilde{x}_i \sin(i\omega_k) \right)^2 \right)$$

soit, en terme matriciel :

$$I(\omega_k) = \frac{\tilde{\mathbf{x}}^t \mathbf{c}_k \mathbf{c}_k^t \mathbf{D} \tilde{\mathbf{x}}}{2} + \frac{\tilde{\mathbf{x}}^t \mathbf{s}_k \mathbf{s}_k^t \mathbf{D} \tilde{\mathbf{x}}}{2} = \frac{\tilde{\mathbf{x}}^t \mathbf{\Pi}_k \tilde{\mathbf{x}}}{2} = \frac{\tilde{\mathbf{x}}^t \mathbf{\Pi}_k \tilde{\mathbf{x}}}{\text{trace}(\mathbf{\Pi}_k)}$$

Le périodogramme représente donc les variances de la variable projetée sur les sous-espaces engendrés par les couples $(\mathbf{c}_k, \mathbf{s}_k)_{1 \leq k \leq \lfloor \frac{n-1}{2} \rfloor}$. Le calcul du périodogramme nécessite le calcul de

la famille des K formes bilinéaires symétriques constituée par les K projecteurs

$$\mathbf{\Pi}_k = (\mathbf{c}_k \mathbf{c}_k^t + \mathbf{s}_k \mathbf{s}_k^t) \mathbf{D}$$

Il passe donc par l'implémentation de la base \mathbf{B} associée à l'analyse spectrale. On définit ensuite la famille des projecteurs par la partition de l'ensemble des vecteurs de la base \mathbf{B} :

$$E = \{1, 2, \dots, n-1\} = \{1, 2\} \cup \dots \cup \{n-3, n-2\} \cup \{n-1\} = E_1 \cup \dots \cup E_{K-1} \cup E_K \text{ si } n \text{ est pair}$$

$$E = \{1, 2, \dots, n-1\} = \{1, 2\} \cup \dots \cup \{n-2, n-1\} = E_1 \cup \dots \cup E_K \text{ si } n \text{ est impair}$$

Pour définir la base, on utilise la proposition faite par Cornillon (1998) selon qui les vecteurs $(\mathbf{c}_k, \mathbf{s}_k)_{1 \leq k \leq \lfloor \frac{n-1}{2} \rfloor}$ sont les vecteurs propres de l'opérateur de voisinage de Méot pour un graphe

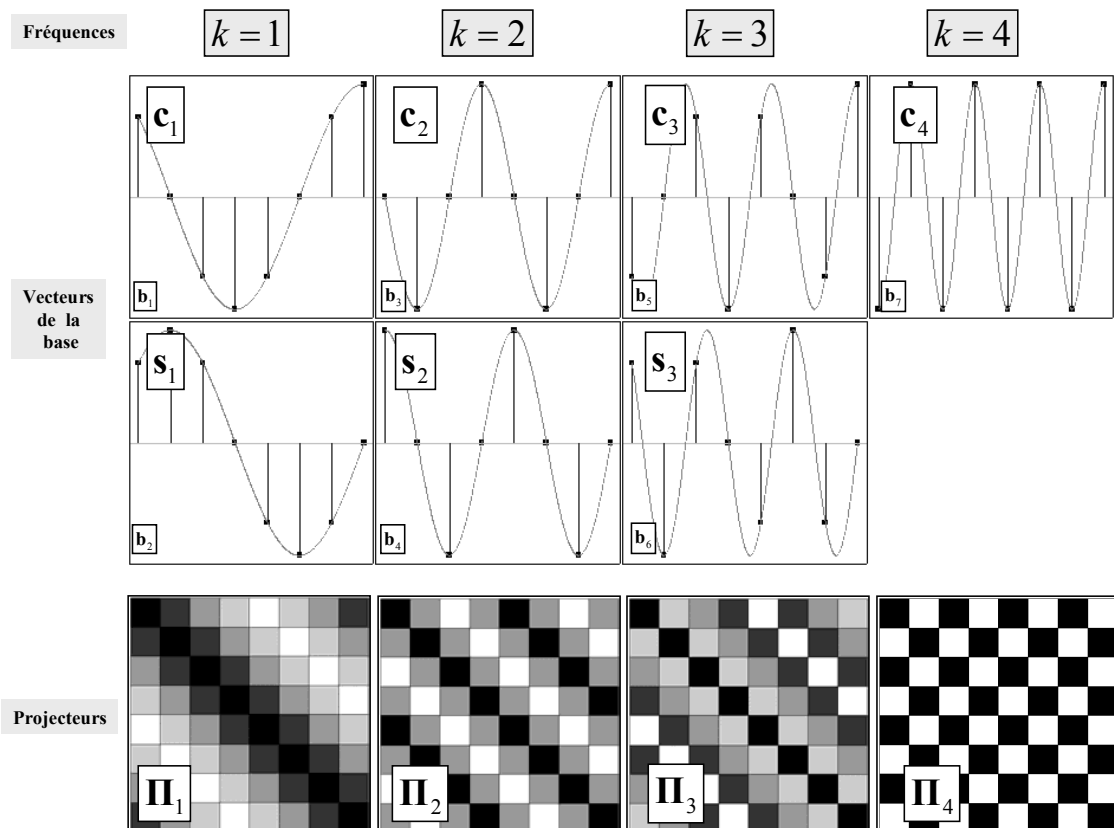
circulaire. La fonction `orthobasis.circ(...)` (Annexe 2.15) donne donc les vecteurs de la base associée à l'analyse spectrale. La fonction `circ2level(...)` (Annexe 2.9) donne la partition permettant d'engendrer la famille des projecteurs associés au calcul du périodogramme à partir de la base \mathbf{B} . La fonction `orthobasis2kfbs(...)` (Annexe 2.5) permet d'engendrer la famille de projecteurs à partir de la base \mathbf{B} et de la partition associée. D'un point de vue numérique, il est bien évident que le coût du calcul par l'implémentation des projecteurs et des formes quadratiques est très important par rapport à la procédure algorithmique FFT (pour une comparaison des deux approches, voir Diggle (1990) p101).

Exemple :

```
n <- 8
orthobas <- orthobasis.circ(n)
dim(orthobas)
[1] 8 7
level <- circ2level(8)
level
[1] A A B B C C D
Levels: A B C D
kfbs <- orthobasis2kfbs(orthobas = orthobas, level = level)
summary(kfbs)
K bilinear symmetric forms
Points : 8      Forms : 4
```

```

All positive forms : TRUE
Call : orthobasis2kfbs(orthobas = orthobas, level = level)
  tr(f) tr(f^2)      sum rank norm
A      2      2  5.548e-17    1    1
B      2      2  1.355e-19    1    1
C      2      2 -1.110e-16    1    1
D      1      1  0.000e+00    1    1
x <- rnorm(8)
x0 <- x - mean(x)
xn <- x0/sqrt((sum(x0**2)/8))
t(xn)%*%xn/8
  [,1]
[1,]  1
mat1 <- as.matrix.kfbs(kfbs, 1)
mat2 <- as.matrix.kfbs(kfbs, 2)
mat3 <- as.matrix.kfbs(kfbs, 3)
mat4 <- as.matrix.kfbs(kfbs, 4)
mat <- list(mat1, mat2, mat3, mat4)
# calcul matriciel du périodogramme
I <- lapply(mat,function(u) t(xn)%*%u%*%xn/sum(diag((u))))
I <- unlist(I)
I
[1] 0.9676 0.8474 0.7156 2.9388
# calcul du périodogramme par la procédure FFT
I <- Mod(fft(xn))^2/length(xn)
I[2:5]
[1] 0.9676 0.8474 0.7156 2.9388
# le périodogramme est lié à la décomposition de la variance
sum(I*c(2,2,2,1))
[1] 8
    
```



5.6. Les bases d'ondelettes à une dimension

L'analyse en ondelettes est une technique d'analyse du signal beaucoup plus récente et largement moins répandue que l'analyse spectrale. Toutefois, ce type d'analyse connaît un regain d'intérêt depuis quelques années et fait l'objet de multiples travaux en mathématiques (Daubechies, 1992), en analyse du signal (Vaidyanathan, 1993) et en statistiques (Percival et Walden (2000) pour l'étude des séries temporelles). La principale motivation est le développement de nouveaux algorithmes de calcul de la transformée en ondelettes, analogue de la transformée de Fourier (Percival, 1993). De plus, cette technique a fait l'objet de nombreuses applications dans des champs scientifiques variés comme la géostatistique (Lark & Webster, 1999, 2001) et l'écologie statistique (Bradshaw & Spies, 1992; Dale & Mah, 1998).

Ce paragraphe a pour objectif de présenter brièvement l'analyse en ondelettes à une dimension en insistant sur l'idée centrale plutôt que sur les détails. Selon Percival (1993), « *the basic idea behind wavelet analysis is to decompose the variance of a series into a number of components, each one of which can be associated with a particular scale at a particular position* ». La transformée en ondelette, à l'instar de la transformée de Fourier, assure la décomposition d'une série en un ensemble de composantes orthogonales. Elle définit les $n-1$ vecteurs de la base \mathbf{B} associée à l'analyse en ondelettes à partir de fonctions mères. Ces fonctions mères portent le nom d'ondelettes (littéralement 'petites ondes') car, contrairement aux fonctions trigonométriques $t \rightarrow \cos(t)$ et $t \rightarrow \sin(t)$ ('grandes ondes'), leurs oscillations sont localisées dans l'espace où le temps et réduites à un intervalle fini (Figure 2.13). Il y a autant de bases \mathbf{B} possibles qu'il existe de fonctions mères $t \rightarrow \psi(t)$ vérifiant :

$$\int_{-\infty}^{\infty} \psi(u) du = 0 \text{ et } \int_{-\infty}^{\infty} \psi^2(u) du = 1$$

Par mesure de simplicité, on se concentrera dans un premier temps exclusivement sur l'ondelette de Haar (Haar, 1910), définie il y a presque 100 ans par :

$$t \rightarrow \psi(t) = \begin{cases} 1, & \text{si } 0 \leq t \leq 1 \\ -1, & \text{si } 1 \leq t \leq 2 \\ 0, & \text{sinon} \end{cases}$$

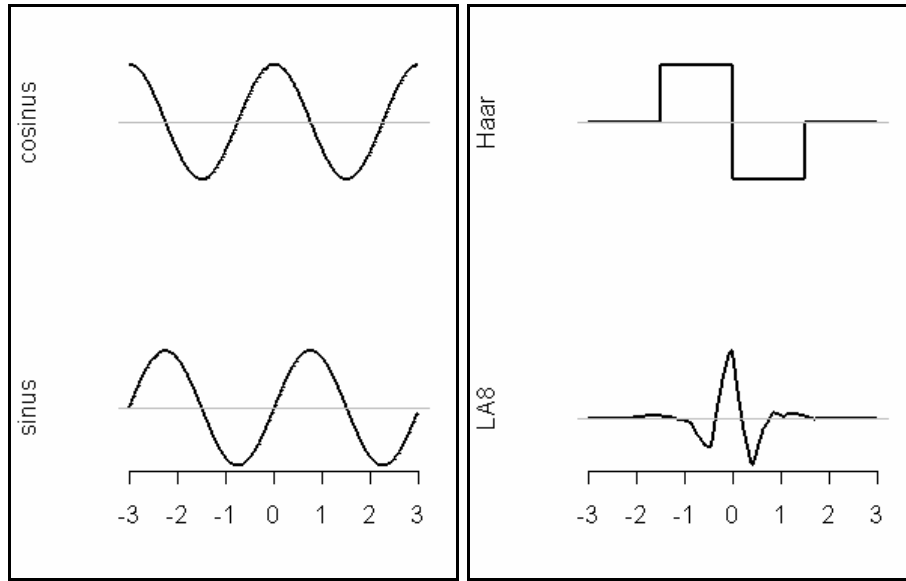


Figure 2.13 : Fonctions trigonométriques (**à gauche**). Ondelettes (**à droite**).

La définition des vecteurs de la base \mathbf{B} passe par la dilatation, la translation et la normalisation des fonctions mères. On retrouve le même principe qu'en analyse spectrale où les fonctions trigonométriques sont dilatées par le paramètre ω_k correspondant à la fréquence de Fourier. Le choix des fréquences implique que les vecteurs obtenus par dilatation des fonctions trigonométriques soient orthogonaux. On retrouve la même contrainte concernant la définition des deux paramètres de dilatation k et de translation u de l'ondelette mère. La solution est donnée, pour une série dont la taille s'exprime comme une puissance de 2, $n = 2^K$ par :

$$\mathbf{g}_i^{(k,u)} = \sqrt{2^{K-k}} \psi\left(\frac{i-u}{k}\right) \text{ avec } \begin{cases} 1 \leq k \leq K \\ 0.5 \leq u \leq n - 2k + 0.5 \end{cases}$$

Un des outils central de l'analyse en ondelettes est le scalogramme. Par définition, les valeurs du scalogramme sont définies par :

$$S(k, u) = \frac{1}{n} \left(\sum_{i=1}^n \tilde{x}_i \mathbf{g}_i^{(k,u)} \right)^2$$

soit, en terme matriciel :

$$S(k, u) = \tilde{\mathbf{x}}^t \mathbf{g}_{(k,u)} \mathbf{g}_{(k,u)}^t \mathbf{D} \tilde{\mathbf{x}} = \tilde{\mathbf{x}}^t \mathbf{\Pi}_{(k,u)} \tilde{\mathbf{x}}$$

Le scalogramme représente donc les variances de la variable projetée sur les sous-espaces engendrés par les vecteurs $\left(\mathbf{g}_{(k,u)} \right)_{\substack{1 \leq k \leq K \\ 0.5 \leq u \leq n - 2k + 0.5}}$. Ce dernier assure la décomposition de la variance aux différentes échelles k pour différentes positions u :

$$\tilde{\mathbf{x}}' \mathbf{D} \tilde{\mathbf{x}} = \sum_{(2)} \tilde{\mathbf{x}}' \mathbf{\Pi}_{(k,u)} \tilde{\mathbf{x}} = \sum_{(2)} S(k,u) = \sum_k \left(\sum_u S(k,u) \right) = \sum_k \tilde{\mathbf{x}}' \mathbf{\Pi}_k \tilde{\mathbf{x}}$$

On peut regrouper les carrés des corrélations par échelles, et représenter le scalogramme uniquement en fonction des échelles k . Si l'on normalise par la trace du projecteur, on obtient une mesure de variance (Percival, 1995), moyenne des carrés de corrélations. Les valeurs de la variance en fonction des échelles k sont analogues aux valeurs du périodogramme pour les fréquences de Fourier ω_k . Le calcul du scalogramme puis des variances aux différentes échelles nécessite d'introduire, d'une part la famille des K formes bilinéaires symétriques constituée par les $n-1$ projecteurs

$$\mathbf{\Pi}_{(k,u)} = \mathbf{g}_{(k,u)} \mathbf{g}_{(k,u)}^t \mathbf{D},$$

d'autre part, la famille des K formes bilinéaires symétriques constituée par les K projecteurs

$$\mathbf{\Pi}_k = \left(\sum_u \mathbf{g}_{(k,u)} \mathbf{g}_{(k,u)}^t \right) \mathbf{D}$$

On commence par l'implémentation de la base \mathbf{B} associée à l'analyse en ondelette. On définit ensuite la famille des projecteurs par la partition de l'ensemble des vecteurs de la base \mathbf{B} :

$$E = \{1, 2, \dots, n-1\} = \{1\} \cup \dots \cup \{n-2\} \cup \{n-1\} = E_1 \cup \dots \cup E_{n-1} \text{ où}$$

$$E = \{1, 2, \dots, n-1\} = \{1, \dots, 2^{K-1}\} \cup \dots \cup \{2^{K-K}\} = E_1 \cup \dots \cup E_K$$

La fonction `orthobasis.wavelet(...)` (Annexe 2.15) donne les vecteurs de la base d'ondelette \mathbf{B} pour l'ensemble des ondelettes mères disponibles dans le package `waveslim`. La fonction `wavelet2level(...)` (Annexe 2.9) donne la partition permettant d'engendrer la famille des projecteurs $\mathbf{\Pi}_k$ associés au calcul du scalogramme à partir de la base \mathbf{B} . La fonction `orthobasis2kfbs(...)` (Annexe 2.5) permet d'engendrer la famille de projecteurs à partir de la base \mathbf{B} et de la partition. D'un point de vue numérique, il est bien évident que le coût du calcul par l'implémentation des projecteurs et des formes quadratiques est très important par rapport à la procédure algorithmique DWT ('discrete wavelet transform'). Cette procédure exploite la rareté des coefficients non nuls dans les bases d'ondelette afin de réduire le coût du calcul, en introduisant successivement les filtres passe haut et passe bas combinés aux ondelettes mères (pour une description détaillée de la procédure, voir le chapitre 4 de l'ouvrage de Percival et Walden (2000)).

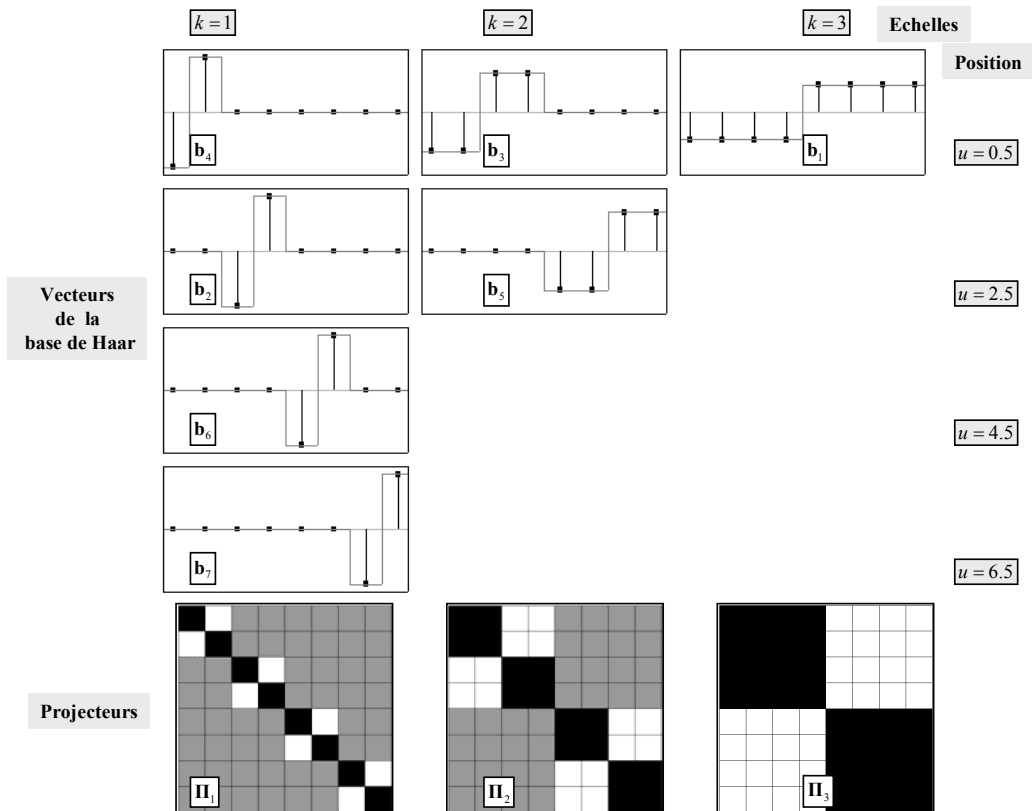
Exemples :

```
n <- 8
# base de Haar
orthobas <- orthobasis.wavelet(n, "haar")
dim(orthobas)
[1] 8 7
```

```

level <- wavelet2level(8)
level
[1] B3 B2 B2 B1 B1 B1 B1
Levels: B3 B2 B1
kfbs <- orthobasis2kfbs(orthobas = orthobas, level = level)
summary(kfbs)
K bilinear symetric forms
Points : 8      Forms : 3
All positive forms : TRUE
Call : orthobasis2kfbs(orthobas = orthobas, level = level)
      tr(f) tr(f^2) sum rank norm
B3      1      1  0    1    1
B2      2      2  0    1    1
B1      4      4  0    1    1
x <- rnorm(8)
x0 <- x - mean(x)
xn <- x0/sqrt((sum(x0**2)/8))
t(xn)%*%xn/8
      [,1]
[1,]      1
mat1 <- as.matrix.kfbs(kfbs, 1)
mat2 <- as.matrix.kfbs(kfbs, 2)
mat3 <- as.matrix.kfbs(kfbs, 3)
mat <- list(mat1, mat2, mat3)
# calcul matriciel des variances à chaque échelle k
S <- lapply(mat, function(u) t(xn)%*%u)%*%xn/sum(diag((u))))
S <- unlist(S)
S
[1] 3.5189 0.4725 0.8840
# calcul des variances par la procédure DWT
S <- dwt(xn, wf = "haar", n.levels = 3)
S <- unlist(lapply(S[-4], function(x) mean(x**2)))
      d1      d2      d3
0.8840 0.4725 3.5189
sum(S*c(1,2,4)) # les valeurs du scalogramme assure
                  # une décomposition de la variance
[1] 8

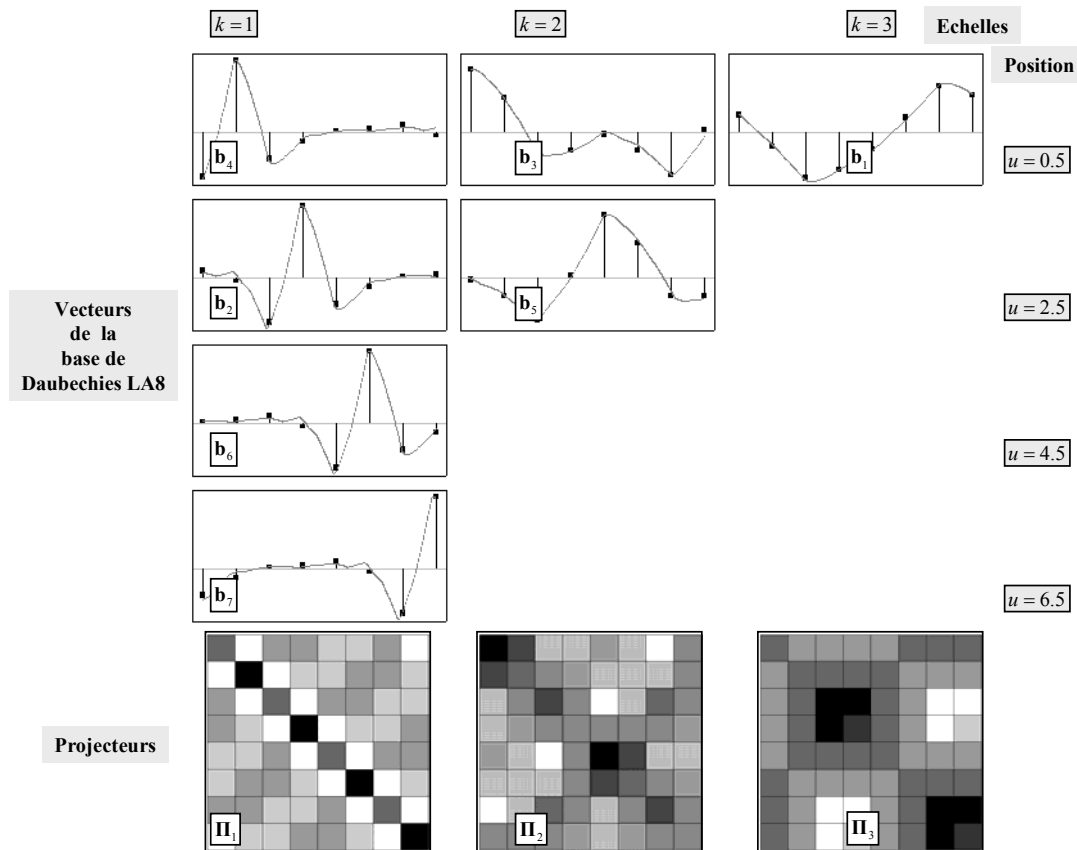
```



```

orthobas <- orthobasis.wavelet(n, "la8")
dim(orthobas)
[1] 8 7
kfbs <- orthobasis2kfbs(orthobas = orthobas, level = level)
summary(kfbs)
K bilinear symmetric forms
Points : 8    Forms : 3
All positive forms : TRUE
Call : orthobasis2kfbs(orthobas = orthobas, level = level)
  tr(f)  tr(f^2)      sum rank norm
B3     1      1  5.551e-17    1    1
B2     2      2 -1.874e-16    1    1
B1     4      4  1.128e-16    1    1

```



6. NORMALISATION DES FORMES BILINÉAIRES

6.1. Introduction

L'objectif de départ était de trouver une manière de décrire la structure d'une variable puis de faire une typologie de ces structures pour plusieurs variables. On a vu qu'il existait de multiples approches pour décrire la structure d'une variable mesurée le long d'un transect, dont la plupart définissent une famille de K formes bilinéaires symétriques. Pour une variable centrée et normée \tilde{x} , il suffit de prendre les formes quadratiques correspondantes normalisées par v_k :

$$\frac{(\tilde{\mathbf{x}}^t \mathbf{A}_k \tilde{\mathbf{x}})_{1 \leq k \leq K}}{v_k}.$$

On retrouve les fonctions structurales classiques (Legendre) : le **variogramme** et le **corrélogramme** pour les formes de Geary/Lebart, Moran/Smouse ou les formes de Hill, le **périodogramme** pour les projecteurs de l'analyse spectrale, **les variances d'ondelette** pour les projecteurs de l'analyse en ondelette, **les variances hiérarchiques** pour les projecteurs de Noy-Meir/Greig-Smith, et l'**orthogram** pour les projecteurs associés à chaque vecteur d'une base orthonormée quelconque, généralisation du **scalogramme** (voir Annexe pour une illustration du calcul de chaque fonction structurale).

Se pose alors la question de la normalisation des formes. En effet, il suffit de multiplier \mathbf{A}_k par une constante pour changer la forme du vecteur $(\tilde{\mathbf{x}}^t \mathbf{A}_k \tilde{\mathbf{x}})_{1 \leq k \leq K}$. La question n'est pas simple, car les normes envisageables sont multiples et dépendent des formes considérées.

6.2. Définitions

6.2.1. Normalisation d'un projecteur

Qu'il s'agisse du périodogramme, des variances d'ondelette, ou des variances hiérarchiques, on retrouve systématiquement la normalisation par $v_k = \text{Trace}(\mathbf{\Pi}_k)$:

$$\frac{\tilde{\mathbf{x}}^t \mathbf{A}_k \tilde{\mathbf{x}}}{\text{Trace}(\mathbf{A}_k^2)} = \frac{\tilde{\mathbf{x}}^t \mathbf{A}_k \tilde{\mathbf{x}}}{\text{Trace}(\mathbf{A}_k)} = \frac{\tilde{\mathbf{x}}^t \mathbf{\Pi}_k \tilde{\mathbf{x}}}{\text{Trace}(\mathbf{\Pi}_k)} = \frac{\tilde{\mathbf{x}}^t \mathbf{\Pi}_k \tilde{\mathbf{x}}}{\dim(EV)}$$

La norme correspond à la dimension du sous-espace de projection. Cette normalisation s'impose car lorsque la forme est positive et $\text{sum}(\mathbf{A}_k) = 0$ elle donne une statistique de la variance (Hill, 1973).

6.2.2. Normalisation des formes de Geary

Pour les formes de Geary, la normalisation naturelle est celle de l'indice de Geary que l'on retrouve pour le calcul des valeurs du semi variogramme :

$$\frac{\tilde{\mathbf{x}}^t \mathbf{A}_k \tilde{\mathbf{x}}}{\text{Trace}(\mathbf{A}_k)} = \frac{\tilde{\mathbf{x}}^t (\mathbf{N}_k - \mathbf{M}_k) \tilde{\mathbf{x}}}{\text{Trace}(\mathbf{N}_k - \mathbf{M}_k)} = \frac{\tilde{\mathbf{x}}^t (\mathbf{N}_k - \mathbf{M}_k) \tilde{\mathbf{x}}}{2m_k}$$

6.2.3. Normalisation des formes de Moran

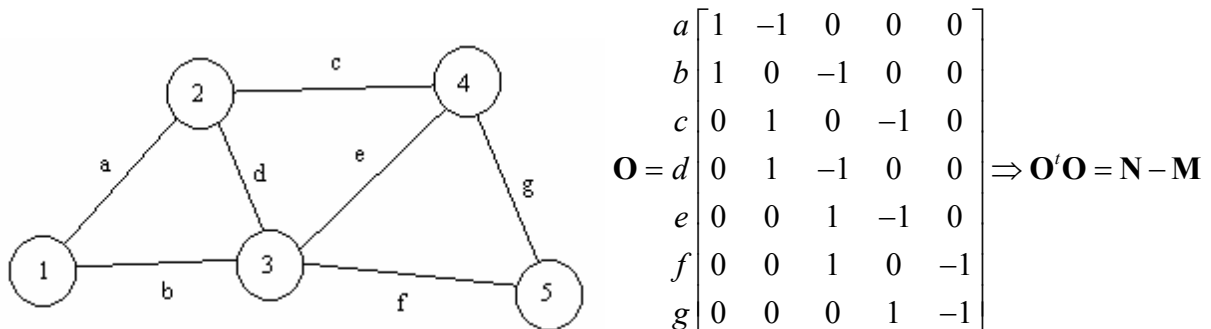
Pour les formes de Moran, la normalisation naturelle est celle de l'indice de Moran :

$$\frac{\tilde{\mathbf{x}}^t \mathbf{A}_k \tilde{\mathbf{x}}}{\text{Trace}(\mathbf{A}_k^2)} = \frac{\tilde{\mathbf{x}}^t \mathbf{M}_k \tilde{\mathbf{x}}}{\text{Trace}(\mathbf{M}_k^2)} = \frac{\tilde{\mathbf{x}}^t (\mathbf{M}_k) \tilde{\mathbf{x}}}{2m_k},$$

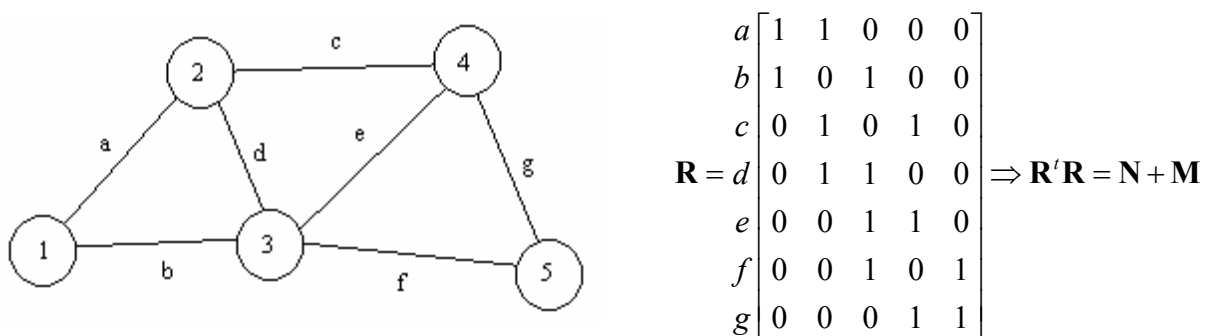
Smouse et Peakall utilisent les formes de Moran mais introduisent un nouveau coefficient dont la normalisation est originale. Ce coefficient est défini par

$$\frac{\tilde{\mathbf{x}}^t \mathbf{M}_k \tilde{\mathbf{x}}}{\tilde{\mathbf{x}}^t \mathbf{N}_k \tilde{\mathbf{x}}} = \frac{\text{Trace}(\tilde{\mathbf{x}} \tilde{\mathbf{x}}^t \mathbf{M}_k)}{\text{Trace}(\tilde{\mathbf{x}} \tilde{\mathbf{x}}^t \mathbf{N}_k)},$$

Ce rapport est étonnant car il fait intervenir les deux matrices fondamentales du graphe de voisinage. La formule est extrêmement simple et semble recouvrir une structure forte. Le résultat est compris entre -1 et +1 parce que les matrices $\mathbf{M}_k + \mathbf{N}_k$ et $\mathbf{N}_k - \mathbf{M}_k$ sont positives. Pour le prouver, ramenons nous aux matrices qui croisent les arêtes du graphe et les sommets du graphe. La première, matrice d'incidence aux arcs (Berge, 1967) donne sur le graphe suivant :



La seconde, matrice d'incidence aux arêtes (Berge, 1967) donne sur le même exemple :



Les deux sont évidemment associées fortement. Soit \mathbf{A} la matrice qui couple les arêtes (en lignes) et les sommets (en colonnes) dont les éléments valent 1 ssi l'arête i a pour sommet le point j (l'autre sommet étant k avec $k > i$). Soit \mathbf{B} la matrice qui couple les arêtes (en lignes) et les sommets (en colonnes) dont les éléments valent 1 ssi l'arête i a pour sommet le point j , l'autre sommet étant k avec $k < i$ alors :

$$\mathbf{O} = \mathbf{A} - \mathbf{B} \text{ et } \mathbf{R} = \mathbf{A} + \mathbf{B}$$

d'où

$$\mathbf{O}'\mathbf{O} = \mathbf{N} - \mathbf{M} = (\mathbf{A}' - \mathbf{B}')(\mathbf{A} - \mathbf{B})$$

$$\mathbf{R}'\mathbf{R} = \mathbf{N} + \mathbf{M} = (\mathbf{A}' + \mathbf{B}')(\mathbf{A} + \mathbf{B})$$

et

$$\mathbf{N} = \mathbf{A}'\mathbf{A} + \mathbf{B}'\mathbf{B}$$

$$\mathbf{M} = \mathbf{A}'\mathbf{B} + \mathbf{B}'\mathbf{A}$$

On observe alors que si une matrice symétrique s'écrit $\mathbf{H} = \mathbf{A}'\mathbf{B} + \mathbf{B}'\mathbf{A}$ on peut lui associer la matrice symétrique $\tilde{\mathbf{H}} = \mathbf{A}'\mathbf{A} + \mathbf{B}'\mathbf{B}$ également symétrique ((Harville, 1997)) et que :

$$-1 \leq \frac{\tilde{\mathbf{x}}'\mathbf{H}\tilde{\mathbf{x}}}{\tilde{\mathbf{x}}'\tilde{\mathbf{H}}\tilde{\mathbf{x}}} \leq 1$$

car $\tilde{\mathbf{H}}$, $\mathbf{H} + \tilde{\mathbf{H}}$ et $\tilde{\mathbf{H}} - \mathbf{H}$ sont positives.

Il suffit alors d'observer qu'une matrice symétrique a toujours une décomposition du type $\mathbf{H} = \mathbf{A}'\mathbf{B} + \mathbf{B}'\mathbf{A}$ par décomposition en valeurs singulières. \mathbf{H} étant symétrique on peut toujours l'écrire $\mathbf{H} = \mathbf{S} + \mathbf{S}'$ où \mathbf{S} est une matrice triangulaire supérieure :

$$\begin{cases} j > i \Rightarrow \mathbf{S}_{ij} = \mathbf{H}_{ij} \\ j = i \Rightarrow \mathbf{S}_{ij} = \mathbf{H}_{ij}/2 \\ j < i \Rightarrow \mathbf{S}_{ij} = 0 \end{cases}$$

alors :

$$\mathbf{H} = \mathbf{S} + \mathbf{S}' = \mathbf{UDV}' + \mathbf{VDU}' \Rightarrow \tilde{\mathbf{H}} = \mathbf{UDU}' + \mathbf{VDV}'$$

Le cas particulier pour la forme de Moran associée à un graphe de voisinage est le coefficient de corrélation de Smouse et Peakall (1999). Le calcul est généralisable à l'ensemble des formes bilinéaires symétriques mais a-t-il une signification ?

Par la suite, nous conservons trois méthodes pour calculer les suites $(\tilde{\mathbf{x}}^t \mathbf{A}_k \tilde{\mathbf{x}} / v_k)_{1 \leq k \leq K}$. La première utilise la norme euclidienne (**méthode VEU**) :

$$Q_{\mathbf{A}_k}(\tilde{\mathbf{x}}) = \frac{\tilde{\mathbf{x}}^t \mathbf{A}_k \tilde{\mathbf{x}}}{\|\mathbf{A}_k\|} = \frac{\tilde{\mathbf{x}}^t \mathbf{A}_k \tilde{\mathbf{x}}}{\sqrt{\text{Trace}(\mathbf{A}_k^2)}}$$

La seconde (**méthode VSP**) utilise la norme introduite par Smouse et Peakall (1999) :

$$S_{\mathbf{A}_k}(\tilde{\mathbf{x}}) = \frac{\tilde{\mathbf{x}}^t \mathbf{A}_k \tilde{\mathbf{x}}}{\tilde{\mathbf{x}}^t \tilde{\mathbf{A}}_k \tilde{\mathbf{x}}}$$

La troisième (**méthode VVP**) utilise la norme spectrale qui est majorée par la norme euclidienne (dite aussi de Frobenius ou de Schur ou de Hilbert-Schmidt (Chatelin, 1988)) :

$$R_{\mathbf{A}_k}(\tilde{\mathbf{x}}) = \frac{\tilde{\mathbf{x}}^t \mathbf{A}_k \tilde{\mathbf{x}}}{\sqrt{\lambda_1(\mathbf{A}_k^2)}} = \frac{\tilde{\mathbf{x}}^t \mathbf{A}_k \tilde{\mathbf{x}}}{\sqrt{\lambda_1(\mathbf{A}_k^2)}}$$

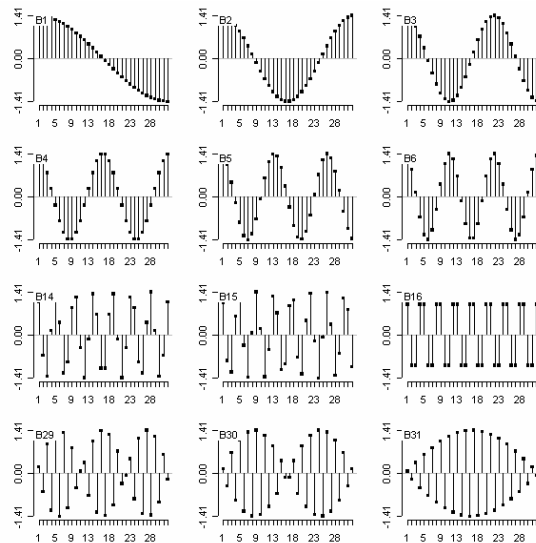
La normalisation spectrale a la propriété de comparer la structure observée à un ensemble de figures de référence (« *templates* ») caractéristiques de chaque famille de formes.

La fonction `val.kfbs(...)` (Annexe 2.10) calcule les suites $(\tilde{\mathbf{x}}^t \mathbf{A}_k \tilde{\mathbf{x}} / v_k)_{1 \leq k \leq K}$ pour les trois types de normalisation considérées.

6.2.4. Comparaison des normes

On décide d'étudier et de comparer le comportement de chaque norme à partir d'un échantillon de diverses fonctions de structures.

```
orthobas <- orthobasis.line(32)
tab <- orthobas[,c(1:6,14:16,29:31)]
dim(tab)
[1] 32 12
par(mar = c(3,3,1,1))
dotchart.line(tab)
```



On calcule les valeurs des formes quadratiques $Q_{\mathbf{A}_k}(\tilde{\mathbf{x}})$, $S_{\mathbf{A}_k}(\tilde{\mathbf{x}})$ et $R_{\mathbf{A}_k}(\tilde{\mathbf{x}})$ associées aux variables présentées ci-dessus pour :

- les formes de Geary

```
knb <- neig2knb(neig(n.line = 32))
geary.kfbs <- knb2kfbs(knb, "Geary")
geary.val <- val.kfbs(tab, geary.kfbs)
names(geary.val)
[1] "veu" "vvp" "vsp"
```

- les formes de Moran

```
knb <- neig2knb(neig(n.line = 32))
moran.kfbs <- knb2kfbs(knb, "Moran")
moran.val <- val.kfbs(tab, moran.kfbs)
```

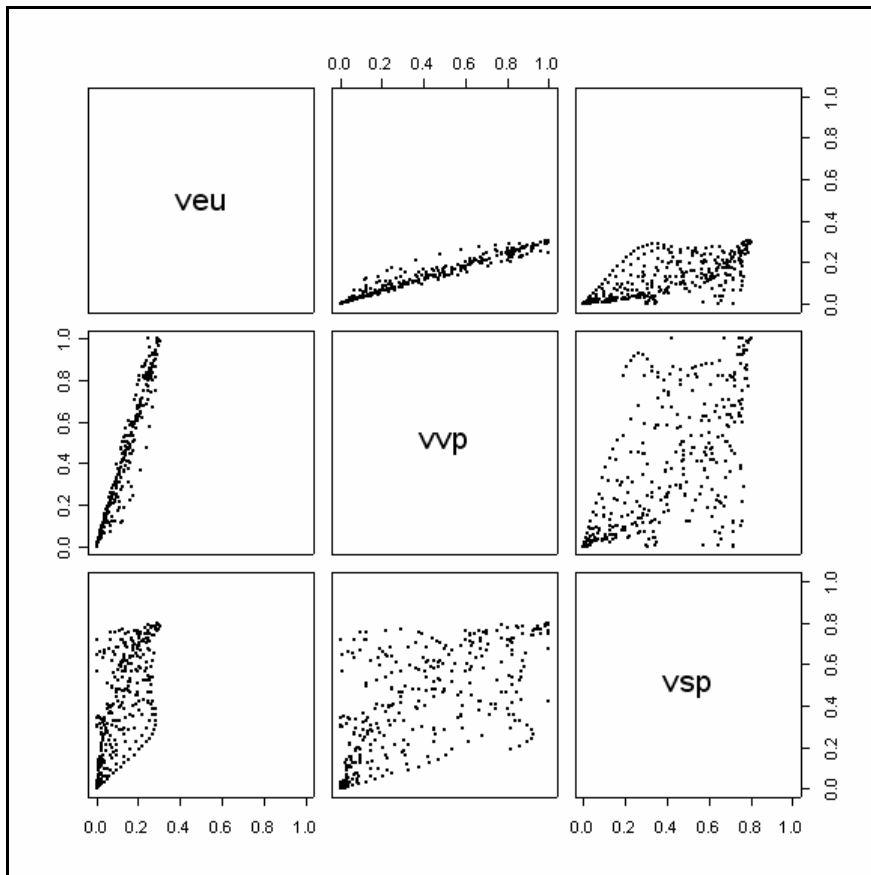
- et les formes de Noy-Meir

```
noy.kfbs <- msbs.kfbs(32, c(1,2,4,8,16,32))
noy.val <- val.kfbs(tab, noy.kfbs)
```

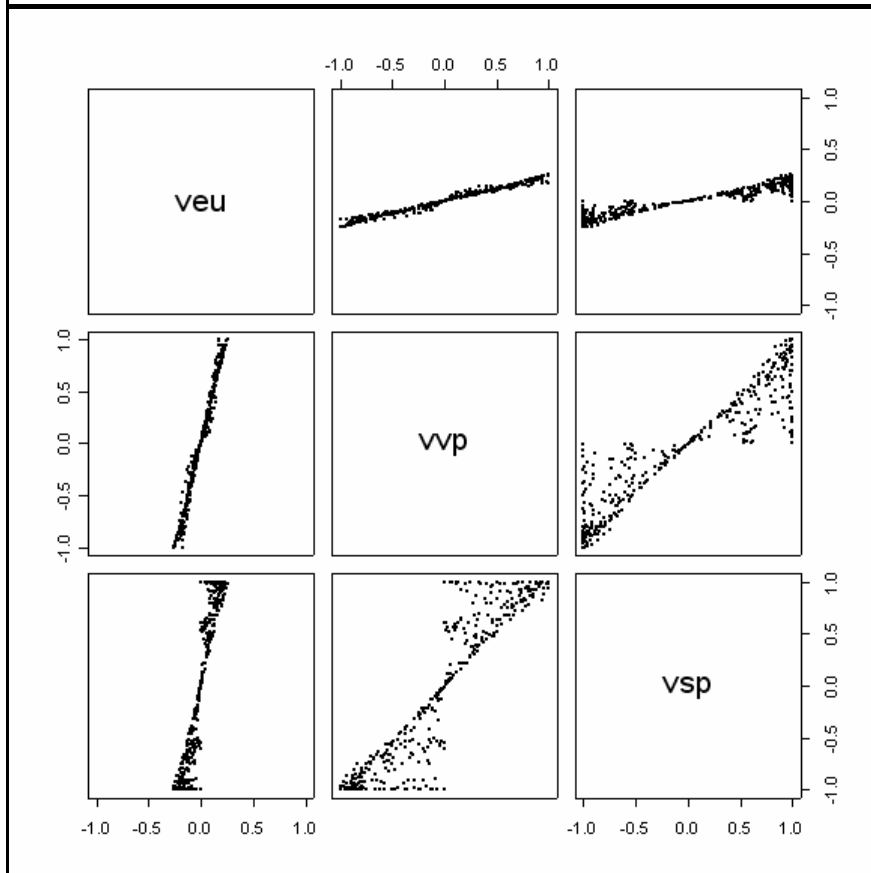
L'objectif de ce calcul est de comparer sur un échantillon de plusieurs variables, le comportement de chacune des trois normes, pour trois familles de formes. On trace alors pour chaque famille de formes, les trois nuages de points dont les coordonnées respectives sont

$$\left(Q_{A_k}(\tilde{\mathbf{x}}_i), S_{A_k}(\tilde{\mathbf{x}}_i)\right), \left(Q_{A_k}(\tilde{\mathbf{x}}_i), R_{A_k}(\tilde{\mathbf{x}}_i)\right) \text{ et } \left(R_{A_k}(\tilde{\mathbf{x}}_i), S_{A_k}(\tilde{\mathbf{x}}_i)\right).$$

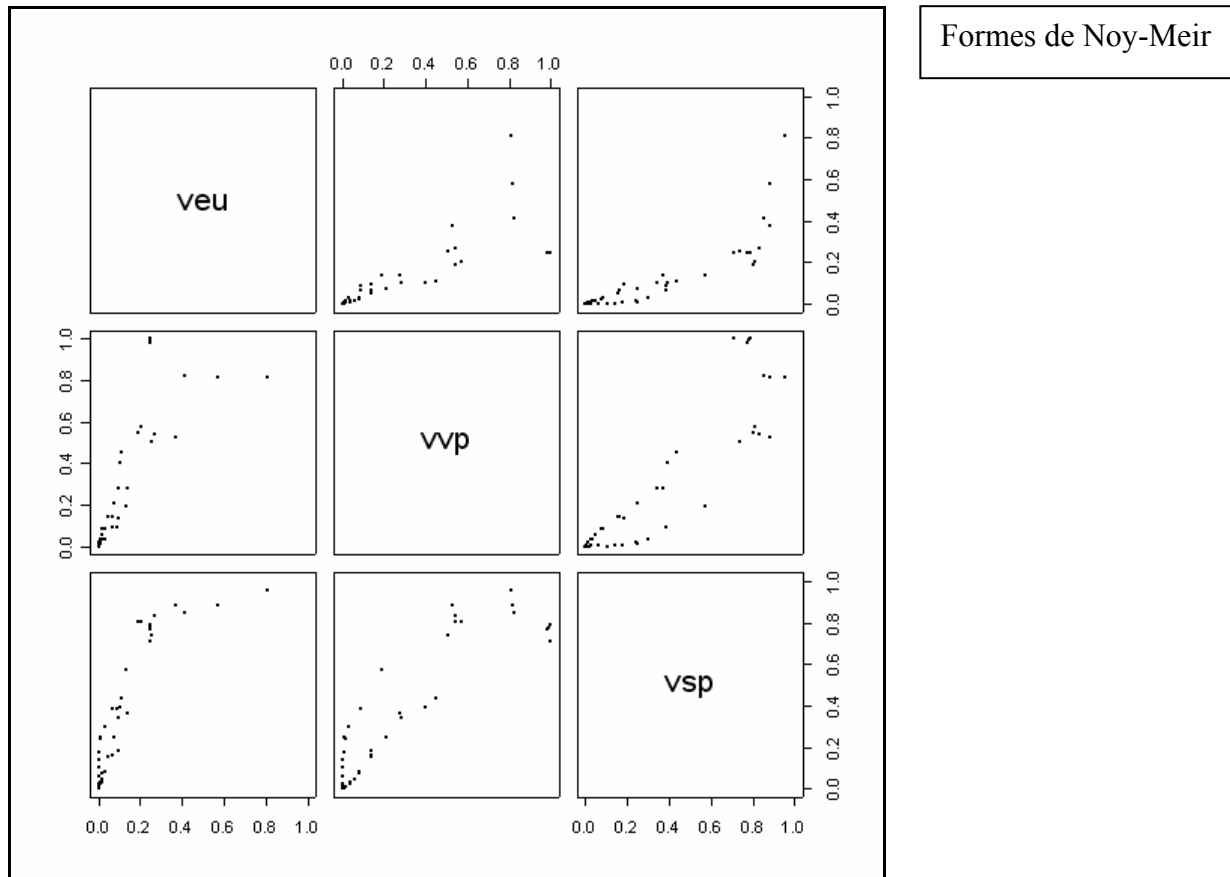
```
pairs.val.kfbs <- function(val) {
  veu <- matrix(unlist(val$veu), ncol=1)
  vvp <- matrix(unlist(val$vvp), ncol=1)
  vsp <- matrix(unlist(val$vsp), ncol=1)
  res <- cbind.data.frame(veu, vvp, vsp)
  pairs(res, xlim = c(-1,1), ylim = c(-1,1), pch = 20)
}
x11()
pairs.val.kfbs(geary.val)
pairs.val.kfbs(moran.val)
pairs.val.kfbs(noy.val)
```

Formes de Geary



Formes de Moran



L'écart entre la norme spectrale et la norme euclidienne est variable. La norme spectrale est toujours plus grande que la norme euclidienne mais dans des proportions très variables. La normalisation de Smouse et Peakall est très particulière aux opérateurs de Moran pour lesquels elle donne pratiquement la norme spectrale. De manière générale, les liens dépendent fortement des familles de formes. La normalisation des formes bilinéaires symétriques n'est pas un problème simple. La normalisation euclidienne doit cependant être rejetée. En effet, dans une famille donnée, l'écart entre norme euclidienne et norme spectrale varie considérablement. Cela veut dire qu'on ne peut comparer deux valeurs d'un variogramme ou d'un corrélogramme pour une même variable, ces derniers étant définis avec la normalisation euclidienne. Pour Moran et Geary, ce n'est cependant pas un problème sensible. La relation entre la norme euclidienne et la norme spectrale est très stable pour ces deux métriques :

```
summary(geary.kfbs)
K bilinear symmetric forms
Points : 32      Forms : 31
All positive forms : TRUE
Call : knb2kfbs(knb = knb, method = "Geary")
      tr(f) tr(f^2) sum rank norm
G1      62    184    0   31 3.990
G2      60    176    0   30 3.962
G3      58    168    0   29 3.919
G4      56    160    0   28 3.848
G5      54    152    0   27 3.802
```

```
G6      52      144  0  26 3.732
G7      50      136  0  25 3.618
G8      48      128  0  24 3.414
G9      46      120  0  23 3.414
G10     44      112  0  22 3.414
G11     42      104  0  21 3.000
...
G31     2        4   0   1 2.000
```

```
summary(moran.kfbs)
K bilinear symmetric forms
Points : 32      Forms : 31
All positive forms : FALSE
Call : knb2kfbs(knb = knb, method = "Moran")
      tr(f) tr(f^2) sum rank  norm
M1      0      62 62  32 1.991
M2      0      60 60  32 1.966
M3      0      58 58  30 1.932
M4      0      56 56  32 1.879
M5      0      54 54  30 1.848
M6      0      52 52  28 1.802
M7      0      50 50  28 1.732
M8      0      48 48  32 1.618
M9      0      46 46  28 1.618
M10     0      44 44  24 1.618
M11     0      42 42  22 1.414
...
M31     0        2  2   2 1.000
```

Par contre, pour les familles de projecteurs, la norme spectrale vaut toujours 1 et la norme euclidienne est la dimension du sous-espace de projection qui varie fortement :

```
summary(noy.kfbs)
K bilinear symmetric forms
Points : 32      Forms : 5
All positive forms : TRUE
Call : msbs.kfbs(n = 32, tbloc = c(1, 2, 4, 8, 16, 32))
      tr(f) tr(f^2) sum rank norm
1_2      16      16 -4.089e-16  16   1
2_4       8       8  1.686e-16   8   1
4_8       4       4 -1.180e-16   4   1
8_16      2       2  6.939e-17   2   1
16_32     1       1  7.980e-17   1   1
```

En conclusion, on peut donc dire que la normalisation spectrale a la fonction claire de comparer chaque variable observée à un ensemble de figures de référence : par conséquent, on la retient définitivement lorsque l'objectif est de réaliser une typologie de structures.

6.3. Typologie de structures

6.3.1. Définitions

Soit le tableau des valeurs des formes pour les variables. Son terme général est

$$y_{ik} = \frac{\tilde{\mathbf{x}}_i^t \mathbf{A}_k \tilde{\mathbf{x}}_i}{v_k}$$

où $\tilde{\mathbf{x}}_i$ est la i ème variable centrée et normalisée au sens de la métrique

canonique \mathbf{D} et \mathbf{A}_k/v_k est où la k ème forme normée au sens de la norme spectrale

$(v_k = \sqrt{\lambda_1(\mathbf{A}_k^2)})$. La variation de $\frac{\tilde{\mathbf{x}}_i^t \mathbf{A}_k \tilde{\mathbf{x}}_i}{v_k}$ en fonction de k a un sens. Elle définit la structure de la variable. La variation de $\frac{\tilde{\mathbf{x}}_i^t \mathbf{A}_k \tilde{\mathbf{x}}_i}{v_k}$ en fonction de i a un sens. Elle définit l'importance des variables à l'échelle choisie. Le tableau traité est homogène. On peut en faire une ACP non centrée.

6.3.2. Illustrations

On considère le jeu de données simulé proposé par Ver-hoef (1989) (Annexe 1.8). L'objectif est clairement affiché : il s'agit de retrouver les différentes échelles de structures et regrouper les variables en fonction de leur structure interne.

```

dotchart.line(gg)
gg.kfbs <- ttlv.kfbs(30, c(1,2,3,5,7,10,15), "Geary")
gg.val <- val.kfbs(gg, gg.kfbs)
gg.pca <- dudi.pca(gg.val$vvvp, center = FALSE)
Select the number of axes: 2
scatter(gg.pca, posi = "bottom")
    
```

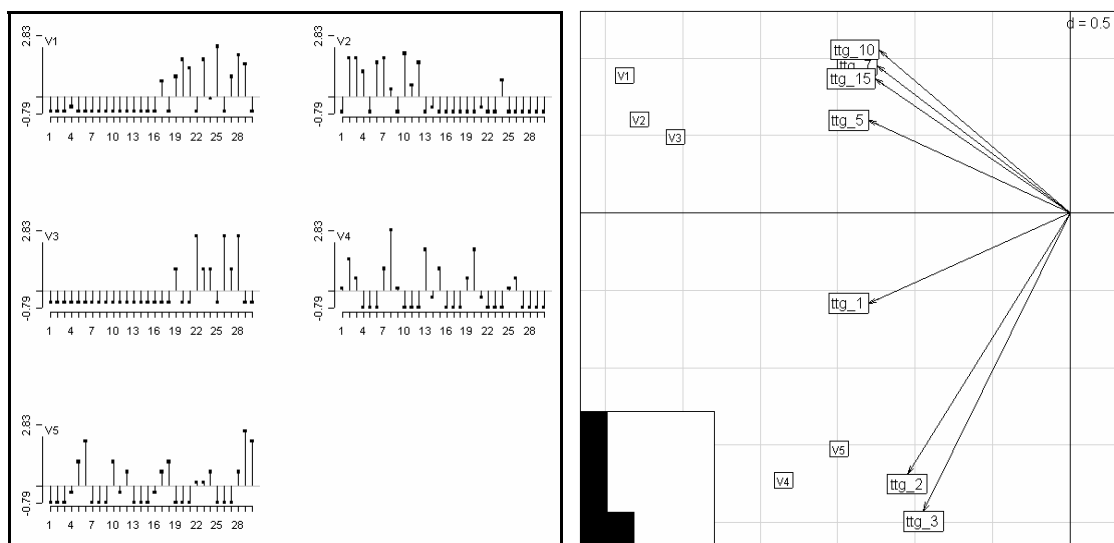


Figure 2.14 : A gauche, représentation graphique des 5 variables simulées. A droite, biplot de l'ACP non centrée du tableau des valeurs des formes quadratiques calculées pour la métrique de Hill.

La typologie est très explicite. Les variables 1, 2 et 3 fonctionnent ensemble : elles sont toutes caractérisées par une structure à grande échelle. Les variables 4 et 5 sont, elles, caractérisées par une structure d'échelle intermédiaire. L'exemple est trivial et ne sert qu'à titre d'illustration. On a reproduit cette expérience sur des données réelles. Les résultats font l'objet du paragraphe suivant.

7. APPLICATIONS AUX DONNÉES D'ALTIMÉTRIE LASER

L'analyse des données d'altimétrie laser a donné lieu à un article publié dans la revue *Remote Sensing of Environment* (Ollier et al., 2003) (Annexe 3.3). On se référera à cet article pour l'ensemble des résultats relatifs à ce problème.

8. DISCUSSION ET PERSPECTIVES

La traduction en termes algébriques de la plupart des méthodes d'analyses multiéchelles a permis de mettre en place les éléments théoriques pour leur intégration en analyse multivariée, en particulier d'aborder les problèmes posés par la typologie de structures. Elle constitue par ailleurs un moyen remarquable d'évaluation des propriétés statistiques des différentes solutions proposées sur des bases mathématiques. Elle illustre la nécessité de l'abstraction mathématique, qui, en assurant une référence théorique commune permet d'ordonner un ensemble de pratiques. Parallèlement à l'article de Dale et al. (2002), qui proposent une ordination de ces pratiques sur des bases conceptuelles (figure 16, p 575), on est alors en mesure de fournir une ordination sur des bases réellement mathématiques (Figure 2.9). De plus, on est désormais capable d'aborder le problème de l'ordination multiéchelle avec de solides arguments, les deux éléments essentiels du problème (les méthodes d'analyse multivariée d'une part et les méthodes d'analyse multiéchelle d'autre part) étant parfaitement maîtrisées tant du point de vue mathématique (triplet statistique $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$ d'une part, famille de formes bilinéaires d'autre part $(\mathbf{A}_k)_{1 \leq k \leq K}$) qu'informatique (classe d'objets 'dudi' d'une part, classe d'objets 'kfb' d'autre part).

Reste à confronter ces outils à des données réelles, afin d'évaluer leur pertinence, indépendamment de leurs propriétés mathématiques. Jusqu'à présent, on s'est limité aux données d'altimétrie laser, mais vu la diversité des plans d'expérience et la diversité des structures biologiques en jeu, on est bien loin d'avoir fait le tour du problème. En particulier, il apparaît bien difficile de privilégier une approche plutôt qu'une autre : le choix d'une famille de formes dépendra implicitement des données et des objectifs recherchés. On a vu par exemple qu'il existait en théorie une infinité de bases orthonormées alors qu'en pratique seules quelques unes sont couramment utilisées. Il n'existe pas de critère absolu permettant de choisir *a priori* une base plutôt qu'une autre : si l'on recherche des périodicités et que les données sont stationnaires, on utilisera plutôt la base de Fourier ; si l'on recherche des changements dans l'organisation de la variance, on utilisera plutôt les bases d'ondelettes ; si l'on cherche à compresser les données, on utilisera une base pour laquelle l'accumulation de la variance sur chaque vecteur propre est optimale...

9. BIBLIOGRAPHIE

Afriat, S.N. (1957) Orthogonal and oblique projectors and the characteristics of pairs of vector spaces. *Proceedings of the Cambridge Philosophical Society, Mathematical and Physical Sciences*, 53, 800-816.

Banet, T.A. & Lebart, L. (1984). Local and Partial Principal Component Analysis (PCA) and Correspondence Analysis (CA). In *COMPSTAT 84* (ed I.A.f.S. Computing.), pp. 113-123. Physica-Verlag, Vienna.

Barbault, R. (1992) *Ecologie des peuplements. Structure, dynamique et évolution* Masson, Paris.

Berge, C. (1967) *Théorie des graphes et ses applications* Dunod, paris.

Bohte, Z., Cepar, D., Kosmelj, K., & Ljubljana, Y.U. (1980) Clustering of time series. *COMPSTAT*.

Borcard, D., Legendre, P., Avois-Jacquet, C., & Tuomisto, H. (2004) Dissecting the spatial structure of ecological data at multiple scales. *Ecology*, 85, 1826-1832.

Boyé, M., Cabaussel, G., & Perrot, Y. (1979). Climatologie. In *Atlas des départements français d'Outre Mer, 4: la Guyane Française* (ed C.a. ORSTOM), pp. 1-4.

Bradshaw, G.A. & Spies, T.A. (1992) Characterizing canopy gap structure in forests using wavelet analysis. *Journal of Ecology*, 80, 205-215.

Brillinger, D.R., Guttorp, P.M., & Schoenberg, F.P. (2002). Point processes, temporal. In *Encyclopedia of Environmetrics* (eds A.H. El-Shaarawi & W.W. Piegorsch), Vol. 3, pp. 1577–1581. John Wiley & Sons, Ltd, Chichester.

Brown, J.H. & Maurer, B.A. (1989) Macroecology: the division of food and space among species on continents. *Science*, 243, 1145-1150.

Chatelin, F. (1988) *Valeurs propres de matrices* Masson, Paris.

Chessel, D. (1992) *Echanges interdisciplinaires en analyse de données écologiques. Mémoire d'habilitation.* Université Lyon 1.

Cornillon, P.-A. (1998) *Prise en compte de proximités en analyse factorielle et comparative.* Thèse, Ecole Nationale Supérieure Agronomique, Montpellier.

Couteron, P. (2001) Using spectral analysis to confront distributions of individual species with an overall periodic pattern in semi-arid vegetation. *Plant Ecology*, 156, 229-243.

Couteron, P. (2002) Quantifying change in patterned semi-arid vegetation by Fourier analysis of digitised aerial photographs. *International Journal of Remote Sensing*, 23, 3407-3425.

Couteron, P., Mahamane, A., & Ouedraogo, P. (1996) Analyse de la structure de peuplements ligneux dans un "fourré tigré" au nord Yatenga (Burkina Faso). Etat actuel et conséquences évolutives. *Annales des Sciences Forestières*, 53, 867-884.

Couteron, P. & Ollier, S. (sous presse) A generalized variogram-based framework for multiscale ordination. *Ecology*.

Couteron, P., Pélissier, R., Mapaga, D., Molino, J.F., & Teillier, L. (2002) Ecological valorisation of a management-oriented forest inventory in French Guiana. *Forest Ecology and Management*.

Dale, M.R.T. (1999) *Spatial pattern analysis in plant ecology* Cambridge University Press.

Dale, M.R.T., Dixon, P., Fortin, M.J., Legendre, P., Myers, D., & Rosenberg, M. (2002) Conceptual and mathematical relationships among methods for spatial analysis. *ecography*, 25, 558-577.

Dale, M.R.T. & Mah, M. (1998) The use of wavelets for spatial pattern analysis in ecology. *Journal of Vegetation Science*, 9, 805-814.

Daubechies, I. (1992) *Ten Lectures on Wavelets* SIAM, Philadelphia.

Delor, C., Perrin, J., Truffert, C., Asfirane, F., & Rossi, P. (1998) Images géophysiques dans le socle guyanais. *Géochronique*, 67, 7-12.

Di Bella, G. & Jona-Lasinio, G. (1996) Including spatial contiguity information in the analysis of multispecific patterns. *Environmental and Ecological Statistics*, 3, 269-280.

Diggle, P.J. (1990) *Time Series: a biostatistical introduction* Clarendon Press, Oxford.

Drake, J.B. & Weishampel, J.F. (2000) Multifractal analysis of canopy height measures in a longleaf pine savanna. *Forest Ecology and Management*, 128, 121-127.

Dungan, J.L., Perry, J., Dale, M.R.T., Citron-Pousty, S., Fortin, M.J., Jakomulska, A., Legendre, A., Miriti, M., & Rosenberg, M.S. (2002) A balanced view of scaling in spatial statistical analysis. *Ecography*, 25, 626-640.

Fisher, N.I. (1993) *Statistical Analysis of Circular Data* Cambridge University Press.

Fortin, M.-J., Dale, M.R.T., & Ver Hoef, J.M. (2002). Spatial analysis in ecology. In *Encyclopedia of Environmetrics* (eds A.H. El-Shaarawi & W.W. Piegorsch), Vol. 2, pp. 2051-2058. John Wiley & Sons, Chichester.

Geary, R.C. (1954) The contiguity ratio and statistical mapping. *The Incorporated Statistician*, 5, 115-145.

Gimaret-Carpentier, C. (1999) Analyse de la biodiversité à partir d'une liste d'occurrences d'espèces : nouvelles méthodes d'ordination appliquées à l'étude de l'endémisme dans les Ghâts occidentaux. Thèse de doctorat, Université Lyon 1.

Goodall, D.W. (1974) A new method for the analysis of spatial pattern by random pairing of quadrats. *Vegetatio*, 53, 153-160.

Greig-Smith, P. (1952) The use of random and contiguous quadrats in the study of the structure of plant communities. *Annals of Botany, London*, 16, 293-316.

Greig-Smith, P. (1961) Data on pattern within plant communities. I The analysis of pattern. *Journal of Ecology*, 49, 695-702.

Greig-Smith, P. & Chadwick, M.J. (1965) Data on pattern within plant communities. III. *Acacia-Capparis semi-desert scrub in the Sudan*. *Journal of Ecology*, 53, 465-474.

Griffith, D.A. (2000) Eigenfunction properties and approximations of selected incidence matrices employed in spatial analyses. *Linear Algebra and its Applications*, 321, 95-112.

Guttorp, P.M., Brillinger, D.R., & Schoenberg, F.P. (2002). Point processes, spatial. In *Encyclopedia of Environmetrics* (eds A.H. El-Shaarawi & W.W. Piegorsch), Vol. 3, pp. 1571-1573. John Wiley & Sons, Ltd, Chichester.

Haar, A. (1910) Zur Theorie der Orthogonalen Funktionensysteme. *Mathematische Annalen*, 69, 331-371.

Harville, D.A. (1997) *Matrix algebra from a statistician's perspective* Springer, New York.

Hérissé, C. (2001). Influences des environnements locaux et régionaux sur l'ichtyofaune: structure en réseau et relation de voisinage. Approche exploratoire. Application au Bassin de la Haute-Saône. DEA analyse et modélisation des systèmes biologiques, Université Claude Bernard, Lyon.

Hill, M.O. (1973) The intensity of spatial pattern in plant communities. *Journal of Ecology*, 61, 225-235.

Hutter, S. (2001). *Etude géomorphologique du massif forestier de Counami*. CIRAD.

Jenkins, G.M. & Watts, D.G. (1968) *Spectral analysis and Its Applications* Holden-Day: San Francisco.

Lark, R.M. & Webster, R. (1999) Analysis and elucidation of soil variation using wavelets. *European Journal of Soil Science*, 50, 185-206.

Lark, R.M. & Webster, R. (2001) Changes in variance and correlation of soil properties with scale and location: analysis using and adapted maximal overlap discrete wavelet transform. *European Journal of Soil Science*, 52, 547-562.

Lavit, C. (1988) *Analyse conjointe de tableaux quantitatifs* Masson, Paris.

Lebart, L. (1969) Analyse statistique de la contiguïté. Publication de l'Institut de Statistiques de l'Université de Paris, 28, 81-112.

Legay, J.M. & Barbault, R. (1995). Une révolution silencieuse dans les sciences de la Nature. In La révolution technologique en écologie (eds J.M. Legay & R. Barbault). Masson.

Leps, J. (1990). Comparison of transect methods for the analysis of spatial pattern. In Spatial Processes in plant Communities (eds F. Krahulec, A.D.Q. Agniew, S. Agniew & H.J. Willems), pp. 71-81. SPB Academic Publishing bv, The Hague, Liblice, Tchécoslovaquie.

Méot, A., Chessel, D., & Sabatier, R. (1993). Opérateurs de voisinage et analyse des données spatio-temporelles. In Biométrie et environnement (eds J.D. Lebreton & B. Asselain), pp. 45-72. Masson, Paris.

Milési, J.P., Egal, E., & Ledru, P. (1995) Les minéralisations du nord de la Guyane Française dans leur cadre géologique. Chronique de la recherche minière, 518, 5-58.

Mugglestone, M.A. & Renshaw, E. (1996) A practical guide to the spectral analysis of spatial point processes. Computational Statistics & Data Analysis, 21, 43-65.

Nelson, R. (1988) Using airborne laser data to estimate forest canopy and stand characteristics. Journal of Forestry, 86, 31-38.

Noy-Meir, I. & Anderson, D.J. (1971). Multivariate pattern analysis, or multiscale ordination: towards a vegetation hologram ? In Statistical Ecology, III Many species populations ecosystems and systems analysis (eds G.P. Patil, E.C. Pielou & W.E. Waters), pp. 208-231. Pennsylvania State University Press.

Ollier, S., Chessel, D., Couteron, P., Pélissier, R., & Thioulouse, J. (2003) Comparing and classifying one-dimensional spatial patterns: an application to laser altimeter profiles. Remote Sensing of Environment, 85, 453-462.

Percival, D. (1993) An introduction to spectral analysis and wavelets, International workshop on advanced mathematical tools in metrology.

Percival, D. (2003). Wavelets. In Encyclopedia of Environmetrics (eds A.H. El-Shaarawi & W.W. Piegorisch). John Wiley & Sons, Ltd, Chichester.

Percival, D.B. (1995). On Estimation of the Wavelet Variance.

Percival, D.B. & Walden, A.T. (2000) Wavelet Methods for Time Series Analysis Cambridge University Press.

Perry, J.N., Liebhold, A.M., Rosenberg, M.S., Dungan, J., Miriti, M., Jakomulka, A., & Citron-Pousty, S. (2002) Illustrations and guidelines for selecting statistical methods for quantifying spatial patterns in ecological data. Ecography, 25, 578-600.

Priestley, M.B. (1981) Spectral analysis and time series Academic Press, London.

Qu, Y., Adam, B., Thornquist, M., Potter, J.D., Thompson, M.L., Yasui, Y., Davis, J., Schellhammer, P.F., Cazares, L., Clements, M.A., Wright, G.L., & Feng, Z. (2003) Data reduction using a discrete wavelet transform in discriminant analysis of very high dimensionality data. *Biometrics*, 59, 143-151.

Renshaw, E. (1997) Spectral techniques in spatial analysis. *Forest Ecology and Management*, 94, 165-174.

Renshaw, E. (2002) Two-dimensional spectral analysis of marked point processes. *Biometrical Journal*, 44, 718-745.

Ripley, B.D. (1978) Spectral analysis and the analysis of pattern in plant communities. *Journal of Ecology*, 66, 965-981.

Ritchie, J.C., Evans, D.L., Jacobs, D., Everitt, J.H., & Wertz, M.A. (1993) Measuring canopy structure with an airborne laser altimeter. *Transaction of the ASAE*, 36, 1235-1238.

Schuster, A. (1898) On the Investigation of Hidden Periodicities with Application to a Supposed 26 Day Period of Meteorological Phenomena. *Terrestrial Magnetism*, 3, 13-41.

Smouse, P. & Peakall, R. (1999) Spatial autocorrelation analysis of individual multiallele and multilocus genetic structure. *Heredity*, 82, 561-573.

Stokes, G.G. (1879) Note on Searching for Periodicities. *Proceedings of the Royal Society for Industrial and Applied Mathematics*, 29, 122.

St-Onge, B. (1999) Estimating individual tree heights of the boreal forest using airborne laser altimetry and digital videography. In *Workshop on mapping surface structure and topography by airborne and spaceborne lasers*, Vol. reference 28. ISPRS, Lajolla, Californie.

St-Onge, B.A., Couture, M., & Alleaume, S. (1998) Forest stand structure mapping using a species-controlled textural approach. In *International Forum on Automated Interpretation of High Spatial Resolution Digital Imagery for Forestry*. in press, Victoria.

Vaidyanathan, P.P. (1993) *Multirate Systems and Filter Banks* Prentice-Hall, New Jersey.

Ver Hoef, J.M., Cressie, N.A.C., & Glenn-Lewin, D.C. (1993) Spatial models for spatial statistics: some unification. *Journal of Vegetation Science*, 4, 441-452.

Ver Hoef, J.M. & Glenn-Lewin, C.G. (1989) Multiscale ordination: a method for detecting pattern at several scales. *Vegetatio*, 82, 59-67.

Watt, A.S. (1947) Pattern and process in plant community. *Journal of Ecology*, 35, 1-22.

Weishampel, J., Sun, G., & Harding, D.J. (1996) Remote sensing of forest canopies. *Selbyana*, 17, 6-14.

STRUCTURE D'UN TRAIT BIOLOGIQUE DANS UN ARBRE PHYLOGÉNÉTIQUE

Développement méthodologique à partir de procédures ad hoc

1.	INTRODUCTION.....	147
2.	LA PHYLOGÉNIE COMME NOUVELLE CLASSE DE DONNÉES.....	151
2.1.	Définitions.....	151
2.2.	La classe d'objets 'phylog'	154
3.	REPRÉSENTATION GRAPHIQUE DES DONNÉES.....	158
3.1.	La fonction <code>symbols.phylog(...)</code>	159
3.2.	La fonction <code>dotchart.phylog(...)</code>	160
3.3.	La fonction <code>table.phylog(...)</code>	161
4.	LA MÉTHODE DES CONTRASTES.....	162
4.1.	Le principe des contrastes phylogénétiques.....	162
4.2.	La métrique phylogénétique.....	169
4.3.	Usage de la méthode des contrastes	175
5.	LE TEST D'ABOUHEIF (1999)	177
5.1.	Principe du test d'Abouheif.....	178
5.2.	Le cas d'une variable quantitative.....	180
5.3.	Le cas d'une variable qualitative.....	182
5.4.	La matrice de proximité A	183
5.5.	Conclusions	188
6.	DU CORRÉLOGRAMME A L'ORTHOGRAM.....	188
7.	DISCUSSION ET PERSPECTIVES	190
8.	BIBLIOGRAPHIE	192

1. INTRODUCTION

Le développement du séquençage de l'information génétique et l'amélioration des méthodes de reconstruction phylogénétique ont conduit à un accroissement significatif du nombre de publications visant à établir les relations phylogénétiques entre espèces à partir de données moléculaires (Le Guyader, 2003). La connaissance de ces relations de parenté permettant une approche évolutive des phénomènes, l'utilisation des données phylogénétiques par les biologistes s'est accrue en parallèle, dans des domaines très différents, tels que l'écologie des communautés ou la génétique du développement (Harvey et al., 1996). La prise en compte des relations de parenté entre espèces est venue enrichir un grand nombre de problématiques écologiques telles que :

1. **l'analyse des corrélations interspécifiques**, entre un trait et l'environnement, ou entre deux traits. Cette dernière a conduit au développement de méthodes particulières, appelées méthodes comparatives, pour tenir compte des relations phylogénétiques entre les espèces. De manière très générale, « *comparative studies identify evolutionary trends by comparing the values of some variable or variables across a range of taxa. The variables may include descriptions of the environments inhabited by the organisms as well as phenotypic characters... From Darwin's time to the present, the comparative method has remained the most general technique for asking questions about common patterns of evolutionary change. The comparative method has, however, changed radically in recent year, with the development of methods based on explicit evolutionary and statistical models...* (Harvey et al., 1996) ». Cette nécessité de prendre en compte les proximités évolutives lors de l'analyse des corrélations entre traits biologiques a d'ailleurs été soulignée très tôt par Darwin selon qui « *we may falsely attribute to correlation of growth, structures which are common to whole groups of species, and which in true are simply due to inheritance ; for an ancient progenitor may have acquired through natural selection some one modification in structure, and, after thousands of generations, some other and independent modification ; and these two modifications having been transmitted to a whole group of descendants with diverse habits, would naturally be thought to be correlated in some necessary manner* (citation de Charles Darwin dans l'origine des espèces (1859), reprise dans Cheverud and Dow (1985)) ». A défaut de phylogénie proprement dite, la taxonomie constitue également une information importante de tout corpus de données écologiques qui résume les proximités entre espèces. Pourtant, cet argument de la « nuisance » phylogénétique dans l'analyse des données sur les populations n'a généralement pas été repris par un argument sur la « nuisance » taxonomique dans l'analyse

des données sur les communautés. Par exemple Statzner et al. (1997) étudient les liens entre traits biologiques et traits écologiques sans supposer que ce lien pourrait bien n'être qu'un sous-produit du lien de chacun des deux traits avec la structure taxonomique des cortèges étudiés (Annexe 1.3). La méthode des contrastes, très largement utilisée par les biologistes (Ackerly, 1997) et initialement proposée par Felsenstein (1985), pose explicitement ce problème et le résout dans le cadre d'un modèle d'évolution donné. Toutefois, comme le font remarquer Martins & Hansen (1996), la plupart des méthodes dérivées de la méthode des contrastes n'ôtent pas des données la corrélation phylogénétique, car cette corrélation est prise en compte au travers d'un modèle d'évolution particulier qui reste peu réaliste. On y reviendra par la suite.

2. **l'analyse de la valeur adaptative** d'un caractère, relation entre un caractère et l'environnement occupé par l'espèce. Elle peut être menée en mettant en évidence une corrélation entre un changement d'habitat et un changement dans la valeur de ce caractère. Une première approche consiste à reconstruire les états ancestraux du caractère à partir de la phylogénie pour déterminer quels phénotypes sont les formes dérivées en réponse à la sélection naturelle (changement de l'environnement). Lorsque les espèces du groupe étudié sont nombreuses, présentent différentes valeurs pour le caractère, et occupent des milieux différents, l'approche comparative peut également être utilisée, en mettant en évidence plusieurs associations indépendantes entre la valeur du caractère et le type de milieu (Martins, 2000).

3. **l'analyse des relations évolutives entre traits**. Comme dans le cas précédent, les relations fonctionnelles entre deux traits peuvent être étudiées en reconstituant les états ancestraux de ces traits ou en utilisant l'approche comparative. Par exemple, Smith et al. (1996) étudient la corrélation entre les traits d'histoire de vie des larves et des adultes chez les Echinodermes, tandis que Podos (2001) s'intéresse à la corrélation entre la morphologie du bec et le répertoire vocal des pinsons de Darwin.

4. **l'analyse de l'histoire des communautés**. Elle passe également par la connaissance des relations de parenté entre espèces et permet d'analyser la composante historique de la structure des communautés. Ainsi Malhotra et al. (1996) cherchent à reconstituer la séquence des événements de colonisation des différentes îles des Canaries par les sous-espèces du lézard *Gallotia galloti*. Losos et al. (1998) étudient l'occupation séquentielle des différentes niches écologiques par les lézards habitant différentes îles des Caraïbes, tandis que Vitt et al. (1999) cherchent à comprendre l'origine des différents régimes alimentaires des lézards

d'Amazonie en analysant leurs relations avec la phylogénie et le microhabitat occupé par les espèces.

5. **l'analyse des facteurs à l'origine de la diversification des lignées et du processus de spéciation.** Elle peut être menée en comparant des taxons qui présentent des richesses spécifiques contrastées et diffèrent par les valeurs d'un caractère. De même, les modalités du processus de spéciation peuvent être étudiées en analysant les relations entre l'écologie, l'aire géographique et la phylogénie des espèces (Barracough et al., 1998; Orr & Smith, 1998). L'identification des facteurs pouvant expliquer le fort taux de diversification de certaines lignées par rapport à d'autres repose sur la comparaison entre des « lignées frères », qui ont divergé à partir d'un même ancêtre commun. Cette approche a plusieurs avantages (Barracough et al., 1998) : (i) les lignées ont évolué de façon indépendante depuis leur divergence, (ii) elles ont le même âge donc leurs richesses spécifiques peuvent être comparées et fournissent une estimation du taux de diversification net (différence entre les événements de spéciation et d'extinction), (iii) ils partagent les traits hérités de leur ancêtre commun qui, sans cela, pourraient brouiller la relation entre le taux de diversification et le facteur étudié. La reconnaissance des groupes frères suppose connue la topologie de l'arbre phylogénétique. Cette approche a permis d'identifier de nombreux facteurs susceptibles d'influencer la diversification des lignées, tels que les relations plantes-insectes (Farrell et al., 1991; Mitter et al., 1998). Lorsque les longueurs de branches sont connues, des approches plus fines sont possibles, permettant notamment de déterminer la direction du changement, accélération ou ralentissement du taux de spéciation dans une lignée (Sanderson & Donoghue, 1996).

6. **l'analyse et la mesure de la biodiversité.** Elles peuvent être appréhendées en intégrant les proximités évolutives et taxonomiques entre espèces (Clarke & Warwick, 1999).

Quel que soit la problématique envisagée, les auteurs cherchent à caractériser ce qui dans la structure d'un tableau est directement lié aux données marginales (Figure 3.1). Les individus statistiques sont des taxons dont on connaît une phylogénie (à gauche) ou la taxonomie (à droite). Le tableau de traits contient l'information biologique, le tableau d'habitats contient des listes de taxons que l'on peut trouver dans certaines conditions de milieu, le tableau de relevés contient des distributions d'abondance effectivement observées. Le problème de la mesure du lien entre la phylogénie et une variable s'étend à celui de la mesure du lien entre une variable et un tableau. Si le premier est résolu, le second suivra.

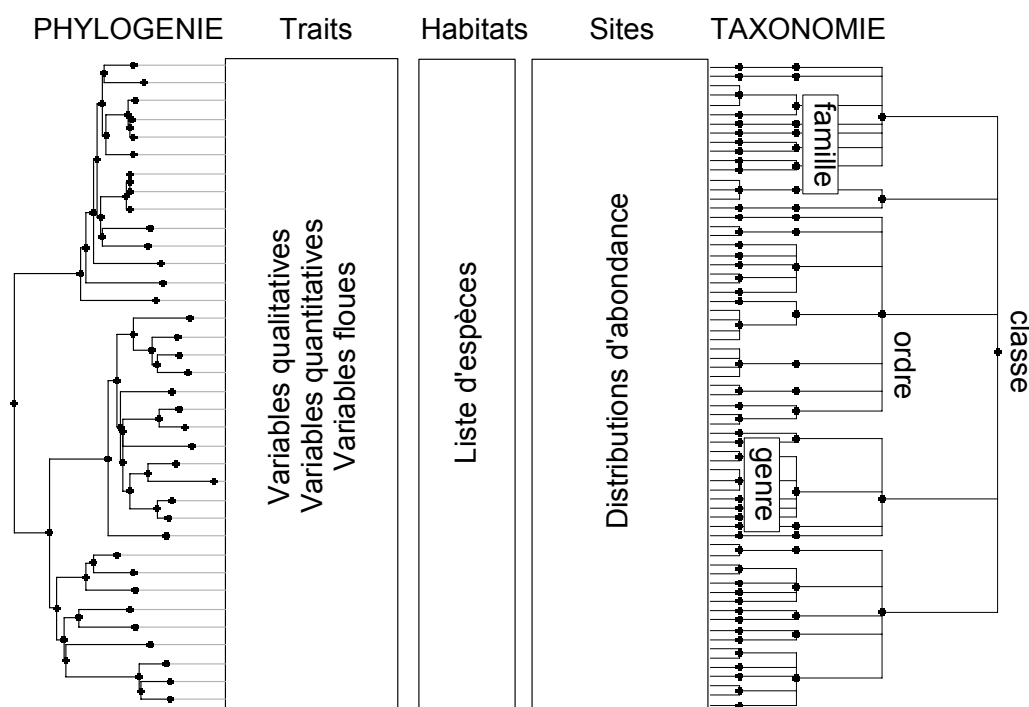


Figure 3.1 : Organisation des données. Les individus statistiques sont des taxons. L'information sur les taxons est organisée sous forme de tableau ; les données évolutives sont organisées sous forme de graphe positionné à la marge de chaque tableau ; on cherche à caractériser ce qui dans la structure du tableau est directement lié aux informations marginales.

Bien souvent, comme dans l'exemple de Statzner et al. (1997), l'information marginale n'est pas prise en considération alors qu'elle caractérise l'ensemble de la structure. De manière générale, la phylogénie (ou la taxonomie) est soit une donnée positive (quel est le lien d'un trait avec la phylogénie, quel est le lien de deux traits dans la phylogénie) soit une donnée négative (quel est le lien d'un trait avec l'environnement, quel est le lien entre deux traits quand on s'est débarrassé des contraintes de la phylogénie). Du point de vue biologique, l'essentiel tient dans la fable de 10 espèces d'Ours dont 5 ont une fourrure courte et 5 ont une fourrure longue. Si les 5 premiers vivent en pays chaud alors que les 5 autres vivent en pays froid, on parle d'adaptation. Si les 5 premiers ont un ancêtre commun alors que les 5 derniers ont un autre ancêtre commun, on parle d'héritage. Si on a les deux, la confusion est totale. De plus, hormis les erreurs d'interprétation que l'on peut commettre, la prise en compte des proximités évolutives pose des problèmes d'ordre statistique. De ce point de vue, l'essentiel tient dans la notion d'autocorrelation phylogénétique (Cheverud & Dow, 1985) qui présuppose que les unités statistiques ne sont plus indépendantes dans la mesure où les espèces sont liées par une histoire commune. Le fait que les espèces étudiées dérivent d'un même ancêtre commun invalide alors les hypothèses statistiques d'indépendance des données, d'égalité des variances et de distribution normale des termes d'erreur.

C'est dans la multiplicité des objectifs biologiques et méthodologiques autour de ces notions qu'il faut voir l'intérêt non négligeable d'intégrer la phylogénie et la taxonomie comme structure canonique en analyse de données. C'est l'objectif de ce chapitre. Dans un premier temps, on définit les classes d'objets et les procédures qui vont nous permettre de manipuler les données phylogénétiques et taxonomiques. On fait ensuite la critique des principales méthodes statistiques de la littérature qui permettent la prise en compte des proximités évolutives en analyse des données écologiques. On revient principalement sur la méthode des contrastes (Felsenstein, 1985), le test non paramétrique d'Abouheif (1999) et l'approche développée par Gittleman et Kot (1990) pour décrire la structure d'un trait biologique dans un arbre phylogénétique.

2. LA PHYLOGÉNIE COMME NOUVELLE CLASSE DE DONNÉES

2.1. Définitions

On considère une phylogénie comme un arbre raciné et valué dont les feuilles sont des OTU (*Operational taxonomic units*), les nœuds de l'arbre étant les HTUs (*Hypothetical taxonomic units*) (Rohlf, 2001). Sur les OTU sont enregistrés une ou plusieurs variables, un ou plusieurs tableaux, des traits biologiques, des distributions d'abondance... La racine est l'ancêtre commun à l'ensemble des p nœuds et des n feuilles. Une branche de l'arbre relie directement un nœud à un autre nœud, ou un nœud à une feuille. Chaque branche définit un sous-arbre raciné au nœud immédiatement inférieur. Un chemin est constitué par l'ensemble des branches qui relie deux unités taxonomiques i et j entre elles. Ce chemin passe par le nœud k , dernier ancêtre commun aux deux unités taxonomiques. Chaque feuille est raccordée à la racine par un chemin unique. Nous admettons les nœuds polytomiques et les nœuds sans bifurcation pour intégrer dans le schéma global les contraintes taxonomiques (Figure 3.2).

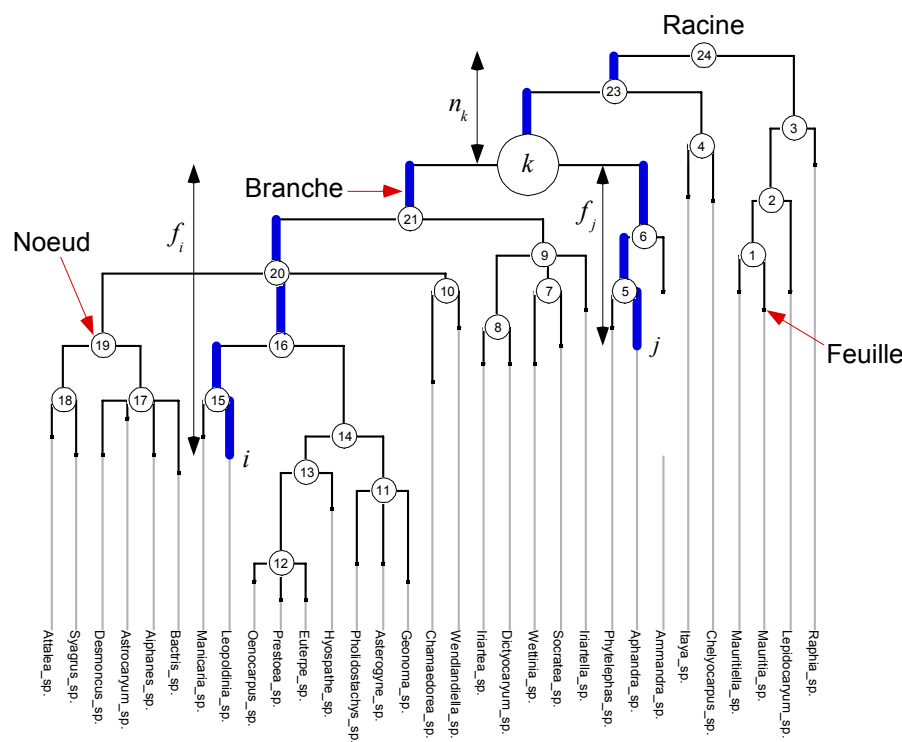


Figure 3.2 : Notions de base dans un arbre phylogénétique où $n = 31$ et $p = 24$

La nature même de l'information contenue dans un arbre phylogénétique est fort diverse car la notion de phylogénie recouvre des logiques fort différentes. Il existe essentiellement trois types de phylogénie (Figure 3.3) :

1. Le premier type dérive des **modèles cladistiques**, mode de classification fondé par Hennig et basé sur une recherche de relation de parenté à l'aide d'états de caractères dérivés partagés. Il retrace les relations de parenté entre espèces. Deux feuilles ont ou n'ont pas un ancêtre commun. Les longueurs de branches n'ont pas de sens : elles sont toutes unitaires.
2. Le second est celui des **modèles historiques**. Chaque ancêtre possède une date d'apparition et toutes les distances à la racine sont égales. L'unité est le myr (million d'années). La contrainte est importante puisque la reconstitution passe par la connaissance systématique des formes fossiles, ce qui est extrêmement rare.
3. Le troisième est celui des **modèles de divergence**, développés en phylogénie moléculaire. Les longueurs de branches sont des estimations de la distance à l'ancêtre en pourcentage de mutations. Les arbres qui en résultent, correspondent à des modèles évolutifs complets et calés sur une échelle de temps. L'utilisation de ces arbres est de plus en plus fréquente en écologie, en particulier pour reconstituer par parcimonie les états ancestraux des traits dans la phylogénie.

Par ailleurs, quand plusieurs distances ont été construites sur une liste de taxons, la phylogénie rend compte d'un consensus entre plusieurs points de vue et/ou plusieurs méthodes. Les espèces qui intéressent le biologiste étant rarement toutes présentes dans les phylogénies publiées, des méthodes ont été développées pour établir des super-arbres synthétisant des arbres phylogénétiques fragmentaires (Sanderson et al., 1993). La base de données phylogénétiques TreeBASE (Sanderson & Donoghue, 1998) a précisément pour objectif la recherche et la combinaison d'arbres et de données provenant de sources différentes. Lorsque les arbres sources sont compatibles, la construction du super-arbre ne pose pas de difficultés. Dans le cas contraire, plusieurs approches sont possibles. Elles n'utilisent que deux types d'information, à savoir la topologie et les racines des arbres sources : les données éventuelles concernant la longueur des branches ne sont pas utilisées.

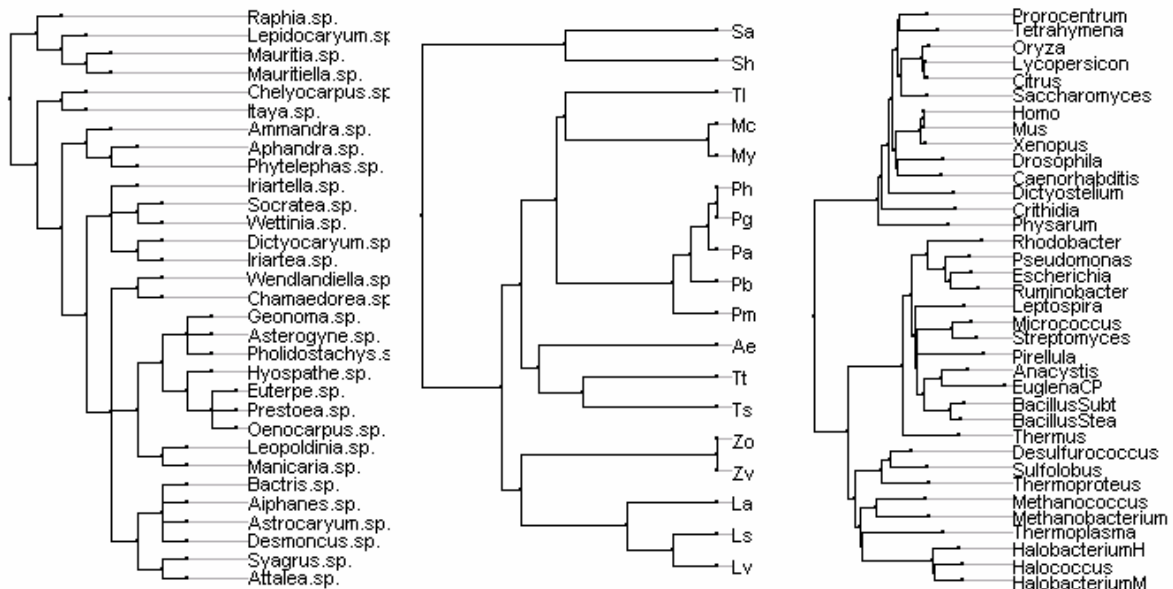


Figure 3.3 : **A gauche**, phylogénie bilan d'expertise, sur 31 genres de palmiers amazoniens. Ces données ont été compilées par C. Gimaret-Carpentier ([data\(newick.org\)](http://data.newick.org)). **Au centre**, modèle hypothétique des relations phylogénétiques de 16 espèces de lézards ([data\(lizards\)](http://data(lizards))). La distance de la base à la racine est de 35 million d'années (Bauwens & Díaz-Uriarte, 1997). **A droite**, phylogénie évaluée d'un ensemble de groupes de séquences d'ARN basée sur les pourcentages de substitutions destinée à la recherche des caractéristiques de l'ancêtre commun ([data\(njplot\)](http://data(njplot))). Cet arbre est l'exemple de base du logiciel njplot (Perrière & Gouy, 1996).

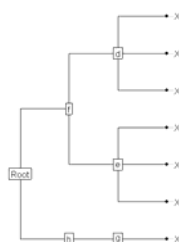
On appellera ici phylogénie, une information exprimée dans un arbre (graphe non orienté, connexe et sans cycle), valué (par défaut toutes les arêtes ont une longueur de 1). Si on définit une racine, l'arbre devient une arborescence. Les phylogénies dérivées d'une distance sont en principe non racinées. On considérera que la présence d'une racine implicite fait partie des données.

Notre objectif est d'intégrer cette structure de données dans la logique de l'analyse linéaire des données soit comme contrainte positive, soit comme élément parasite. Il s'agit d'étendre les analyses inter et intra-classes associées à une partition, à des structures plus complexes telles que les taxonomies ou les phylogénies. Pour cela on définit dans R une nouvelle classe d'objets 'phylog' qui permet de gérer l'importation et la manipulation d'arbre phylogénétique dans l'environnement de travail du logiciel.

2.2. La classe d'objets 'phylog'

La classe d'objets 'phylog' définit une structure d'association entre n points qui s'exprime sous forme d'arborescence. Les arbres sont entrés par le format 'Newick' (<http://evolution.genetics.washington.edu/phylip/newicktree.html>), très employé dans les logiciels dédiés à l'élaboration de phylogénies. La fonction `newick2phylog(...)` (Annexe 2.14) en fait des listes de la classe 'phylog' (Annexe). Les fonctions `plot2phylog(...)` (Annexe 2.20) et `radial2phylog(...)` (Annexe 2.20) assurent la représentation graphique des objets de cette classe :

```
# on peut construire un arbre manuellement
tre <- "(((1,2,3)d,(4,5,6)e)f,((7)g)h);"
phy <- newick2phylog(tre)
phy
Phylogenetic tree with 7 leaves and 6 nodes
$class: phylog
$call: newick2phylog(x.tre = tre)
$tre: ((X1,X2,X3)d,(X4,X5,X6)e)f,((X7)g)h)Root;
```



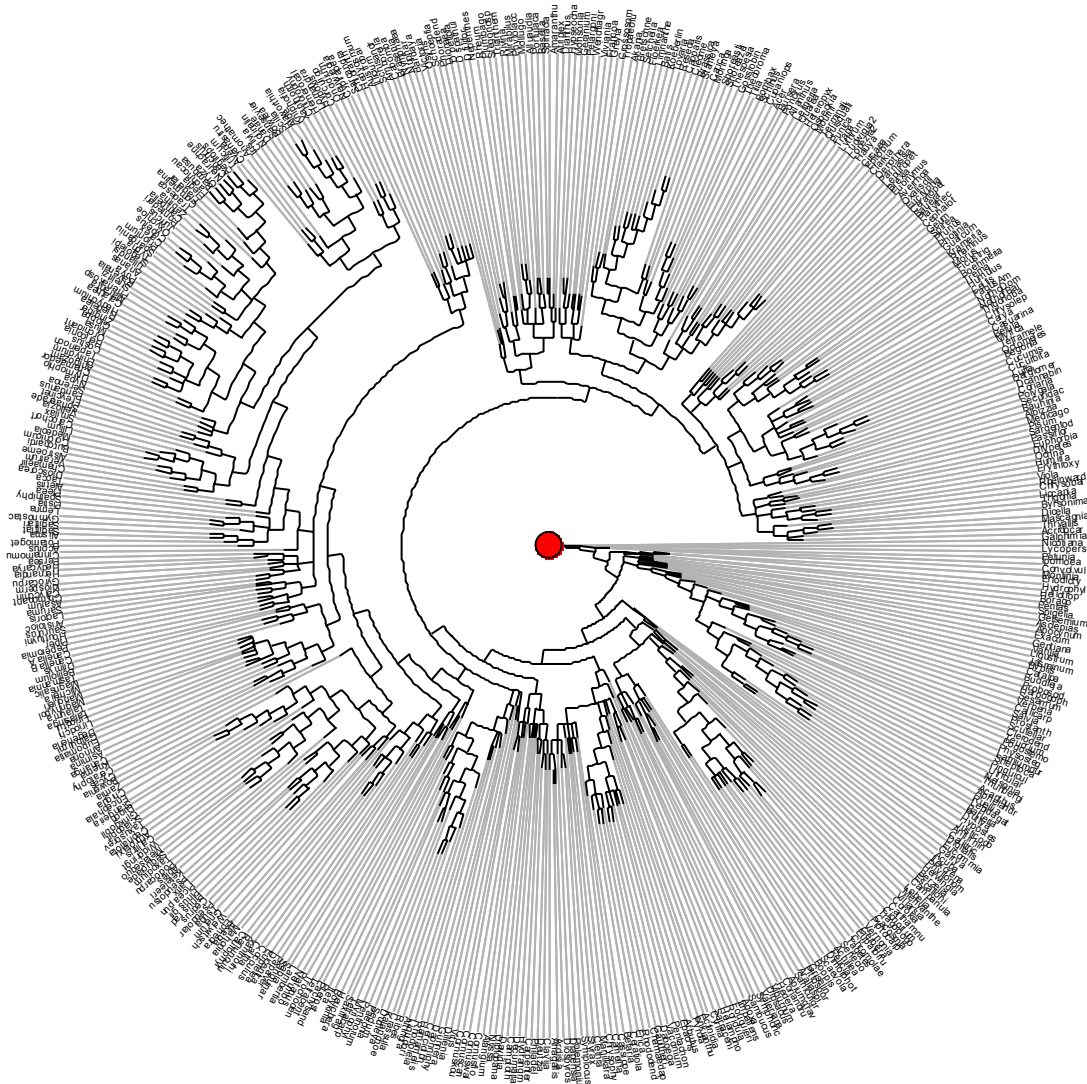
	class	length	content
\$leaves	numeric	7	length of the first preceeding adjacent edge
\$nodes	numeric	6	length of the first preceeding adjacent edge
\$parts	list	6	subsets of descendant nodes
\$paths	list	13	path from root to node or leave
\$droot	numeric	13	distance to root

	class	dim	content
\$Wmat	matrix	7-7	W matrix : root to the closest ancestor
\$Wdist	dist	21	Nodal distances
\$Wvalues	numeric	6	Eigen values of QWQ/sum(Q)
\$Wscores	data.frame	7-6	Eigen vectors of QWQ '1/n' normed
\$Amat	matrix	7-7	Topological proximity matrix A
\$Avalues	numeric	6	Eigen values of QAQ matrix
\$Adim	integer	1	number of positive eigen values of QAQ
\$Ascores	data.frame	7-6	Eigen vectors of QAQ '1/n' normed
\$Aparam	data.frame	6-4	Topological indices for nodes
\$Bindica	data.frame	7-6	class indicator from nodes
\$Bscores	data.frame	7-6	Topological orthonormal basis '1/n' normed
\$Bvalues	numeric	6	xtWx values for orthonormal basis
\$Blabels	character	6	Nodes labelling from orthonormal basis

`plot(phy)`

```
# on peut également récupérer les chaînes des caractères au format Newick
# exemple d'arbre consensus de 8975 arbres sur 500 espèces de plantes
# diffusion à http://www.cis.upenn.edu/~krice/treezilla/
```

```
tre <- newick.eg[[10]]
tre # chaîne de 1374 caractères qui résume une matrice de distance 500x500
"(Nicotiana,((((((((((((((((Galphimia,Acridocar ... Petunia),Lycopersi);"
phy <- newick2phylog(tre)
radial.phylog(phy, circ = 1.70, clabel.leaves = 0.3,cnodes = 0, cleaves = 0)
```



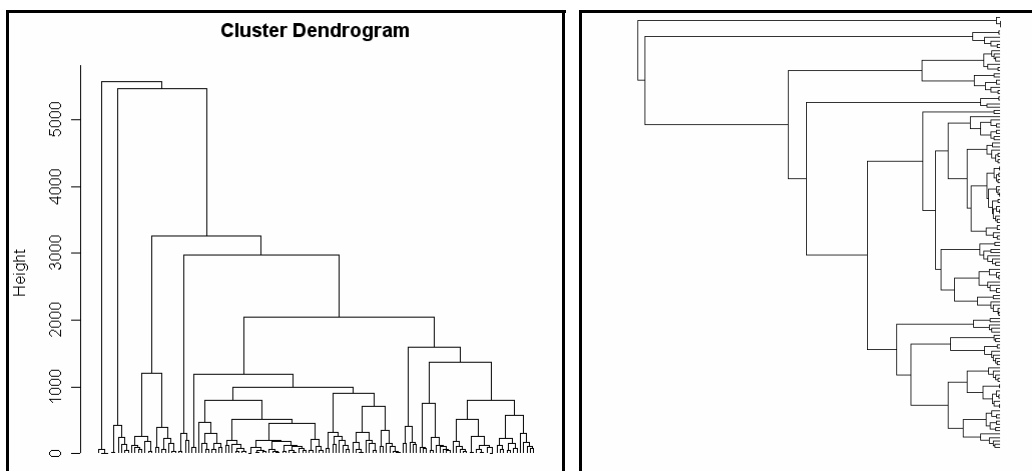
Cette classe d'objets ne recouvre pas que les phylogénies proprement dites. A défaut de phylogénie, on a vu en introduction que la taxonomie est également une information importante de tout corpus de données écologiques. D'un point de vue structurel, une taxonomie peut-être vue comme un ensemble de partitions emboîtées qui représente, comme la phylogénie, une information qui peut être introduite dans l'analyse des données. C'est pourquoi on insère une procédure `taxo2phylog(...)` (Annexe 2.14) qui fait le lien entre taxonomie et phylogénie en terme de structure de données.

```
taxo <- as.taxo(bsetal197$taxo)
bsetal.phy <- taxo2phylog(taxo, FALSE)
```


appealing properties: it attempts to capture phylogenetic diversity rather than simple richness of species and is more closely linked to functional diversity ». L'opération renvoie la taxonomie à la phylogénie et la phylogénie à la distance entre feuilles d'un arbre. On verra que la distance phylogénétique est facilement euclidienne et que la porte ouverte par les auteurs débouche sur une mesure euclidienne de la biodiversité (Champely & Chessel, 2002) donc sur une méthode typologique euclidienne associée dans l'axiomatic de Rao (1982). Elle contient bien des perspectives et fait l'objet de la thèse de Sandrine Pavoine. La classe de données taxonomiques 'taxo' (Annexe 2.24) et la classe de données phylogénétiques 'phylog' sont donc voisines et réfléchir simultanément sur les deux a un sens statistique et un sens biologique. On manipulera donc par la suite les taxonomies par l'intermédiaire d'objets de la classe 'phylog'.

Entre les deux structures de données, on trouve les hiérarchies de partitions générées par les classifications ascendantes hiérarchiques. La fonction `hclust2phylog(...)` (Annexe 2.14) fait le lien entre ces structures et les phylogénies et exprime à nouveau une position commune en terme de structure des données. Le passage inverse n'est possible que pour les phylogénies ne comportant que des bifurcations :

```
X <- prep.fuzzy.var(bsetal97$biol,bsetal97$biol.blo)
dudil <- dudi.fca(X,scan = F)
hcl <- hclust(dudil$tab^2, "ward")
plot(hcl, hang = -1, labels = FALSE)
phy <- hclust2phylog(hcl)
plot(phy, clabel.leaves = 0, f.phylog = 0.9, cleaves = 0)
```



On retiendra que la classe d'objets 'phylog' définit donc une structure d'association entre n points qui s'exprime sous forme d'arborescence. On recouvre ainsi :

- les généalogies réelles ou estimées, partiellement ou totalement résolues ;
- les phylogénies temporelles dont les longueurs de branches estiment les dates de naissance des entités taxonomiques ;
- les taxonomies portant sur n taxa avec un nombre de niveaux quelconques ;
- les classifications ascendantes hiérarchiques.

3. REPRÉSENTATION GRAPHIQUE DES DONNÉES

Curieusement, si les représentations des arbres phylogénétiques sont universellement présentes, la représentation graphique simultanée des phylogénies et des données d'observation semble fort rare. Il y a pourtant tout à gagner à voir deux structures de données l'une en face de l'autre.

A titre d'illustration, on considère les traits d'histoires de vie de poissons téléostéens étudiés dans Rochet et al. (2000) (Figure 3.4, Annexe 1.14). Par la suite on travaillera sur le logarithme de ces variables. L'arbre représente une synthèse des publications récentes sur le sujet. Les critères de classification des espèces sont principalement morpho-anatomiques, quoique pondérés par des études moléculaires. La longueur des branches, qui représente dans ce type d'arbre le temps écoulé entre deux différenciation ne sont pas connues à l'heure actuelle. De plus, le taux d'évolution varie selon les traits, voire selon les branches envisagées. L'arbre n'est donc pas valué : la longueur des branches est partout la même, et arbitrairement égale à un. L'objectif est de différencier ce qui, dans les relations entre traits biologiques, relève de la proximité évolutive des espèces, de ce qui n'en relève pas. Les auteurs ont mis en œuvre un modèle autorégressif pour estimer la composante phylogénétique.

Il apparaît indispensable, si l'on veut se faire une idée sérieuse sur la structure des données, de fournir avant toute analyse statistique, une représentation graphique de ces dernières. Par analogie, on peut dire qu'il ne viendrait pas à l'idée d'un géostatisticien de se lancer dans la modélisation de la structure spatiale de variables géoréférencées sans avoir au préalable établi une cartographie des données. On définit plusieurs stratégies.

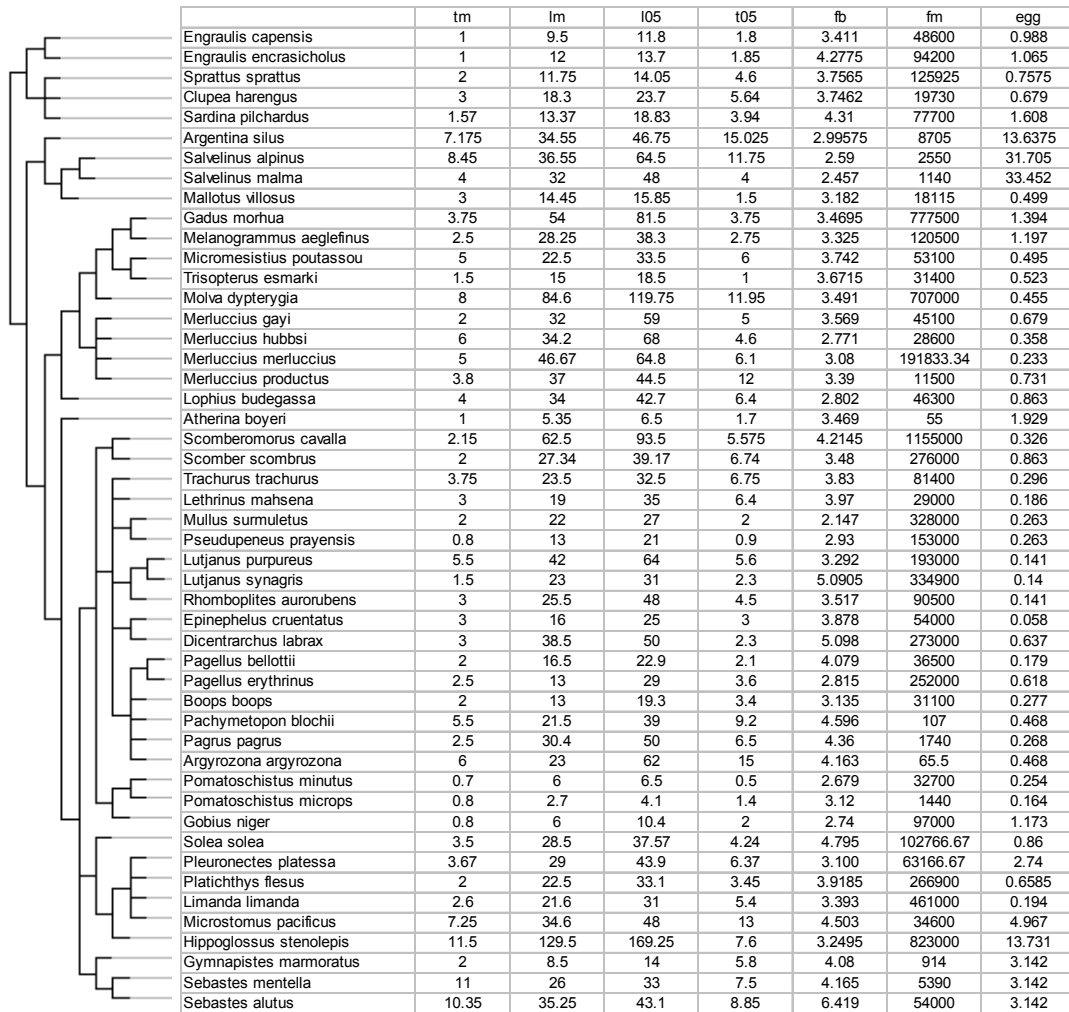
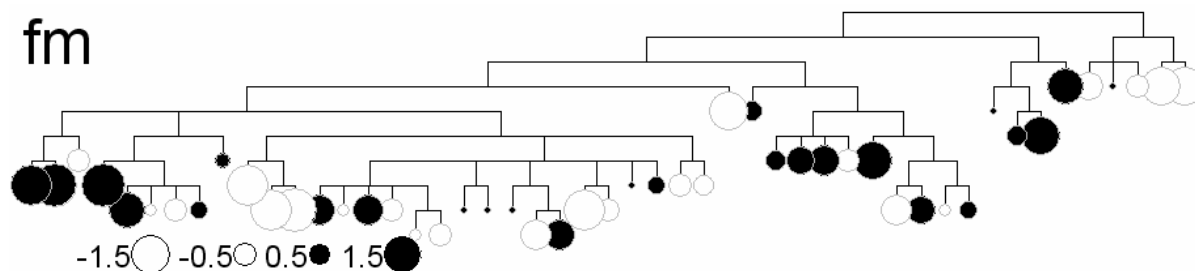


Figure 3.4 : Traits biologiques et phylogénie de poissons marins (d’après Rochet (2000)). Les traits correspondent respectivement à l’âge de maturité sexuelle (**tm**), la longueur à la maturité sexuelle (**lm**), la longueur au temps ‘5% de survivants’ (**i05**), le temps écoulé entre la maturité sexuelle et le temps ‘5% de survivants’ (**t05**), la mesure de l’accroissement de la fécondité avec la taille des femelles (**fb**), fécondité à la maturité (**fm**), et le volume des œufs (**egg**).

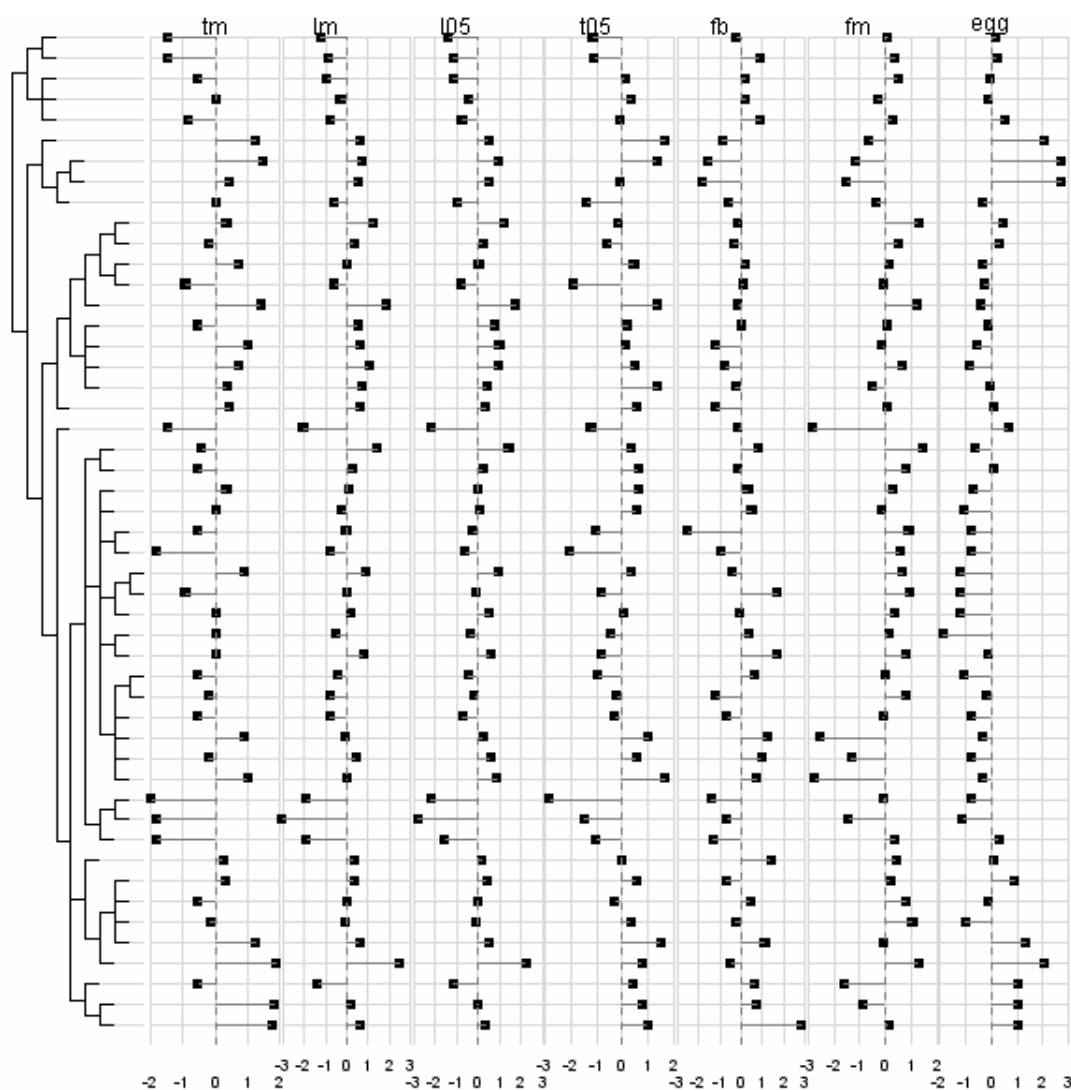
3.1. La fonction `symbols.phylog(...)`

Avec la fonction `symbols.phylog(...)` (Annexe 2.22), les traits (variables quantitatives) sont centrés et normés avant d’être représentés. On choisit de placer des symboles au niveau des feuilles, dont la taille est proportionnelle à la valeur prise par chaque trait pour le taxon considéré (noire pour les valeurs positives et blanche pour les négatives).



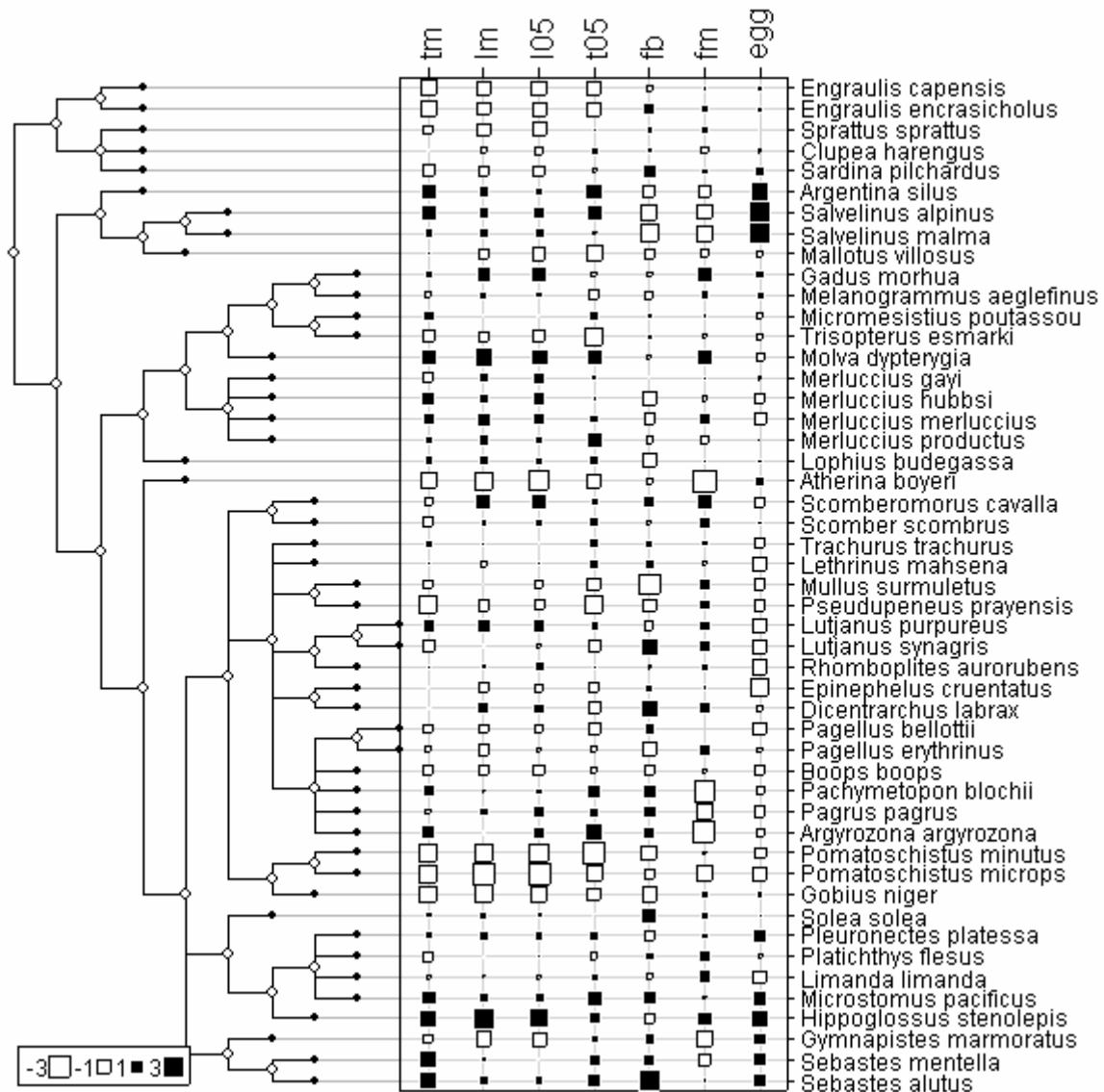
3.2. La fonction `dotchart.phylog(...)`

Avec la fonction `dotchart.phylog(...)` (Annexe 2.23), on choisit de placer en face de chaque espèce un point dont la position relative sur une grille indique la valeur prise par chaque trait. Cette représentation s'inspire du dotplot de Cleveland (1994), défini comme « *the graphical method for measurements that have labels* ». Elle permet la représentation d'une ou plusieurs variables quantitatives simultanément.



3.3. La fonction `table.phylog(...)`

Avec la fonction `table.phylog(...)` (Annexe 2.23), on donne une représentation simultanée des valeurs de plusieurs traits sous la forme de symboles. Cette fonction est une généralisation de la fonction `symbols.phylog(...)` à un tableau de traits.



Ces trois représentations graphiques donnent une première idée sur l'organisation de la variabilité de chaque trait le long de la phylogénie. On repère rapidement les traits fortement structurés, les espèces originales, les traits ayant la même structure ainsi que les espèces ayant le même profil. Elles posent clairement la question de la mesure statistique du lien entre la phylogénie et une variable ou un ensemble de variables :

- l'organisation des données du tableau est-elle indépendante par rapport à l'arbre ?

- si dépendance il y a, quelle est sa nature ou son intensité ?

Par exemple, Gittleman et al. (1996) cherchent à comparer la plasticité évolutive de différentes catégories de traits, faisant l'hypothèse que les traits comportementaux devraient être les plus labiles et les traits morphologiques les plus fortement corrélés à la phylogénie, les traits d'histoire de vie occupant une position intermédiaire. La question n'est pas simple. Des solutions à ces problèmes ont été proposées dans la littérature et sont regroupées sous le terme de méthodes comparatives (Harvey & Pagel, 1991). La plupart, à l'exception des travaux de Cornillon (2000) et de Rolhf (2001), relève de la statistique *ad hoc*. A partir d'une revue critique de ces procédures rencontrées dans la littérature, on cherche à définir des procédures canoniques permettant la prise en compte des proximités évolutives en analyse des données.

4. LA MÉTHODE DES CONTRASTES

4.1. Le principe des contrastes phylogénétiques

La méthode des contrastes phylogénétiques indépendants (PIC, *phylogenetic independant contrasts*), introduite par Felsenstein (1985), est la méthode comparative la plus utilisée en biologie comparative. C'est la première méthode statistique qui propose de prendre en compte l'influence des proximités évolutives sur les corrélations entre traits : « *my intention is to point out a serious statistical problem with all numerical studies that involves a comparison of two phenotypes across a range of species or higher taxa, or a comparison of one phenotype with an environmental variable. It arises from the fact that species are part of a hierarchically structured phylogeny, and thus cannot be regarded for statistical purposes as if drawn independently from the same distribution* ».

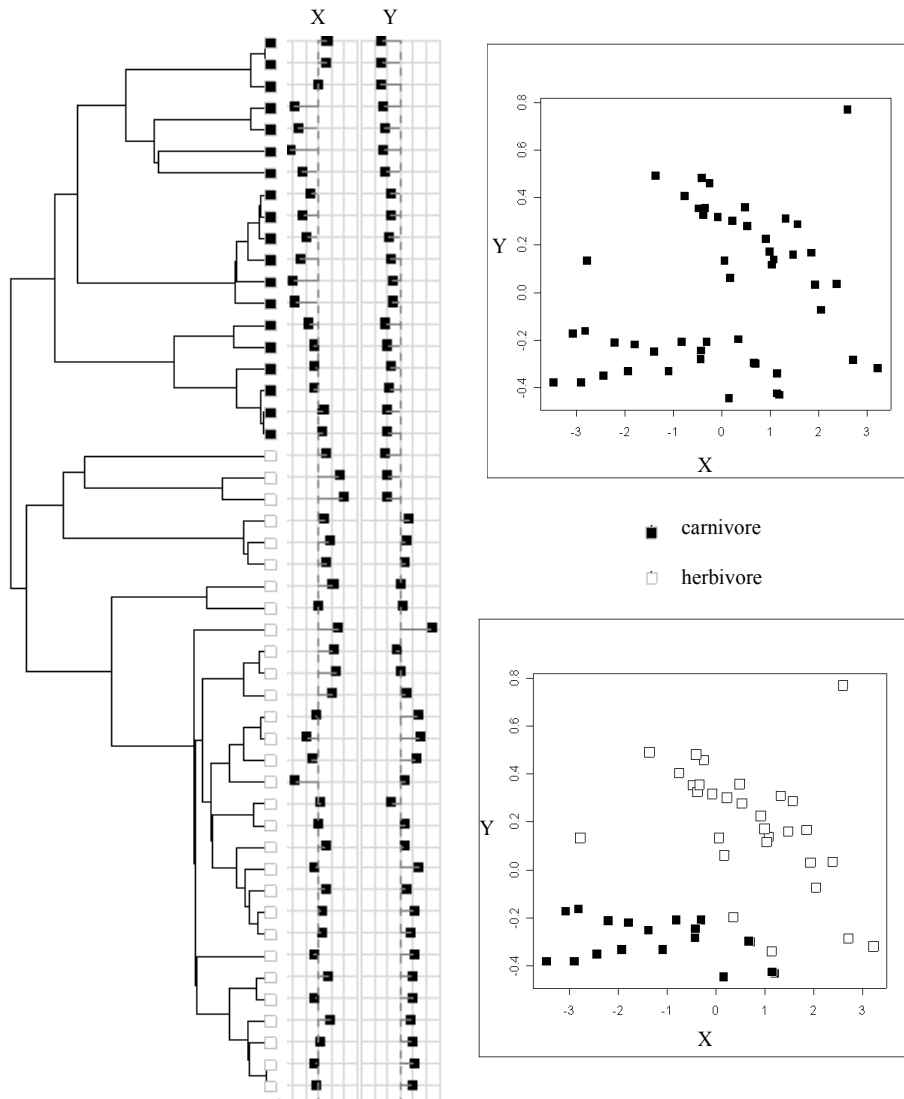


Figure 3.5 : A gauche, représentation graphique de la phylogénie avec ses des deux groupes monophylétiques (carnivore/herbivore) et les deux traits biologiques (poids- X et rapport métatarse/fémur- Y , voir Annexe 1.4). Les deux traits sont corrélés (en haut, à droite), bien que la corrélation à l'intérieur de chaque groupe soit faible (en bas, à droite).

Felsenstein (1985) part du constat suivant, que l'on illustre à partir de données réelles publiées par Garland et Janis (1993). « *Suppose that the data turned out to look like in Figure 3.5 : The phylogeny shows that a large number of species consist actually of two groups of closely related species (carnivore/herbivore). There appears to be a significant regression of Y on X . If the points are distinguished according to which monophyletic group they come from, we can see that there is two clusters. Within each of these groups there is non significant regression of one character on the other...* ».

De manière plus générale, la phylogénie constitue un facteur de confusion pouvant modifier sérieusement les propriétés des statistiques classiquement utilisées telle que la corrélation. La méthode des contrastes phylogénétiques indépendants a justement été

proposée dans le but d'éliminer la composante phylogénétique lors de la mesure de la corrélation entre deux traits. La mise en œuvre de cette méthode demande une lecture et une interprétation attentive du texte original, son créateur (Felsenstein, 1985) ayant opté pour une présentation intuitive. Soit $\mathbf{X} = (X_1, X_2, \dots, X_n)^t$ une variable aléatoire dans \mathbb{R}^n avec X_i une variable aléatoire réelle associée à chaque feuille i . Chaque trait \mathbf{x} peut être vu comme une réalisation de la variable aléatoire \mathbf{X} .

La méthode des contrastes est intimement liée à la modélisation de la variable aléatoire \mathbf{X} par un mouvement brownien (marche au hasard en temps continu). Le principe du mouvement brownien est illustré sur la Figure 3.6 : à partir d'un exemple tiré de l'ouvrage de Felsenstein (2004). Le modèle repose sur les assertions suivantes :

- l'évolution le long d'une branche suit un mouvement brownien standard de dérive nulle dont la variance est proportionnelle à la longueur de la branche (par exemple, $X_1 - X_{11} \sim N(0, \sigma^2 0.3)$).
- les évolutions après bifurcation sont indépendantes (par exemple, $(X_1 - X_{11})$ et $(X_2 - X_{11})$ sont indépendantes).
- quand l'arbre est enraciné, on définit la valeur du trait au niveau de la racine par μ

Avec ces hypothèses, on peut facilement définir la loi des feuilles de l'arbre :

$$\mathbf{X} = (X_1, X_2, \dots, X_n)^t \sim N(\mu \mathbf{1}_n, \sigma^2 \mathbf{W}),$$

où \mathbf{W} est la matrice de variance covariance du modèle associant à chaque couple de feuilles la distance à la racine du premier ancêtre commun aux deux feuilles (Figure 3.6 :). Sur la diagonale de \mathbf{W} , on trouve en particulier les distances à la racine de chaque feuille.

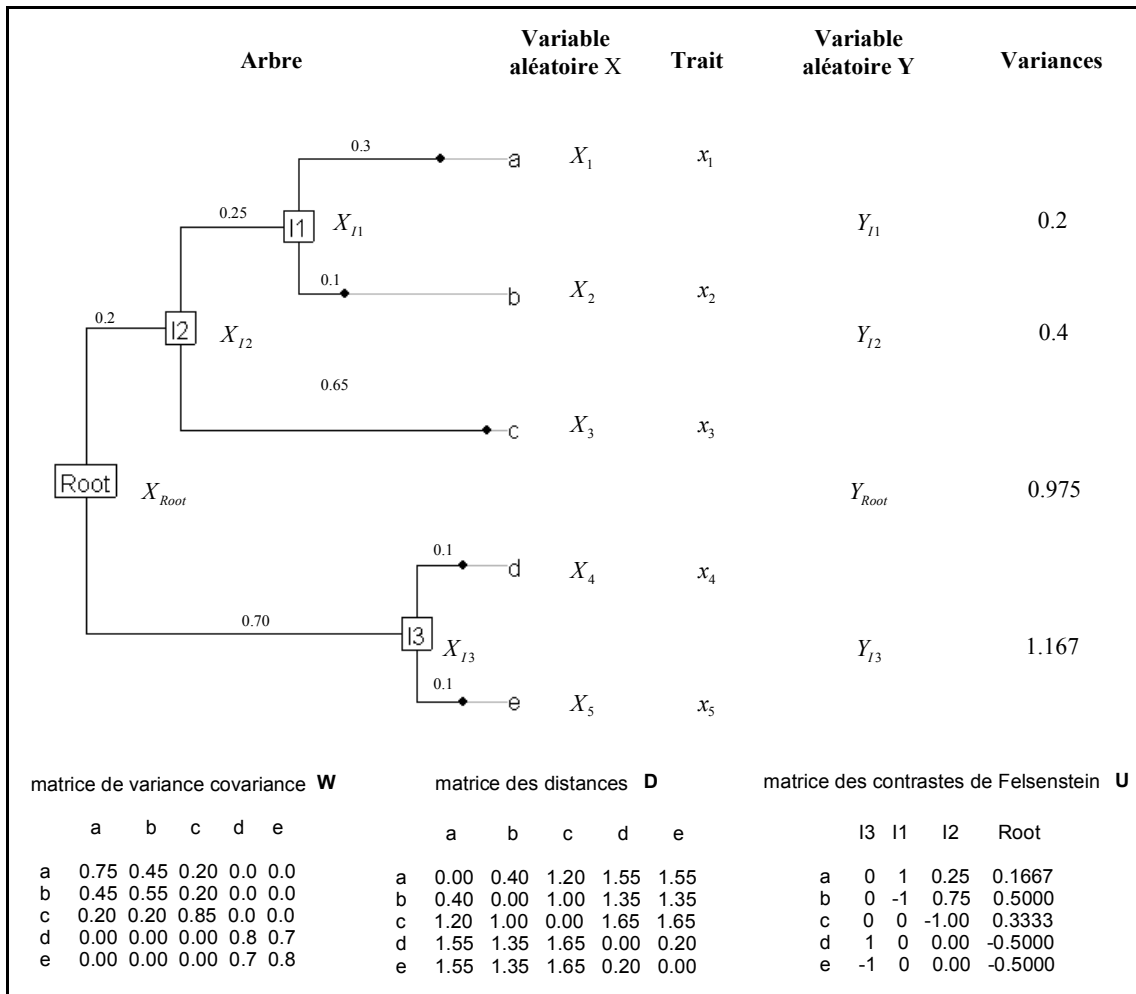
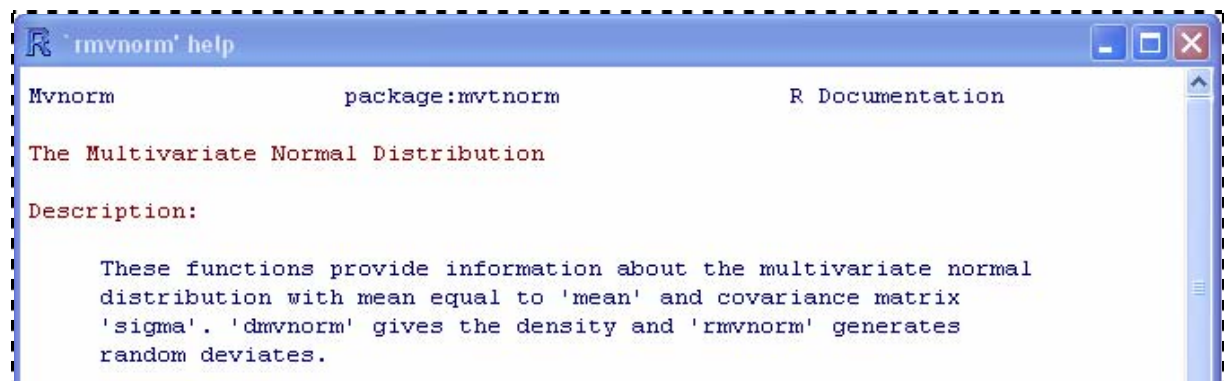


Figure 3.6 : Le modèle brownien et la définition des contrastes de Felsenstein.

On peut alors facilement simuler plusieurs traits le long d'une phylogénie sous l'hypothèse d'un mouvement brownien à partir de la fonction `rmvnorm(...)` de la librairie

`mvtnorm` :

```
library(mvtnorm)
help(rmvnorm)
```

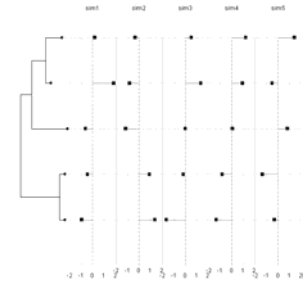
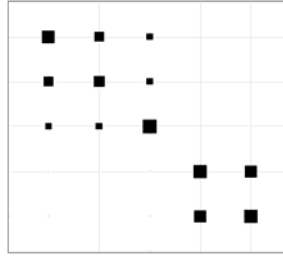


```
tre <- c("((a:0.3,b:0.1)I1:0.25,",
        "c:0.65)I2:0.2,(d:0.1,",
```

```

"e:0.1)I3:0.7)Root;")
phy <- newick2phylog(tre)
sim <- rmvnorm(5, rep(0, 5),
sigma = phy$Wmat)
sim <- t(sim)
sim <- as.data.frame(sim)
row.names(sim) <- letters[1:5]
names(sim) <- paste(rep("sim",
5), 1:5, sep = "")
table.value(phy$Wmat,
cleg = 0, clabel.row = 0,
clabel.col = 0)
dotchart.phylog(phy, sim,
ceti = 0.75, csub = 0.75)

```



En statistique, dans le cadre du modèle linéaire, un contraste \mathbf{u} est un vecteur de \mathbb{R}^n qui définit une nouvelle variable aléatoire comme combinaison linéaire des variables aléatoires

$X_i : \mathbf{u}'\mathbf{X} = \sum_{i=1}^n u_i X_i$. C'est cette définition que nous adopterons par la suite quand on parlera des contrastes. Pour éviter toute confusion, on distinguera le vecteur \mathbf{u} (contraste), la variable aléatoire $\mathbf{u}'\mathbf{X}$ et la valeur prise $\sum_{i=1}^n u_i x_i$ par cette variable quand X_i a pris au cours d'une expérience la valeur x_i (score du contraste selon Felsenstein (2003)).

La définition des contrastes \mathbf{u} par Felsenstein (1985), reprise par Rohlf (2001), est purement algorithmique : elle assure, sous l'hypothèse du mouvement brownien l'indépendance des variables aléatoires Y_i associées aux $n-1$ nœuds de la phylogénie. C'est dans ce sens que Felsenstein (1985) les appellent des contrastes indépendants, dans la mesure où leur définition assure l'indépendance des variables aléatoires qui leurs sont associées. En aucune mesure, les vecteurs \mathbf{u} de Felsenstein (1985) ne sont censés être indépendants du point de vue algébrique. La même remarque prévaut lorsqu'il parle de la variance des contrastes : pour Felsenstein (1985), les contrastes sont normalisés lorsque les variables aléatoires qui leurs sont associés ont une variance de 1 à la constante σ^2 près.

Le calcul des contrastes peut être illustré à partir de l'exemple présenté sur la Figure 3.6. On s'intéresse dans un premier temps, au contraste associé au nœud II. D'après les hypothèses du mouvement brownien :

$$\begin{cases} (X_1 - X_{II}) \sim N(0, 0.3\sigma^2) \\ (X_2 - X_{II}) \sim N(0, 0.1\sigma^2) \\ (X_1 - X_{II}) \text{ et } (X_2 - X_{II}) \text{ sont indépendantes} \end{cases}$$

On peut alors définir une variable aléatoire Y_{I1} associée au contraste $\mathbf{u}'_{I1} = (1, -1, 0, 0, 0) / \sqrt{0.3 + 0.1}$ par

$$Y_{I1} = \frac{X_1 - X_2}{\sqrt{0.3 + 0.1}} = \frac{(X_1 - X_{I1}) - (X_2 - X_{I1})}{\sqrt{0.3 + 0.1}} = \mathbf{u}'_{I1} \mathbf{X}.$$

D'après les hypothèses du mouvement brownien $Y_{I1} \sim N(0, \sigma^2)$. Le raisonnement est le même pour le nœud $I3$ mais pour pouvoir continuer la démarche avec le nœud $I2$, il faut définir un estimateur de la variable aléatoire X_{I1} . Felsenstein (1985) propose d'utiliser une combinaison linéaire des variables portées par les deux feuilles a et b :

$$\hat{X}_{I1} = aX_1 + bX_2.$$

Sous l'hypothèse du mouvement brownien, a et b doivent vérifier les deux conditions suivantes :

$$\begin{cases} E(\hat{X}_{I1}) = aE(X_1) + bE(X_2) \\ \text{cov}(X_1 - X_2, \hat{X}_{I1}) = 0 \end{cases} \Leftrightarrow \begin{cases} a + b = 1 \\ aw_a - bw_b = 0 \end{cases} \Leftrightarrow \begin{cases} a = \frac{w_b}{w_a + w_b} \\ b = \frac{w_a}{w_a + w_b} \end{cases}$$

avec w_a et w_b représentant respectivement les longueurs de branches qui mènent des feuilles a et b au nœud II . Ainsi, il existe une variable aléatoire \hat{X}_{I1} associée au vecteur $\mathbf{v}'_{I1} = (w_b, w_a, 0, 0, 0) / (w_a + w_b) = (0.1, 0.3, 0, 0, 0) / (0.3 + 0.1)$ définie par

$$\hat{X}_{I1} = \frac{0.1X_1 + 0.3X_2}{0.3 + 0.1} = \mathbf{v}'_{I1} \mathbf{X}.$$

La variable \hat{X}_{I1} a la même espérance que X_{I1} mais sa variance est légèrement plus grande.

Cornillon (2000) montre en effet que $\text{var}(\hat{X}_{I1}) = \text{var}(X_{I1}) + \frac{w_a w_b}{w_a + w_b}$. Ainsi, lorsque l'on

calcule le contraste au nœud $I2$, à partir de l'estimation du trait au nœud II , on doit tenir compte de cette différence de variance dans le calcul afin que la variable associée au contraste

$I2$ ait une variance égale à σ^2 . On redéfinit alors la valeur w_{I1} par $w_{I1}^\wedge = w_{I1} + \frac{w_a w_b}{w_a + w_b}$.

Felsenstein (1985) introduit ainsi de manière algorithmique la définition de $n-1$ contrastes \mathbf{u} rangés en colonnes dans une matrice de contrastes \mathbf{U} de dimension $n \times (n-1)$ (Figure 3.6 :, ce qui définit une nouvelle variable aléatoire multidimensionnelle :

$$\mathbf{Y} = (Y_1, Y_2, \dots, Y_{n-1})^t = \mathbf{U}' \mathbf{X}.$$

A partir de la définition algorithmique des contrastes de Felsenstein (1985), on peut également introduire une matrice \mathbf{V} de dimension $n \times (n-1)$ définissant les estimateurs du trait au niveau des nœuds internes. La fonction `phylog.pic(...)` (Annexe 2.17) calcule pour toute phylogénie résolue les matrices \mathbf{U} et \mathbf{V} :

```
tre <- "((a:0.3,b:0.1)I1:0.25,c:0.65)I2:0.2,(d:0.1,e:0.1)I3:0.7)Root;"
phy <- newick2phylog(tre)
pic <- phylog2pic(phy)
summary(pic)
```

	Length	Class	Mode	
contrastes	4	data.frame	list	# $\mathbf{u}^i \sqrt{\mathbf{var}_i}$
prediction	4	data.frame	list	# $\mathbf{v}^i \sqrt{\mathbf{var}_i}$
variance	4	-none-	numeric	# \mathbf{var}

```
pic$contrastes
  I3 I1  I2  Root
a  0  1  0.25 0.1667
b  0 -1  0.75 0.5000
c  0  0 -1.00 0.3333
d  1  0  0.00 -0.5000
e -1  0  0.00 -0.5000
pic$prediction
  I3  I1  I2  Root
a 0.0 0.25 0.1667 0.1071
b 0.0 0.75 0.5000 0.3214
c 0.0 0.00 0.3333 0.2143
d 0.5 0.00 0.0000 0.1786
e 0.5 0.00 0.0000 0.1786
pic$variance
  I3  I1  I2  Root
0.200 0.400 0.975 1.167
```

On retiendra que l'on peut expliciter la matrice des contrastes \mathbf{U} . Elle est de dimension $n \times (n-1)$ et le nom d'un contraste correspond au nœud auquel il est associé. Pour une phylogénie résolue, c'est-à-dire sans nœud polytomique, on a exactement le même nombre de contrastes que le nombre de nœuds. La matrice des contrastes assure la définition au niveau des nœuds de variables aléatoires indépendantes sous l'hypothèse du mouvement brownien.

On peut donc représenter, pour un trait donné, les scores des contrastes $\mathbf{y} = \mathbf{U}' \mathbf{x}$ au niveau de chaque nœud de l'arbre (Figure 3.7). Sous l'hypothèse du mouvement brownien, les scores sont les réalisations d'une variable aléatoire $\mathbf{Y} \sim N(0\mathbf{1}_{n-1}, \sigma^2 \mathbf{Id}_{n-1})$.

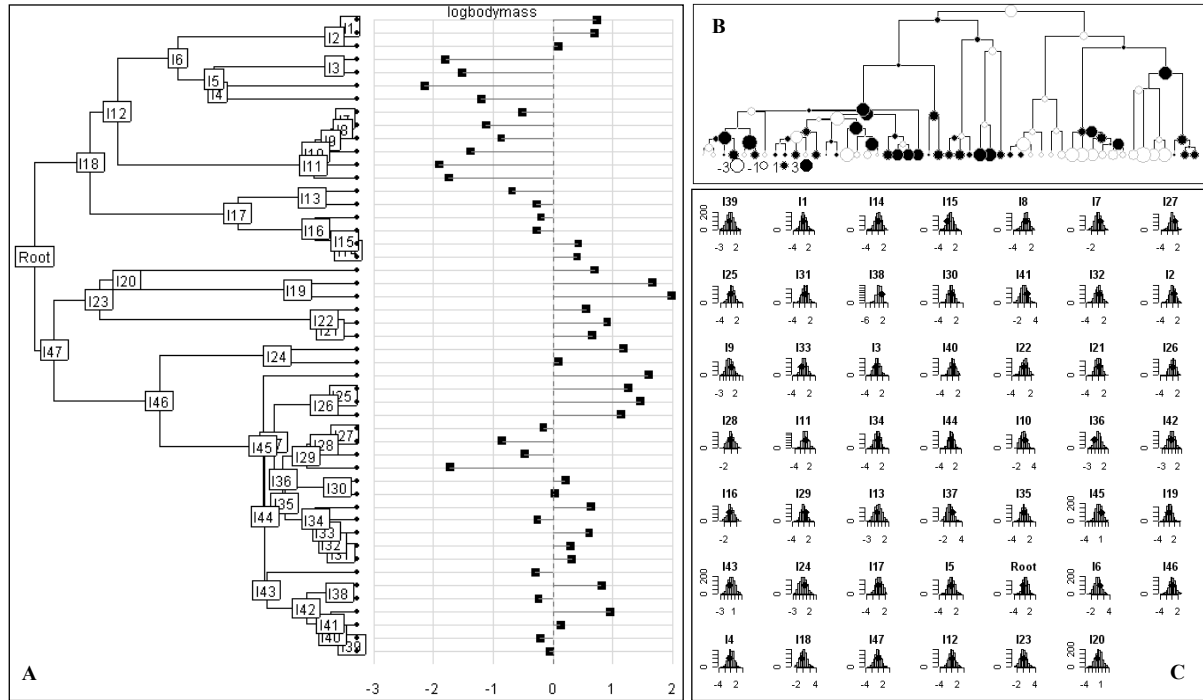


Figure 3.7 : **A.** Représentation de la variable poids du corps pour 49 espèces de mammifères (Annexe 2.4). **B.** Représentation simultanée des valeurs de la variable et des scores des contrastes. **C.** Distributions d'échantillonnage des scores des contrastes sous l'hypothèse du mouvement brownien. Les scores observés pour la variable considérée sont représentés sur chaque histogramme.

4.2. La métrique phylogénétique

On aurait pu donner une tout autre définition de la matrice des contrastes \mathbf{U} , beaucoup plus générale, simplement en remarquant que l'on cherche $n-1$ contrastes \mathbf{u} rangés en colonnes dans une matrice de contrastes \mathbf{U} définissant une nouvelle variable aléatoire multidimensionnelle :

$$\mathbf{Y} = (Y_1, Y_2, \dots, Y_{n-1})^t = \mathbf{U}^t \mathbf{X} \text{ avec } \mathbf{X} = (X_1, X_2, \dots, X_n)^t \sim N(\mu \mathbf{1}_n, \sigma^2 \mathbf{W}).$$

On peut montrer que $\text{var}(\mathbf{Y}) = \sigma^2 \mathbf{U}^t \mathbf{W} \mathbf{U}$, c'est-à-dire que les variables aléatoires associées aux contrastes \mathbf{U} sont indépendantes si les vecteurs \mathbf{u} forment une famille \mathbf{W} -orthogonale. Pour remplacer les variables X_i covariantes par des combinaisons linéaires Y_i qui ne le sont plus, il suffit de prendre une base orthonormée de \mathbb{R}^n au sens de \mathbf{W} (une base qui vérifie $\mathbf{U}^t \mathbf{W} \mathbf{U} = \mathbf{Id}_n$). Il existe une infinité de solutions dont la matrice des contrastes proposée par Felsenstein (1985). On peut en effet vérifier numériquement que la base obtenue algorithmiquement selon la procédure de Felsenstein (1985) forme bien une base orthogonale au sens de \mathbf{W} :

```
tre <- "((a:0.3,b:0.1)I1:0.25,c:0.65)I2:0.2,(d:0.1,e:0.1)I3:0.7)Root;"
phy <- newick2phylog(tre)
```

```
pic <- phylog2pic(phy)
U <- as.matrix(pic$contrastes)
U <- t(t(U)/sqrt(pic$variance))
round(t(U)%*%phy$Wmat%*%U, 4)
```

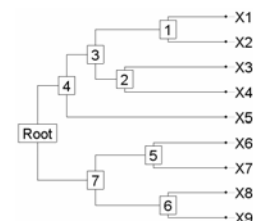
	I3	I1	I2	Root
I3	1	0	0	0
I1	0	1	0	0
I2	0	0	1	0
Root	0	0	0	1

Toutefois, les contrastes de Felsenstein (1985) perdent cette propriété (Rohlf, 2001) dès que la phylogénie contient d'autres bifurcations que des dichotomies. De même, si l'on change de modèle d'évolution, on obtient une autre matrice de variance-covariance définie en fonction de \mathbf{W} (Hansen & Martins, 1996), et le calcul des scores associés aux contrastes par l'approche algorithmique n'a plus de sens alors que l'on peut toujours obtenir des contrastes indépendants en respectant cette propriété générale.

La matrice \mathbf{W} , prend donc le statut particulier de matrice de variance-covariance d'un modèle multidimensionnel normal. Elle contient hors diagonale les covariances attendues entre les *OTUs* dans le modèle de divergence associée à une marche au hasard en temps continu. Elle est à la base des méthodes PIC (*Phylogenetic independent contrasts*) et PGLS (*phylogenetic generalized least-squares*) comparées dans Rohlf(2001) qui l'appelle Σ et indique qu'elle s'appelle \mathbf{B} dans Martins et Hansen (1997) et \mathbf{C} dans Garland et Ives (2000). On va voir dans ce paragraphe que son intérêt est bien plus général dans la mesure où c'est la matrice d'un produit scalaire associé à la distance phylogénétique, ainsi qu'une matrice définissant des proximités phylogénétiques entre feuilles.

En effet, on peut se passer radicalement du modèle théorique associé à \mathbf{W} pour donner à cette matrice un statut de produit scalaire dans \mathbb{R}^n . Pour les illustrations on peut utiliser l'exemple fictif de Martins et Hansen (1997) :

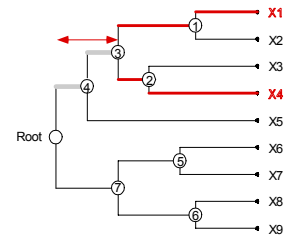
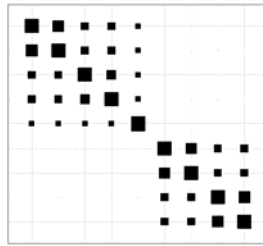
```
marthans.tre <- newick.eg[[14]]
marthans.phy <- newick2phylog(marthans.tre)
plot.phylog(marthans.phy ,
  labels.nodes = c(as.character(1:7), "Root"),
  clabel.nodes = 2, clabel.leaves = 2, f = 0.8)
```



Chaque couple de feuille (i, j) définit un premier ancêtre commun qui a une distance à la racine h_{ij} . Soit \mathbf{W} la matrice $n \times n$ des distances h_{ij} à la racine du premier ancêtre commun. On rappelle que les termes de la diagonale sont alors définis par les distances à la racine de chaque feuille :

```
marthans.phy$Wmat
table.value(marthans.phy$Wmat, clabel.row = 0, clabel.col = 0, cleg = 0)
```

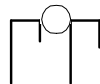
	X1	X2	X3	X4	X5	X6	X7	X8	X9
X1	13	9	4	4	2	0	0	0	0
X2	9	13	4	4	2	0	0	0	0
X3	4	4	13	6	2	0	0	0	0
X4	4	4	6	13	2	0	0	0	0
X5	2	2	2	2	13	0	0	0	0
X6	0	0	0	0	0	13	8	4	4
X7	0	0	0	0	0	8	13	4	4
X8	0	0	0	0	0	4	4	13	9
X9	0	0	0	0	0	4	4	9	13



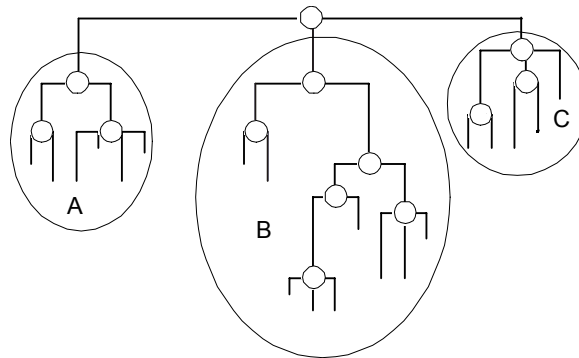
Cette matrice n'a que des valeurs propres strictement positives :

```
eigen(marthans.phy$Wmat)$values
[1] 29.60 29.52 13.48 12.60 11.81 7.00 5.00 4.00 4.00
```

Cette propriété est très générale. Elle est vraie pour un arbre élémentaire du type :



Dans ce cas W est diagonale à valeurs positives sur la diagonale. Si elle est vraie pour plusieurs arbres (A, B et C), elle est également vraie pour l'arbre formés par ces trois sous-arbres :



En effet, la matrice W est obtenue par assemblage des matrices associées aux sous-arbres réunis à ce niveau. Les distances à la racine sont toutes augmentées, dans un sous-arbre, de la même quantité, à savoir la longueur de la branche connectant ce sous-arbre, par exemple ($\mathbf{1}_{mm}$ désigne la matrice carrée $m \times m$ ne contenant que des 1) :

$$\mathbf{W} = \begin{bmatrix} \boxed{\mathbf{W}_A + d_A \mathbf{1}_{f_A f_A}} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boxed{\mathbf{W}_B + d_B \mathbf{1}_{f_B f_B}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boxed{\mathbf{W}_C + d_C \mathbf{1}_{f_C f_C}} \end{bmatrix}$$

Les valeurs propres de \mathbf{W} sont celles des matrices du type $\mathbf{W}_A + d_A \mathbf{1}_{f_A f_A}$. Or :

$$(\mathbf{W}_A + d_A \mathbf{1}_{f_A f_A}) \mathbf{u} = \lambda \mathbf{u} \Rightarrow \mathbf{W}_A \mathbf{u} + d_A (\mathbf{1}_{f_A}^t \mathbf{u}) \mathbf{1}_{f_A} = \lambda \mathbf{u}$$

donc :

$$\mathbf{u}^t \mathbf{W}_A \mathbf{u} + d_A (\mathbf{1}_{f_A}^t \mathbf{u}) \mathbf{u}^t \mathbf{1}_{f_A} = \lambda \mathbf{u}^t \mathbf{u} \Rightarrow \lambda \|\mathbf{u}\|^2 = \mathbf{u}^t \mathbf{W}_A \mathbf{u} + d_A (\mathbf{1}_{f_A}^t \mathbf{u})^2$$

Si les valeurs propres de \mathbf{W}_A sont toutes positives, la matrice \mathbf{W}_A est positive et les valeurs propres de $\mathbf{W}_A + d_A \mathbf{1}_{f_A f_A}$ sont positives, donc la propriété est vraie pour \mathbf{W} . La matrice \mathbf{W} est donc celle d'un produit scalaire.

La première conséquence est que la distance entre deux feuilles devient :

$$\delta_{ij}^2 = \|m_i - m_j\|_{\mathbf{W}}^2 = \|m_i\|^2 + \|m_j\|^2 - 2 \langle m_i | m_j \rangle = w_{ii} + w_{jj} - 2w_{ij}$$

\mathbf{W} est symétrique, définie et positive. C'est une métrique euclidienne de matrice \mathbf{W} dans la base canonique. Il est donc logique de prendre pour distance nodale :

$$d'_{ij} = \|e_i - e_j\|_{\mathbf{W}} = \sqrt{\delta_{ij}} = \sqrt{w_{ii} + w_{jj} - 2w_{ij}}$$

Or deux feuilles i et j ont un premier ancêtre commun k . La longueur du plus court chemin menant de i à j est la distance nodale entre i et j classiquement utilisée dans la littérature. Cette quantité s'écrit :

$$d_{ij} = w_{ii} + w_{jj} - 2w_{ij} = d_{ij}'^2$$

Il suffit d'ajouter et de retrancher deux fois la distance de l'ancêtre commun à la racine.

```
marthans.phy$Wdist # D'
      x1      x2      x3      x4      x5      x6      x7      x8
x2 2.828
x3 4.243 4.243
x4 4.243 4.243 3.742
x5 4.690 4.690 4.690 4.690
x6 5.099 5.099 5.099 5.099 5.099
x7 5.099 5.099 5.099 5.099 5.099 3.162
x8 5.099 5.099 5.099 5.099 5.099 4.243 4.243
x9 5.099 5.099 5.099 5.099 5.099 4.243 4.243 2.828
```

```
marthans.phy$Wdist**2 # D
      x1 x2 x3 x4 x5 x6 x7 x8
```

```

X2 8
X3 18 18
X4 18 18 14
X5 22 22 22 22
X6 26 26 26 26 26
X7 26 26 26 26 26 10
X8 26 26 26 26 26 18 18
X9 26 26 26 26 26 18 18 8

```

La distance ordinairement utilisée est donc un carré de distance euclidienne et la mesure de diversité de Clarke et Warwick (1999) est un cas particulier de la mesure de diversité de Champely et Chessel (2002). On peut donc définir une méthode d'ordination canoniquement associée, qui permet d'introduire dans la quantification de la différence entre deux sites de recensement des mesures dépendantes des différences taxonomiques ou phylogénétiques (c'est un des objectifs de la thèse de Sandrine Pavoine actuellement en cours).

La seconde conséquence est que la phylogénie définit un produit scalaire dans l'ensemble des variables mesurées sur les feuilles de l'arbre. C'est exactement ce que fait un graphe de voisinages entre sites sur l'ensemble des variables mesurées dans ces sites (Thioulouse et al., 1995). Graphes de voisinage, distances euclidiennes et phylogénies s'introduisent donc dans l'analyse des données comme métrique euclidienne. Les vecteurs propres de voisinage (*ibidem*) donnent pour les graphes de voisinage des composantes cartographiables. Les vecteurs propres de phylogénie font de même (Figure 3.8) :

```

u <- gridrowcol(3,3)      # exemple spatial
par(mfrow = c(3,3))
s.label(u$xy, neig = u$neig,
       cneig = 2, grid = FALSE,
       inc = FALSE, clab = 2,
       label = paste("X", 1:9, sep = ""))
for(i in 1:8) s.value(u$xy, u$orthobasis[,i],
                    neig = u$neig, cleg = 0,
                    cneig = 2, grid = FALSE,
                    inc = FALSE, csi = 1.5)
table.value(neig2mat(u$neig))

u <- marthans.phy        # exemple phylogénétique
table.phylog(u$Wscores, u, cleg = 0,
            clabel.row = 0.75, clabel.col = 0, csi = 1.4)
table.value(u$Wmat, row.labels = paste("X", 1:9, sep = ""),
            col.labels = paste("X", 1:9, sep = ""), cleg = 0)

```

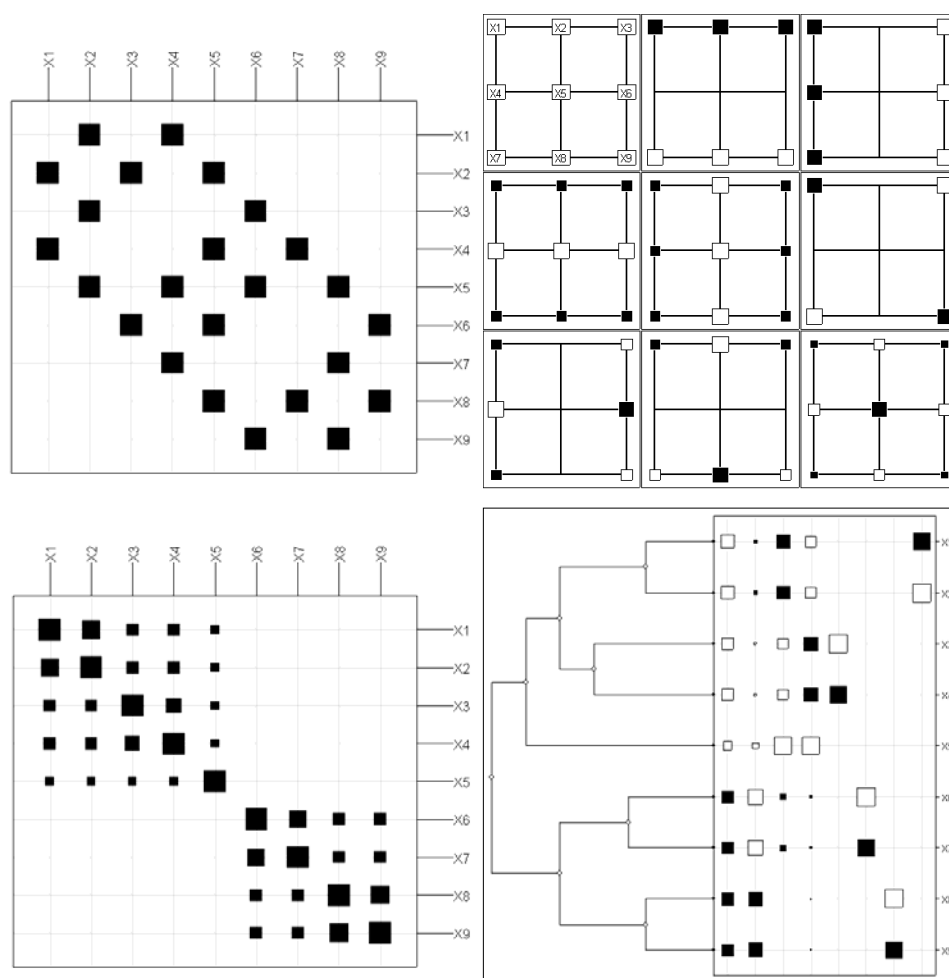


Figure 3.8 : A gauche, matrices de proximité (**M** pour le graphe de voisinage, **W** pour la phylogénie). A droite, figures de référence associées aux matrices de proximité.

Il a été observé très tôt la correspondance existante entre la question des phylogénies et la question des structures spatiales. Pour la petite anecdote, lorsque Louis XV confia à son botaniste, Bernard de Jussieu, le soin d'organiser pour le Trianon de Versailles un jardin botanique qui soit le reflet de la classification naturelle, ce dernier eut une idée de génie. Il tenta d'élaborer une carte du jardin, ce qui deviendra par la suite le « *Système de Trianon* », publié en 1789 par Antoine Laurent (Le Guyader, 2003). Chaque espèce y était représentée par une petite surface. Souhaitant placer côte à côte les espèces qui se ressemblent le plus, il réalisa une surface plus grande, un petit bosquet qui représentait le genre. Les différents genres étaient alors réunis en une parcelle de plus grande taille, appelée famille... Les proximités entre espèces, genres familles, c'est-à-dire les proximités taxonomiques étaient donc matérialisées, selon le « *Système de Trianon* », par des proximités spatiales. D'un point de vue statistique, le lien entre méthodes d'étude des structures spatiales et des structures phylogénétiques s'est fait bien plus tardivement. Il est introduit par l'intermédiaire des

modèles auto-régressifs par Cheverud et Dow (1985) puis complété par Gittleman et Kot (1990). Dans ce dernier cas, la matrice \mathbf{W} est introduite comme matrice de proximité dans le calcul de l'indice de Moran et du corrélogramme.

On retiendra de ce paragraphe qu'à une phylogénie est canoniquement associé un produit scalaire défini par la matrice \mathbf{W} , dont le terme général correspond à la longueur des branches aux premiers ancêtres communs. Cette matrice \mathbf{W} est la matrice des variances covariances sous l'hypothèse d'un modèle évolutif brownien. La racine de la longueur du chemin le plus court associé à un couple de feuilles définit une matrice de dissimilarités \mathbf{D} euclidienne. \mathbf{D} et \mathbf{W} ont donc un intérêt indépendamment de tout modèle évolutif.

4.3. Usage de la méthode des contrastes

Une fois la méthode des contrastes définie, on peut s'interroger sur son usage. Pour répondre, il suffit d'une expérience extrêmement simple.

Considérons un échantillon aléatoire simple d'une loi normale, c'est-à-dire une variable aléatoire $\mathbf{X} = (X_1, X_2, \dots, X_n)$ qui suit une loi normale multivariée de paramètres :

$$E(\mathbf{X}) = \mu \mathbf{1}_n^t \quad \text{Var}(\mathbf{X}) = \sigma^2 \mathbf{I}_n$$

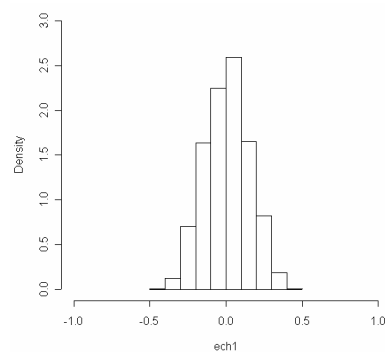
et une variable $\mathbf{Y} = (Y_1, Y_2, \dots, Y_n)$ qui suit une loi normale multivariée de paramètres :

$$E(\mathbf{Y}) = \nu \mathbf{1}_n^t \quad \text{Var}(\mathbf{Y}) = \tau^2 \mathbf{I}_n$$

On connaît la distribution d'échantillonnage théorique de la statistique $\text{corr}(\mathbf{X}, \mathbf{Y})$ mais pour l'étudier, on génère simplement un échantillon de 1000 tirages de paramètres $\mu = \nu = 0, \mathbf{1}_n^t \sigma^2 = \tau^2 = 1$. On donne l'histogramme de la distribution d'échantillonnage obtenue ainsi que les principaux quantiles :

```
fun <- function(x) {
  u <- rnorm(98)
  u <- matrix(u, 49, 2)
  u <- cor(u)[1,2]
  return(u)
}

ech1 <- unlist(lapply(1:1000, fun))
hist(ech1, proba = T, main = "",
      xlim = c(-1, 1), ylim = c(0, 3))
quantile(ech1, c(0.01, 0.05, 0.95, 0.99))
1%      5%      95%     99%
-0.3123 -0.2374  0.2405  0.3209
```

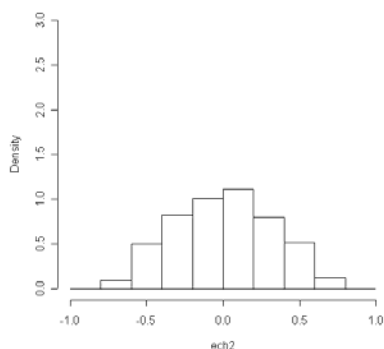


On refait exactement la même chose pour deux échantillons indépendants, évoluant selon le modèle d'évolution brownien : $\text{Var}(\mathbf{X}) = \sigma^2 \mathbf{W}$ et $\text{Var}(\mathbf{Y}) = \tau^2 \mathbf{W}$.

```
phy <- newick2phylog(carniherbi49$tre2)

fun <- function(x){
u <- rmvnorm(2, rep(0, 49), phy$Wmat)
u <- t(u)
u <- cor(u)[1,2]
return(u)
}

ech2 <- unlist(lapply(1:1000, fun))
hist(ech2, proba = T, main = "",
      xlim = c(-1, 1), ylim = c(0,3))
quantile(ech2, c(0.01, 0.05, 0.95, 0.99))
      1%      5%      95%      99%
-0.6717 -0.5271  0.5515  0.6666
```

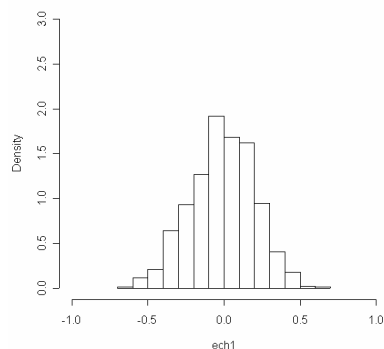


L'autocorrélation induite par la phylogénie est sans conteste une nuisance importante. Dans le deuxième cas, une corrélation de 0.5 (positive ou négative) n'est pas significative, alors que dans le premier cas elle apparaît comme tout à fait extraordinaire. L'autocorrélation phylogénétique perturbe fortement l'analyse statistique de la corrélation entre les traits, ce qui confirme ce que l'on avait déjà souligné à travers la Figure 3.5 :. Les auteurs préconisent de prendre la corrélation entre les scores des contrastes pour s'affranchir de la nuisance induite par l'autocorrélation phylogénétique. Dans quelle mesure, cette pratique est-elle pertinente ? Afin de répondre à cette question, on étudie la distribution d'échantillonnage de la corrélation entre les scores des contrastes $\text{corr}(\mathbf{U}'\mathbf{X}, \mathbf{U}'\mathbf{Y})$, pour les deux situations considérées :

```
# échantillons aléatoires simples simulés selon le modèle Gaussien classique

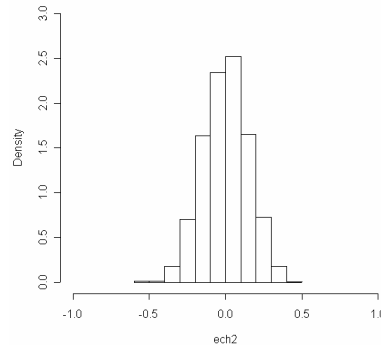
pic <- phylog2pic(phy)
fun <- function(x){
u <- rnorm(98)
u <- matrix(u, 49, 2)
u <- t(pic$contrast) %*% u
u <- u/sqrt(pic$variance)
u <- cor(u)[1,2]
return(u)
}

ech1 <- unlist(lapply(1:1000, fun))
hist(ech1, proba = T, main = "",
      xlim = c(-1, 1), ylim = c(0,3))
quantile(ech1, c(0.01, 0.05, 0.95, 0.99))
      1%      5%      95%      99%
-0.5214 -0.3788  0.3183  0.4448
```



```
# échantillons aléatoires simples simulés selon le modèle Gaussien brownien
fun <- function (x){
u <- rmvnorm(2, rep(0, 49), phy$Wmat)
u <- t(u)
u <- t(pic$contrastes)*%u
u <- u/sqrt(pic$variance)
u <- cor(u)[1,2]
return(u)
}

ech2 <- unlist(lapply(1:1000, fun))
hist(ech2, proba = T, main = "",
      xlim = c(-1, 1), ylim = c(0,3))
quantile(ech2,c(0.01,0.05,0.95,0.99))
  1%      5%      95%      99%
-0.3266 -0.2477  0.2408  0.3303
```



L'illustration est parlante. Quand le modèle brownien est vrai, la correction par les contrastes est pertinente dans la mesure où la corrélation entre les traits diminue et sa distribution correspond bien à celle de deux traits indépendants. Par contre, sur des variables normales, l'application des contrastes est nuisible puisqu'elle induit une corrélation artificielle entre les deux variables. Un bon nombre d'articles utilisant la méthode des contrastes comme une simple recette de cuisine, interprète des corrélations qui sont certainement induites par la procédure elle-même. Avant d'appliquer un modèle, il faut en avoir vérifié les hypothèses. Une recette de cuisine n'est réussie que dans la mesure où l'on mélange les bons ingrédients...

5. LE TEST D'ABOUHEIF (1999)

Ainsi, si l'on suspecte qu'une variable est liée à la phylogénie, on peut employer les corrections phylogénétiques en calculant les scores des contrastes car cette pratique modifie toutes les estimations. Sinon, il ne faut pas les employer. Tester si une variable est liée à la phylogénie constitue donc une étape préliminaire à toute analyse comparative. La publication récente de Blomberg et al. (2003), donne une revue assez complète des tests proposés dans la littérature. Le test d'Abouheif (1999) est un des tests les plus utilisés. Il ne prend en compte que la topologie des arbres phylogénétiques, les longueurs de branches n'étant pas prises en considération. De plus, le test d'Abouheif est le seul à être défini pour une variable quantitative comme pour une variable qualitative. Il est donc très général, particulièrement simple, ce qui en fait un excellent outil pour une première phase exploratoire. Il a été introduit de manière très intuitive, en adaptant deux tests non paramétriques contre l'absence de dépendance sérielle. L'idée est intéressante mais la mise en œuvre très empirique. C'est un

bel exemple d'une pratique de la statistique *ad hoc*. On propose une relecture critique du test d'Abouheif, à partir de laquelle on définit une procédure canonique.

5.1. Principe du test d'Abouheif

Le problème consiste à donner une mesure de la ressemblance entre espèces voisines pour un trait considéré, le voisinage entre espèces étant introduit par la phylogénie. Abouheif (1999) s'est inspiré de deux statistiques introduites dans un contexte non phylogénétique pour tester l'absence de dépendance sérielle dans une série univariée :

- pour les variables quantitatives, il utilise la statistique S_1 associée au test de von Neumann et al. (1941), basée sur la somme des carrés des différences pour tous les couples de mesures successives.
- pour les variables qualitatives, il utilise la statistique S_2 associée au 'run' test (Sokal & Rohlf, 1969), basée sur le nombre de séquences homogènes rencontrées dans une série à deux ou plusieurs modalités.

Pour introduire le voisinage entre espèces, il part de la remarque fondamentale suivante : la structure de données "phylogénie" est du point de vue graphique très particulière, puisqu'elle admet une multitude de représentations dont aucune n'est canonique. En effet, toute phylogénie peut être représentée de P manières différentes, suite à la permutation des branches au niveau des nœuds (Figure 3.9).

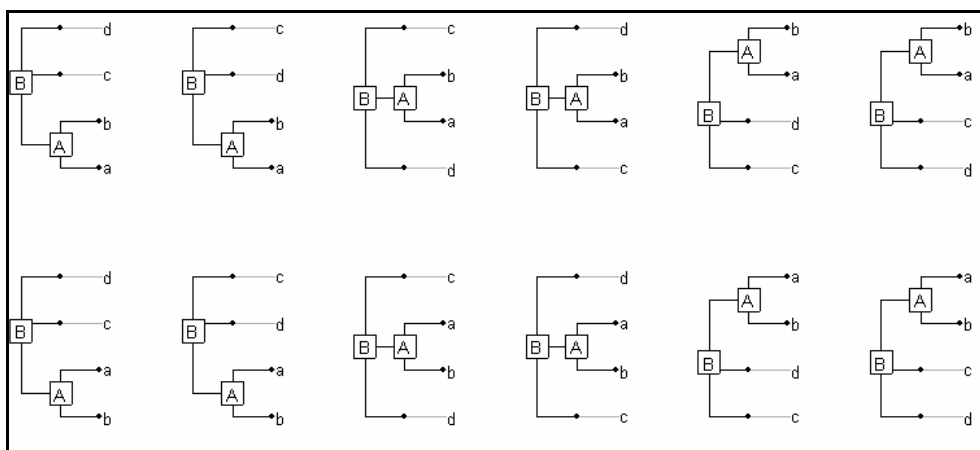


Figure 3.9 : Ensemble des représentations graphiques d'un arbre phylogénétique ($P = 12$)

Chaque permutation sélectionne une permutation des n feuilles, modifiant l'ordre des données. Les statistiques S_1 associée au test de von Neumann et al. (1941) et S_2 associée au

‘run’ test(Sokal & Rohlf, 1969) prennent donc des valeurs différentes suivant la représentation considérée (Figure 3.10). Abouheif propose alors de prendre la moyenne des valeurs prises par les statistiques S_1 et S_2 pour l’ensemble des P représentations possibles du même arbre. L’idée est excellente, mais la mise en œuvre un peu moins. En effet, dès que le nombre de feuilles devient un peu grand, le nombre de représentations possibles devient vite très important ce qui rend impossible le calcul explicite de toutes les valeurs prises par chacune des deux statistiques. Abouheif considère alors une solution approchée et fait le calcul à partir d’un échantillon de 1000 représentations prises au hasard parmi l’ensemble des P représentations possibles. D’un point de vue calculatoire, la solution est satisfaisante car la convergence est assez rapide. D’un point de vue théorique, elle n’est pas justifiée car le calcul peut se faire indépendamment de la détermination de toutes les représentations. En effet, ce qu’Abouheif ignore, c’est qu’il calcule sans le savoir une statistique de Moran et définit implicitement une nouvelle matrice de proximité **A** entre les feuilles qui possède des propriétés statistiques intéressantes.

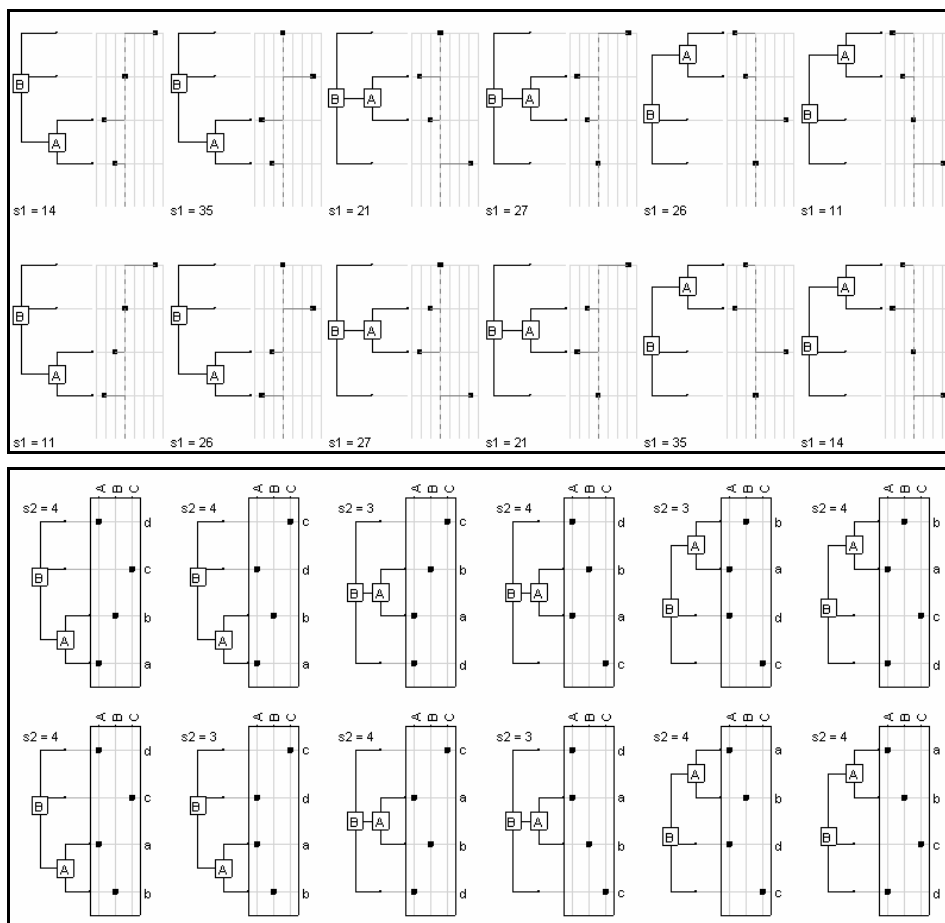


Figure 3.10 : En haut, les 12 représentations possibles des valeurs d’une variable quantitative ($x \leftarrow c(-1, -2, 0, 3)$) en face de la phylogénie. La statistique S_1 prend des valeurs différentes suivant la représentation. En bas, les 12 représentations possibles du tableau disjonctif complet d’une variable qualitative ($x \leftarrow c("A", "B", "C", "d")$). La statistique S_2 prend des valeurs différentes suivant la représentation.

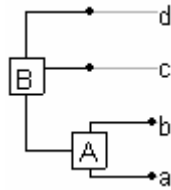
5.2. Le cas d'une variable quantitative

La statistique proposée par Abouheif pour une variable quantitative x est définie par :

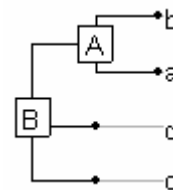
$$C_{mean} = \frac{1}{P} \sum_p \left(1 - \frac{\eta_p}{2} \right) \text{ avec } \eta_p = \frac{S_{1p}}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n-1} (x_{\tau_p(i)} - x_{\tau_p(i+1)})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

où l'indice τ_p est celui de l'ordre des feuilles pour une des P représentations de la phylogénie.

Par exemple, pour les deux représentations suivantes, on a :



$$\tau_1^t = (1, 2, 3, 4)$$

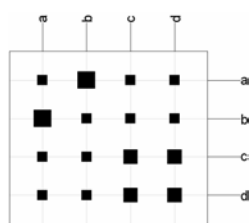


$$\tau_5^t = (3, 4, 1, 2)$$

La statistique C_{mean} peut se réécrire sous la forme d'une statistique de Moran qui est étroitement liée à une statistique de Geary. En effet, elle s'écrit sous la forme

$$C_{mean} = 1 - \frac{\frac{1}{P} \sum_p S_{1p}}{2 \sum_i (x_i - \bar{x})^2} = 1 - \frac{\sum_i \sum_j a_{ij} (x_i - x_j)}{2 \sum_i (x_i - \bar{x})^2}$$

où a_{ij} est le terme général d'une matrice de proximité \mathbf{A} défini comme la fréquence des représentations qui placent la feuille i juste avant la feuille j sur l'arbre phylogénétique. Plus ce terme est grand, plus les feuilles sont proches du point de vue évolutif. Cette matrice est symétrique, elle a une marge uniforme et sa somme égale n . Elle vaut, pour l'exemple considéré (Figure 3.9) :



$$12 \times \mathbf{A} = \begin{bmatrix} 2 & 6 & 2 & 2 \\ 6 & 2 & 2 & 2 \\ 2 & 2 & 4 & 4 \\ 2 & 2 & 4 & 4 \end{bmatrix} \begin{matrix} a \\ b \\ c \\ d \end{matrix}$$

Sur les 12 représentations, la feuille a (1 ère ligne) est :

- 2 fois au dessous de toutes les autres
- 6 fois juste au dessus de b
- 2 fois juste au dessus de c

- 2 fois juste au dessus de d

On peut alors réécrire la statistique C_{mean} sous la forme matricielle :

$$C_{mean} = 1 - \frac{n \sum_i \sum_j a_{ij} (x_i - x_j)}{2 \sum_i \sum_j a_{ij} \sum_i (x_i - \bar{x})^2} = 1 - \frac{\mathbf{z}^t (\mathbf{N} - \mathbf{A}) \mathbf{z}}{\mathbf{1}_n^t \mathbf{A} \mathbf{1}_n}$$

avec $z_i = \frac{x_i - \bar{x}}{1/n \sum_i (x_i - \bar{x})^2}$ et $\mathbf{N} = \text{Diag}(\mathbf{A} \mathbf{1}_n) = \mathbf{I}d_n$

La matrice \mathbf{A} ayant pour marge la pondération uniforme, la statistique C_{mean} est exactement une statistique de Moran :

$$C_{mean} = 1 - \mathbf{z}^t \left(\frac{\mathbf{N} - \mathbf{A}}{\mathbf{1}_n^t \mathbf{A} \mathbf{1}_n} \right) \mathbf{z} = 1 - \mathbf{z}^t \left(\frac{\mathbf{I}d_n}{n} - \frac{\mathbf{A}}{\mathbf{1}_n^t \mathbf{A} \mathbf{1}_n} \right) \mathbf{z} = \mathbf{z}^t \frac{\mathbf{A}}{\mathbf{1}_n^t \mathbf{A} \mathbf{1}_n} \mathbf{z}$$

Le test d' Abouheif (1999) pour une variable quantitative est donc un cas particulier du test de Moran utilisant une matrice de proximité \mathbf{A} induite par la phylogénie. On peut vérifier sur l'exemple de la Figure 3.10 (**en haut**) que les deux calculs 'selon la logique du test d'Abouheif et selon la logique du test de Moran) donnent bien la même chose :

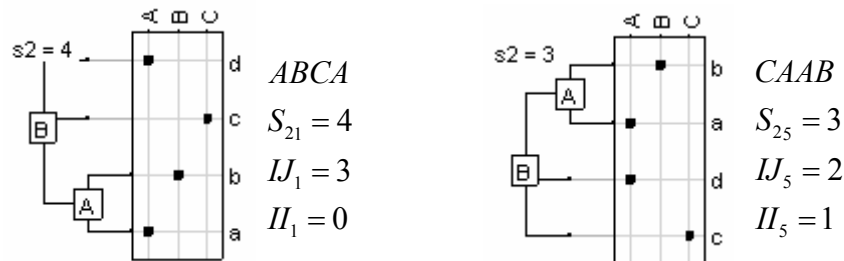
```
# calcul selon l'approche d'Abouheif
x <- c(-1,-2,0,3)
mean(x)
[1] 0
eta <- c(14, 35, 21, 27, 26, 11, 11, 26, 27, 21, 35, 14)
eta <- eta/sum(x**2)
Cmean <- mean(1-eta/2)
Cmean
[1] 0.2024

# calcul par l'indice de Moran
A <- matrix(c(2,6,2,2,6,2,2,2,2,2,4,4,2,2,4,4), nrow = 4)/12
A
      [,1] [,2] [,3] [,4]
[1,] 0.1667 0.5000 0.1667 0.1667
[2,] 0.5000 0.1667 0.1667 0.1667
[3,] 0.1667 0.1667 0.3333 0.3333
[4,] 0.1667 0.1667 0.3333 0.3333
gearymoran(A, as.data.frame(x))
class: krandtest
test number: 1
permutation number: 999
  test obs  P(X<=obs) P(X>=obs)
1 x      0.202 1      0.235
```

Pour vérifier l'égalité des deux approches, on a implémenté la matrice \mathbf{A} à la main à partir de l'observation des P représentations. En fait, cette matrice \mathbf{A} peut avoir une expression analytique assez simple pour toutes les phylogénies, qu'elles soient résolues ou non.

5.3. Le cas d'une variable qualitative

La statistique S_{2p} proposée par Abouheif pour une variable qualitative \mathbf{x} est égale au nombre de séquences homogènes d'une même modalité pour une représentation p donnée. Elle est intimement liée à deux autres statistiques : le nombre de fois où la modalité change lorsque l'on passe d'une feuille à la suivante (statistique que l'on note IJ_p) et le nombre de fois où la modalité reste inchangée (statistique que l'on note II_p). En effet, on a tout d'abord pour chaque représentation p : $IJ_p + II_p = n-1$ et $IJ_p + 1 = S_{2p}$ donc $S_{2p} = n - II_p$. Par exemple, pour les deux représentations suivantes, on a :



Or la valeur de ces deux nouvelles statistiques s'exprime facilement sous la forme d'un produit matriciel intégrant le tableau disjonctif de la variable qualitative et la matrice de proximité \mathbf{A} . Les statistiques $(II_p)_{1 \leq p \leq P}$ et $(IJ_p)_{1 \leq p \leq P}$ s'avèrent être l'équivalent des statistiques de Moran et Geary pour une variable qualitative. En effet, si l'on note \mathbf{X} le tableau disjonctif associé à la variable \mathbf{x} , on a :

$$tr(\mathbf{X}'\mathbf{A}\mathbf{X}) = 1 + \frac{1}{P} \sum_p II_p \quad \text{et} \quad tr\left(\mathbf{X}'\left(\frac{\mathbf{Id}_n}{n} - \mathbf{A}\right)\mathbf{X}\right) = 1 - n + \frac{1}{P} \sum_p IJ_p$$

On retrouve donc l'égalité précédente, cette fois-ci avec une variable qualitative :

$$tr\left(\mathbf{X}'\left(\frac{\mathbf{Id}_n}{n} - \mathbf{A}\right)\mathbf{X}\right) + tr(\mathbf{X}'\mathbf{A}\mathbf{X}) = 1.$$

Toutefois, afin d'être cohérent avec la définition de l'indice de Moran défini pour une variable quantitative, il faut tenir compte de la normalisation et de la pondération des colonnes qui n'est pas uniforme pour une variable qualitative. En effet, on définit ici un test global en ajoutant la contribution de chaque indicatrice de classes, c'est-à-dire chaque colonne du tableau disjonctif. Dans le cas d'une variable quantitative, on n'a pas ce problème là, puisque l'on a une seule variable dont le poids est 1. Il faut donc définir une pondération des colonnes

en intégrant la pondération \mathbf{Q} et la transformation $\mathbf{X} \rightarrow \mathbf{X}_c$ associée à l'analyse des correspondances multiples. On a toujours :

$$\text{tr}\left(\mathbf{Q}\mathbf{X}_c^t\left(\frac{\mathbf{Id}_n}{n} - \mathbf{A}\right)\mathbf{X}_c\right) + \text{tr}\left(\mathbf{Q}\mathbf{X}_c^t\mathbf{A}\mathbf{X}_c\right) = \text{tr}\left(\mathbf{Q}\mathbf{X}_c^t\left(\frac{\mathbf{Id}_n}{n}\right)\mathbf{X}_c\right) = 1.$$

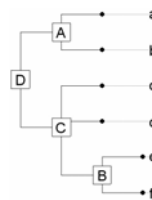
Je doute qu'Abouheif ait fait ce choix là car il est impossible de savoir ce qu'il a choisi, son programme n'étant pas accessible et son article ne le précisant pas. On sait juste que sa statistique est liée à la statistique S_2 . Celle que l'on vient de définir par $\text{tr}\left(\mathbf{Q}\mathbf{X}_c^t\mathbf{A}\mathbf{X}_c\right)$ est très générale. Elle généralise le point de vue d'Abouheif à une variable qualitative ainsi qu'à un triplet statistique de manière canonique et non empirique. Par la même, on donne une solution au test multivarié contre l'absence de structure phylogénétique. Pour cela, il suffit d'avoir l'expression analytique de la matrice de proximité \mathbf{A} .

5.4. La matrice de proximité \mathbf{A}

Cette matrice est une matrice de proximité définissant des proximités entre les couples de feuilles portées par un arbre phylogénétique. Pour chaque couple, le terme général a_{ij} est défini comme la fréquence des représentations compatibles avec la phylogénie qui placent les deux feuilles i et j l'une en dessus de l'autre. Pour les termes de la diagonale a_{ii} , la proximité des feuilles est alors définie comme la fréquence des représentations compatibles qui placent la feuille i au dessous de toutes les autres.

On considère l'arbre phylogénétique suivant

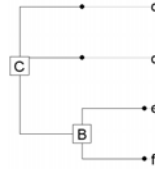
```
tre <- "((a,b)A,(c,d,(e,f)B)C)D;"
phy <- newick2phylog(tre)
plot(phy, clabel.nodes = 2,
      clabel.leaves = 2, cleaves = 1.5)
```



L'ensemble des feuilles est noté $L = \{a, b, c, d, e, f\}$. Le nombre total de feuilles est $n = \text{card}(L) = 6$. L'ensemble des nœuds est noté $N = \{A, B, C, D\}$. Le nombre total de nœuds est $p = \text{card}(N) = 4$. Le plus court chemin qui conduit d'une feuille à une autre permet de définir un ensemble ordonné de nœuds. Par exemple au chemin qui mène de d à b , noté (d, C, D, A, b) , est associé l'ensemble $P_{db} = \{C, D, A\}$. De même, le chemin qui conduit de la racine à une feuille définit l'ensemble ordonné des ancêtres associés à une feuille. Par

exemple à la feuille e est associé le chemin (e, B, C, D) et l'ensemble des nœuds $P_{ee} = \{B, C, D\}$. A chaque nœud est enraciné un sous-arbre qui permet de définir l'ensemble des descendants directs d'un nœud. Par exemple, au nœud C est associé l'ensemble des descendants directs $DD_C = \{c, d, B\}$ auquel correspond le sous-arbre :

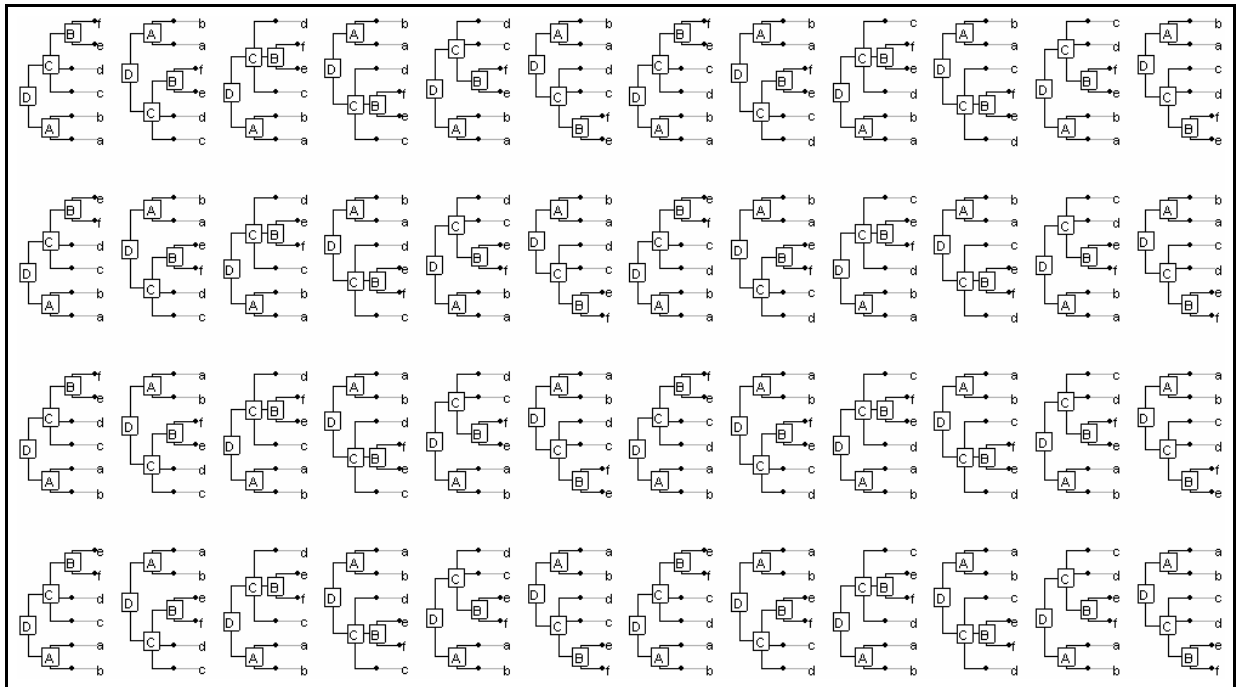
```
subtree <- phylog.extract(phy, "C")
subtree
[1] "(c:1,d:1,(e:1,f:1)B:1)C;"
subphy <- newick2phylog(subtree)
plot(subphy, clabel.nodes = 2,
      clabel.leaves = 2, cleaves = 1.5)
```



Le nombre de descendants direct d'un nœud est noté $dd_c = \text{card}(DD_C) = 3$. Il est directement impliqué dans le calcul du nombre de représentations compatibles avec la topologie de la phylogénie. Le nombre total de représentations P est défini par le produit du nombre de permutations possibles au niveau de chaque nœud, c'est-à-dire

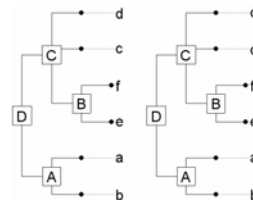
$P = \prod_{i \in N} dd_i! = 2!2!3!2! = 48$. On peut tracer les 48 présentations possibles :

```
enum <- enum.phylog(phy)
par(mfrow=c(4,12))
fun <- function(x) {
  plot(phy, x, clabel.nodes = 1.5,
       clabel.leaves = 1.5,
       cleaves = 1.5, f = 0.75)
}
apply(enum, 1, fun)
```



On peut désormais définir, parmi ces représentations compatibles, et pour chaque couple de feuilles (i, j) , le nombre de représentations qui placent 2 feuilles l'une juste au-dessus de l'autre. Par exemple, pour le couple (e, a) , il y a exactement deux représentations : (d, c, f, e, a, b) et (c, d, f, e, a, b) .

```
par(mfrow=c(1,2))
plot(phy, enum[29,], clabel.nodes = 2,
     f = 0.75, clabel.leaves = 2,
     cleaves = 1.5)
plot(phy, enum[35,], clabel.nodes = 2,
     f = 0.75, clabel.leaves = 2
     , cleaves = 1.5)
```



Les deux représentations possibles sont liées à la permutation au niveau du nœud C des feuilles c et d . Toutes les autres permutations entraînent la perte de la proximité des deux feuilles. C'est lié aux faits que le chemin (e, B, C, D, A, a) qui mène de la feuille e à la feuille a passe par tous les nœuds, et chaque nœud à l'exception de C est un nœud portant une dichotomie. On conçoit alors que le nombre de représentations va être lié au chemin qui relie deux feuilles ainsi qu'au nombre de descendants directs de chaque nœud. De plus, le nombre de représentations pour les couples (i, j) et (j, i) est identique car $P_{ij} = P_{ji}$, d'où la symétrie de la matrice d'Abouheif.

Pour un couple de feuilles on peut alors conjecturer le nombre de représentations qui placent les deux feuilles côte à côte avec i au dessus de j . Il est égal au nombre de

permutations qui laissent inchangée la position des deux feuilles. Ce nombre de permutations est lié d'une part aux permutations intervenant au niveau des nœuds qui n'appartiennent pas au chemin P_{ij} , d'autre part aux permutations intervenant au niveau des nœuds qui appartiennent au chemin P_{ij} et qui portent au moins 3 descendants directs. Pour les nœuds k qui n'appartiennent pas au chemin, le nombre de permutations possibles est $dd_k!$. Pour ceux qui appartiennent au chemin, le nombre de permutations est $(dd_k - 1)!$. Ainsi, le nombre de permutations qui laissent inchangées la position des deux feuilles est égal à $I = \prod_{k \in P_{ij}} (dd_k - 1)! \prod_{k \in N - P_{ij}} dd_k!$. La fréquence des représentations pour un couple (i, j) est donc définie par

$$a_{ij} = \frac{I_{ij}}{p} = \frac{\prod_{k \in P_{ij}} (dd_k - 1)! \prod_{k \in N - P_{ij}} dd_k!}{\prod_{k \in N} dd_k!} = \frac{1}{\prod_{k \in P_{ij}} dd_k}$$

C'est le terme général de la matrice de proximité **A**. Pour les termes a_{ii} de la diagonale, on retrouve bien la fréquence des représentations pour lesquelles la feuille i est en dessous de toutes les autres. Cette matrice est par définition symétrique et bi-stochastique (à pondération uniforme par ligne et par colonne). Ses termes sont tous strictement positifs. Les termes sont d'autant plus grands que les espèces sont proches d'un point de vue évolutif. La proximité entre espèces s'exprime selon une logique radicalement différente de la logique propre à la matrice **W** (Figure 3.11). L'utilisation de chacune des matrices, en particulier pour les tests statistiques tels que le test de Moran, risque donc de donner des résultats forts différents (Tableau 3.1). J'aurais tendance à dire que la matrice **A** est beaucoup plus raffinée que la matrice **W**. Par conséquent, son utilisation en analyse de données est parfaitement légitime et doit être préférée à **W**, surtout lorsque les longueurs de branches sont inconnues.

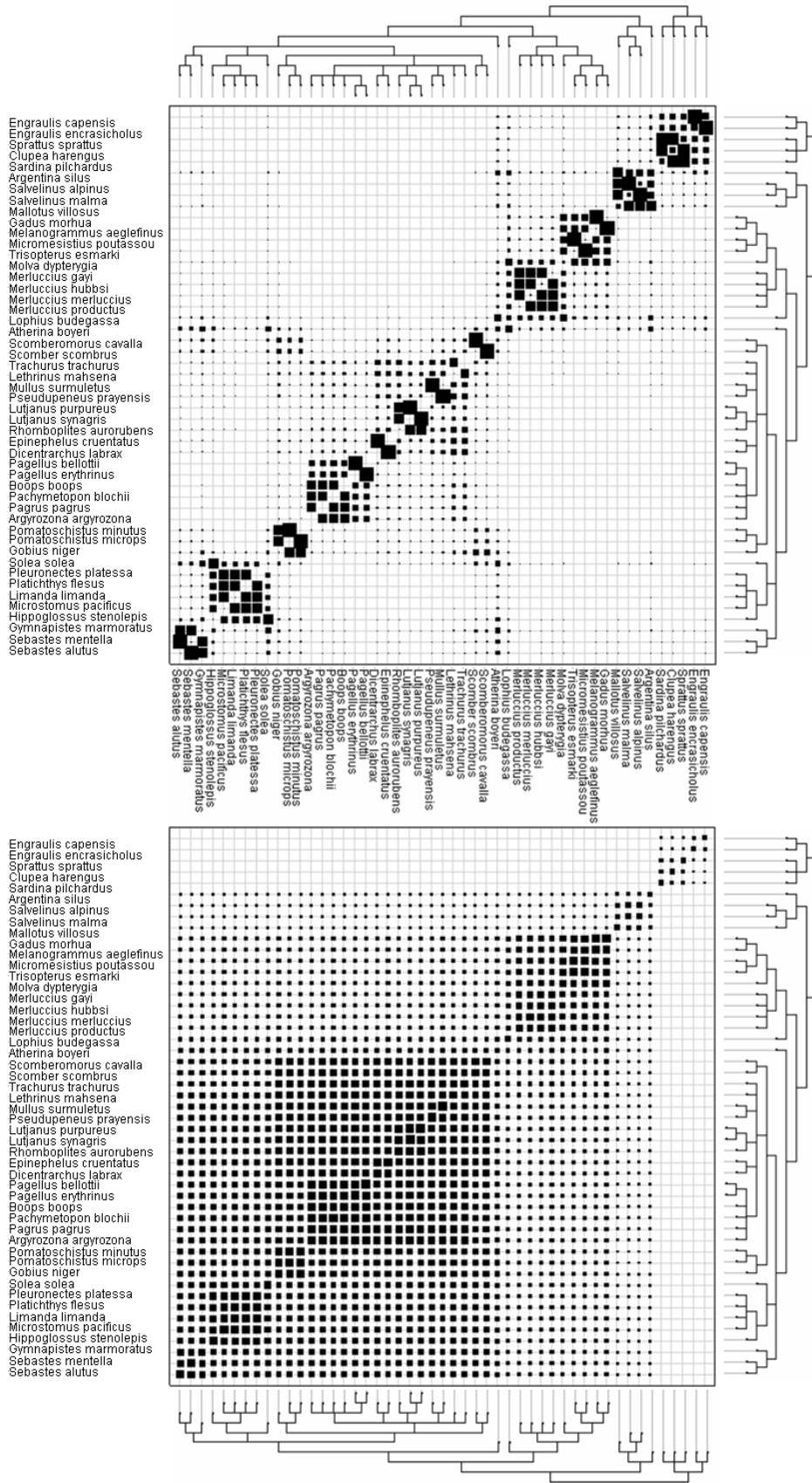


Figure 3.11 : Matrices de proximités associées à la phylogénie des poissons marins téléostéens (Annexe 1.14). **En haut**, matrice de proximité **A** associée au test d'Abouheif. **En bas**, matrice de proximité **W** associée au modèle brownien.

5.5. Conclusions

Nous venons de définir, à partir de l'idée d'Abouheif, un test multivarié contre l'absence de structure phylogénétique. Lorsque le tableau n'a qu'une variable quantitative, on retrouve exactement le test de Moran-Geary introduit au chapitre 1. Lorsque le tableau n'a qu'une variable qualitative, l'application du test de Moran-Geary sur le tableau disjonctif pondéré par les pondérations de l'analyse des correspondances multiples donne un test basé sur le nombre de suites homogènes d'une même modalité. C'est déjà un test global qui porte sur l'inertie. Lorsque le tableau est un mélange de variables qualitatives et quantitatives, on peut généraliser l'idée en définissant un test global sur l'inertie du schéma. Soit $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$ un triplet statistique. Soit \mathbf{A} la matrice de proximité associée à la phylogénie. Le test portera sur

la statistique $tr\left(\mathbf{QX}'_c\mathbf{D}^{\frac{1}{2}}\mathbf{A}\mathbf{D}^{\frac{1}{2}}\mathbf{X}_c\right)$. On retrouve alors la décomposition de l'inertie du schéma :

$$tr\left(\mathbf{QX}'_c\mathbf{D}^{\frac{1}{2}}\mathbf{A}\mathbf{D}^{\frac{1}{2}}\mathbf{X}_c\right) + tr\left(\mathbf{QX}'_c\mathbf{D}^{\frac{1}{2}}(\mathbf{Id}_n - \mathbf{A})\mathbf{D}^{\frac{1}{2}}\mathbf{X}_c\right) = tr\left(\mathbf{QX}'_c\mathbf{D}^{\frac{1}{2}}\mathbf{Id}_n\mathbf{D}^{\frac{1}{2}}\mathbf{X}_c\right).$$

Moran
Geary
Classique

L'analyse du quadruplet $(\mathbf{X}, \mathbf{Q}, \mathbf{D}, \mathbf{A})$ sous sa forme Geary ou Moran définit alors une analyse sous contrainte phylogénétique. Le problème a été posé explicitement dans plusieurs articles (Ackerly & Reich, 1999; Buskirk, 1997; Clobert et al., 1998). Il a été envisagé concrètement lors d'une collaboration avec Léonor Palmeira au cours de son DEA et les résultats ont été partiellement publiés dans son rapport (Palmeira, 2003-2004). On ne peut pas encore tirer de conclusion sur la pertinence de cette approche d'un point de vue pratique et il faudra multiplier les essais sur des données et des problèmes de nature variés. Toutefois, d'un point de vue théorique, le fait que la matrice \mathbf{A} soit bi-stochastique implique que les deux points de vue (Geary et Moran) sont réconciliés sans avoir besoin d'introduire la pondération de voisinage (Thioulouse et al., 1995). Cela présente de très gros avantages : d'une part, cela permet d'introduire sans problème les pondérations lignes propres à chaque analyse multivariée; d'autre part, l'indice de Moran correspond à un vrai coefficient de corrélation : il est donc toujours compris entre -1 et 1 ; enfin, la moyenne et la variance restent constantes par permutation, ce qui permet la mise en œuvre de tests non paramétriques sans avoir à recalculer la moyenne et la variance à chaque permutation.

6. DU CORRÉLOGRAMME A L'ORTHOGRAMME

Ces tests globaux peuvent nous renseigner sur l'existence d'un lien entre un tableau et une phylogénie. Ils sont forts utiles de ce point de vue là mais ils ne nous renseignent pas sur la nature de ce lien, en particulier à quel niveau et comment il s'établit. Par exemple, si l'on applique le test d'Abouheif (test de Moran avec la matrice **A**) pour les sept traits biologiques étudiés dans l'article de Rochet et al. (2000), présentés Figure 3.4, on constate qu'ils sont tous plus ou moins significativement liés à la phylogénie (Tableau 3.1). A partir de l'étude de la représentation graphique (Figure 3.4), il est très difficile de repérer de visu des différences de structures entre traits biologiques. Il faut donc des outils pour appréhender la forme et l'intensité de la structure d'un trait dans un arbre phylogénétique. C'est l'objectif de ce paragraphe.

<pre> phy <- newick2phylog(mjrochet\$tre) tab <- log(mjrochet\$tab) tab0 <- data.frame(scalewt(tab)) </pre>			
<pre> # test de Moran avec la matrice A gearymoran(phy\$Aamat, tab0) class: krandtest test number: 7 permutation number: 999 test obs P(X<=obs) P(X>=obs) 1 tm 0.298 0.999 0.003 2 lm 0.323 1 0.001 3 105 0.341 1 0.001 4 t05 0.198 0.981 0.021 5 fb 0.251 0.996 0.006 6 fm 0.268 0.995 0.007 7 egg 0.486 1 0.001 </pre>		<pre> # test de Moran avec la matrice W gearymoran(phy\$Wmat, tab0) class: krandtest test number: 7 permutation number: 999 test obs P(X<=obs) P(X>=obs) 1 tm 0.048 0.954 0.048 2 lm 0.054 0.968 0.034 3 105 0.05 0.952 0.05 4 t05 0.033 0.769 0.233 5 fb 0.039 0.9 0.102 6 fm 0.03 0.704 0.298 7 egg 0.096 1 0.001 </pre>	

Tableau 3.1 : Test de Moran dans sa version non paramétrique pour les sept traits biologiques de l'exemple mjrochet (Annexe 1.14). A gauche, c'est la matrice **A** qui joue le rôle de matrice de proximité. Le test est équivalent au test d'Abouheif. A droite, c'est la matrice **W** qui joue le rôle de matrice de proximité. On constate sur cet exemple, que le test de Moran est très sensible au choix de la matrice de proximité.

Gittleman et Kot (1990) ont été les premiers à poser explicitement ce problème : « *a given set of comparative data forces us to confront two important questions : are there phylogenetic effects in the data ? where is there phylogenetic correlation ?* ». Ils ont donné une solution en adaptant le corrélogramme classiquement utilisé en statistique spatiale (Sokal, 1979), aux données phylogénétiques. Cette solution reste peu satisfaisante pour deux raisons principales.

D'une part la mise en œuvre est très empirique : les auteurs repartent de l'indice de Moran, proposent plusieurs bricolages audacieux pour qu'il ressemble à un coefficient de corrélation, élèvent la matrice de proximité à une puissance arbitraire pour rendre les tests plus sensibles, puis ramènent à 0 les termes inférieurs à une valeur seuil. A chaque valeur seuil, correspond un coefficient du corrélogramme. En introduisant le droit de faire ce que

bon lui semble à quiconque, en particulier de chercher des valeurs pour rendre les tests plus sensibles, on rend les calculs le moins reproductibles possible et la méthode est donc restée inemployée. C'est encore de la statistique *ad hoc*.

D'autre part, le corrélogramme fait partie des trois grandes classes de fonctions structurelles (Legendre & Legendre, 1998) permettant de décrire la structure d'une variable. Il est basé sur le calcul de plusieurs statistiques globales, intégrant pour une classe de distances donnée l'ensemble des couples d'unités statistiques appartenant à cette classe. Il est donc très peu sensible aux variations locales que l'on peut rencontrer entre couples d'unités statistiques appartenant à la même classe (Anselin, 1995).

C'est pourquoi on a choisi une toute autre stratégie. A partir de la définition d'une base orthonormée canoniquement associée à la phylogénie, on propose de décrire le lien entre un trait et la phylogénie au travers de la décomposition de la variance du trait sur les vecteurs de la base. On y associe plusieurs tests non paramétriques contre l'absence de structure. Cette approche s'inscrit dans la lignée de l'analyse spectrale et de l'analyse en ondelettes et fait partie de la classe des transformées orthonormales. Elle est détaillée dans Ollier (sous presse) (Annexe 3.4) et illustrée par un poster (Annexe 3.5).

7. DISCUSSION ET PERSPECTIVES

Les outils proposés dans ce chapitre sont de natures descriptives. Ils permettent d'établir rapidement un premier diagnostic sur la structure des données mais dans aucune mesure l'analyse exploratoire ne permet d'inférer un modèle à partir de ces données. Selon Yoccoz (1994), « *une bonne description des données est absolument nécessaire à l'identification de structures, et pour une identification correcte du modèle. Cependant, une compréhension des structures (i.e, les traduire en terme de processus : de quoi à pourquoi) nécessite des modèles théoriques des processus. C'est souvent loin d'être possible...Plusieurs raisons peuvent être invoquées, entre autres : différents processus peuvent conduire à la même structure (au moins partiellement), et la nature stochastique de ces processus rend délicate l'interprétation d'une seule réalisation de ces processus* ». Plusieurs modèles théoriques de processus évolutifs ont été proposés dans la littérature (Hansen & Martins, 1996). Le plus simple d'entre eux, le modèle brownien (dérive aléatoire), reste le plus utilisé bien qu'il soit complètement irréaliste d'un point de vue biologique. De plus, dans la plupart des cas, il n'y a aucun effort sérieux d'évaluation des hypothèses et les auteurs ne se posent même pas la question de la qualité de leur ajustement aux données. Pourtant, en établissant des liens entre la forme de l'orthogramme et les modèles théoriques, soit par simulation dans un premier temps, soit

mathématiquement dans un second temps, il serait possible d'évaluer la pertinence d'un modèle en fonction des données.

Par ailleurs, on a posé jusque là le problème de la description de la structure d'un trait dans un arbre phylogénétique. On sait caractériser le lien entre un trait et une phylogénie par une décomposition de la variance sur les vecteurs de références d'un arbre phylogénétique. Qu'en est-il de la description d'un tableau dans un arbre phylogénétique ? Comment mesurer le lien entre deux ou plusieurs traits en tenant compte des proximités évolutives ? Plusieurs pistes sont possibles, mais une fois encore, la solution dépendra des données et du problème posé. Si tous les traits présentent la même structure, caractérisée par une forte ressemblance pour deux grands groupes d'espèces isolés d'un point de vue évolutif, on envisagera une analyse inter-classe (vs intra-classe). Si tous les traits présentent la même structure, caractérisée par une dérive aléatoire, on envisagera une analyse sous contrainte en utilisant la matrice de proximité \mathbf{W} . Dans tous les autres cas, le problème reste ouvert...

8. BIBLIOGRAPHIE

Abouheif, E. (1999) A method for testing the assumption of phylogenetic independence in comparative data. *Evolutionary Ecology Research*, 1, 895-909.

Ackerly, D.D. (1997) Plant life histories: a meeting of phylogeny and ecology. *Tree Physiology*, 12, 7-9.

Ackerly, D.D. & Reich, P.B. (1999) Convergence and correlations among leaf size and function in seed plants: a comparative test using independent contrasts. *American Journal of Botany*, 86, 1272-1281.

Anselin, L. (1995) Local indicators of spatial association-LISA. *Geographical Analysis*, 27, 93-115.

Barracough, T.G., Vogler, A.P., & Harvey, P.H. (1998) Revealing the factors that promote speciation. *Philosophical Transactions of the Royal Society London B*, 353, 241-249.

Bauwens, D. & Díaz-Uriarte, R. (1997) Covariation of life-history traits in lacertid lizards: a comparative study. *The American Naturalist*, 149, 91-111.

Blomberg, S.P., Garland, T., & Ives, A.R. (2003) Testing for phylogenetic signal in comparative data. *Evolution*, 57, 717-745.

Buskirk, J.V. (1997) Independent evolution of song structure and note structure in American wood warblers. *Proceedings of the Royal Society B*, 264, 755-761.

Champely, S. & Chessel, D. (2002) Measuring biological diversity using Euclidean metrics. *Environmental and Ecological Statistics*, in press.

Cheverud, J. & Dow, M.M. (1985) An autocorrelation analysis of genetic variation due to lineal fission in social groups of rhesus macaques. *American Journal of Physical Anthropology*, 67, 113-122.

Clarke, K.R. & Warwick, R.M. (1999) The taxonomic distinctness measure of biodiversity: weighting of step lengths between hierarchical levels. *Marine Ecology - Progress Series*, 184, 21-29.

Cleveland, W.S. (1994) *The elements of graphing data* AT&T Bell Laboratories, Murray Hill, New Jersey.

Clobert, J., Garland, T., & Barbault, R. (1998) The evolution of demographic tactics in lizards: a test of some hypotheses concerning life history evolution. *Journal of Evolutionary Biology*, 11, 329-364.

Cornillon, P.-A., Pontier, D., & Rochet, M.J. (2000) Autoregressive models for estimating phylogenetic and environmental effects: accounting for within-species variations. *Journal of Theoretical Biology*, 202, 247-256.

Darwin, C. (1859) *L'origine des espèces* Flammarion.

Farrell, B.D., Dussourd, D.E., & Mitter, C. (1991) Escalation of plant defense : do latex and resin canals spur plant diversification?. *The American Naturalist*, 138, 881-900.

Felsenstein, J. (1985) Phylogenies and the comparative method. *The American Naturalist*, 125, 1-15.

Felsenstein, J. (2004) *Infering phylogenies* Sinauer, Sunderland.

Garland, T.J. & Ives, A.R. (2000) Using the past to predict the present: confidence intervals for regression equations in phylogenetic comparative methods. *The American Naturalist*, 155, 346-364.

Garland, T.J. & Janis, C.M. (1993) Does metatarsal/femur ratio predict maximal running speed in cursorial mammals? *Journal of Zoology*, 229, 133-151.

Gittleman, J.L., Anderson, C.G., Kot, M., & Luh, H.K. (1996). Phylogenetic lability and rates of evolution: A comparison of behavioral, morphological and life history traits. In *Phylogenies and the Comparative Method in Animal Behaviour* (ed E.P. Martins), pp. 166-205. Oxford University Press, Oxford.

Gittleman, J.L. & Kot, M. (1990) Adaptation: statistics and a null model for estimating phylogenetic effects. *Systematic Zoology*, 39, 227-241.

Hansen, T.F. & Martins, E.P. (1996) Translating between microevolutionary process and macroevolutionary patterns: the correlation structure of interspecific data. *Evolution*, 50, 1404-1417.

Harvey, P.H., Leigh Brown, A.J., Maynard Smith, J., & Nee, S., eds. (1996) *New Uses for New Phylogenies*. Oxford University Press, Oxford.

Harvey, P.H. & Pagel, M. (1991) *The Comparative Method in Evolutionary Biology* Oxford University Press.

Le Guyader, H. (2003) *Classification et évolution Le Pommier*.

Legendre, P. & Legendre, L. (1998) *Numerical ecology*, 2nd English edition edn. Elsevier Science BV, Amsterdam.

Losos, J.B., Jackman, T.R., Larson, A., Queiroz, d., & L., R.-S.K. (1998) Contingency and determinism in replicated adaptive radiations of island lizards. *Science*, 279, 2115-2118.

Malhotra, A., Thorpe, R.S., Black, H., Daltry, J.C., & W., W. (1996). Relating geographic patterns to phylogenetic process. In *New Uses for New Phylogenies* (eds P.H. Harvey, A.J. Leigh Brown, J. Maynard Smith & N. S.). Oxford University Press, Oxford.

Martins, E.P. (1996) Phylogenies, spatial autoregression, and the comparative method: a computer simulation test. *Evolution*, 50, 1750-1765.

- Martins, E.P.** (2000) Adaptation and the comparative method. *Tree*, 15, 296-299.
- Martins, E.P. & Hansen, T.F.** (1997) Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. *The American Naturalist*, 149, 646-667.
- Mitter, C., Farrell, B., & Wiegmann, B.** (1998) The phylogenetic study of adaptive zones: has phytophagy promoted insect diversification? *The American Naturalist*, 132, 107-128.
- Ollier, S., Couteron, P., & Chessel, D.** (sous presse) Orthonormal transforms to describe and test the phylogenetic signal. *Biometrics*.
- Orr, M.R. & Smith, T.B.** (1998) Ecology and speciation. *Trends in Ecology and Evolution*, 13, 502-506.
- Palmeira, L.** (2003-2004). Influence des substitutions dépendantes du voisinage sur les méthodes reconstruction phylogénétique, Lyon.
- Perrière, G. & Gouy, M.** (1996) WWW-Query: An on-line retrieval system for biological sequence banks. *Biochimie*, 78, 364-369.
- Podos, J.** (2001) Correlated evolution of morphology and vocal signal structure in Darwin's finches. *Nature*, 409, 185-188.
- Rao, C.R.** (1982) Diversity and dissimilarity coefficients: a unified approach. *Theoretical Population Biology*, 21, 24-43.
- Rochet, M.J., Cornillon, P.-A., Sabatier, R., & Pontier, D.** (2000) Comparative analysis of phylogenic and fishing effects in life history patterns of teleost fishes. *Oikos*, 91, 255-270.
- Rohlf, F.J.** (2001) Comparative methods for the analysis of continuous variables: geometric interpretations. *Evolution*, 55, 2143-2160.
- Sanderson, M.J., Baldwin, B.G., Bharatan, G., Campbell, C.S., Ferguson, D., Porter, C., Von Dohlen, C., Wojciechowski, M.F., & Donoghue, M.J.** (1993) The growth of phylogenetic information and the need for a phylogenetic database. *Systematic Biology*, 42, 562-568.
- Sanderson, M.J. & Donoghue, M.J.** (1996) Reconstructing shifts in diversification rates on phylogenetic trees. *Trends in Ecology and Evolution*, 11, 15-20.
- Sanderson, M.J. & Donoghue, M.J.** (1998) Phylogenetic supertrees: assembling the trees of life. *Trends in Ecology and Evolution*, 13, 105-109.
- Smith, A.B., Littlewood, D.T.J., & Wray, G.A.** (1996). Comparative evolution of larval and adult life-history stages and small subunit ribosomal RNA amongst post-Palaeozoic

echinoids. In *New Uses for New Phylogenies* (eds P.H. Harvey, A.J. Leigh Brown, J. Maynard Smith & S. Nee). Oxford University Press, Oxford.

Sokal, R.R. (1979). Ecological parameters inferred from spatial correlograms. In *Contemporary quantitative ecology and related econometrics* (eds G.P. Patil & M. Rosenzweig), pp. 167-196. International Co-operative Publishing House, Fairland.

Sokal, R.R. & Rohlf, F.J. (1969) *Biometry* Third edition. W.H. Freeman and Company, New-York.

Statzner, B., Hoppenhaus, K., Arens, M.-F., & Richoux, P. (1997) Reproductive traits, habitat use and templet theory: a synthesis of world-wide data on aquatic insects. *Freshwater Biology*, 38, 109-135.

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

Vitt, L.J., Zani, P.A., & Esposito, M.C. (1999) Historical ecology of Amazonian lizards : implications for community ecology. *Oikos*, 87, 286-294.

Yoccoz, N. (1994). Déduction et inférence en biologie des populations: le rôle des modèles. L'exemple des petits mammifères et de leurs variations cycliques.

CONCLUSION

Au terme de ce travail, il paraît utile de dresser un bilan et de signaler quelques perspectives. Nous nous sommes efforcés dans les pages précédentes d'apporter des éléments de réponse à la question suivante : quand et comment intégrer les contraintes spatiales, temporelles et évolutives en analyse des données écologiques ? Ceci nous a conduit à faire appel au langage algébrique et géométrique utilisé en analyse de données, afin de rendre compte de la diversité des méthodes proposées pour intégrer les proximités. Je propose de terminer cette thèse par un bref résumé des résultats « techniques », en insistant davantage sur la démarche qui a permis d'aboutir à ce travail. J'en profite pour décrire comment s'est instauré le dialogue interdisciplinaire au cours de ma thèse. Enfin, à titre de perspectives, j'évoque les développements de méthode multidimensionnelle multiéchelle envisageables à plus long terme.

Dans cette thèse, bien que la consultation statistique ait parfois été l'unité du dialogue interdisciplinaire (comme dans le chapitre deux), la donnée n'a pas toujours été le premier élément initiateur de ce dialogue. En effet, un logiciel comme R, gratuit et complètement transparent a largement contribué à renouveler sa nature. La disponibilité de R en tant que logiciel libre, permet aux utilisateurs d'examiner, de modifier et d'améliorer le code source de R, puis de partager ces changements avec les autres. Elle assure ainsi un développement coopératif, et permet la réalisation d'un dialogue au-delà des rencontres effectives entre personnes. On peut alors dire, que les programmes, au même titre que la donnée, soulèvent des questions pertinentes et constituent du matériel expérimental pour le biométricien. Dans un sens, ils contribuent à la construction théorique et sont également destinés à provoquer la prise en compte de problèmes statistiques nouveaux. Le chapitre un en est une illustration : c'est en traduisant la possibilité de combiner deux procédures informatiques à travers la fonction `multispati(...)` (Annexe 2.12) que l'idée de généraliser l'approche de Wartenberg (1985), à l'ensemble des analyses multidimensionnelles, a émergé.

Bien que le développement de l'informatique renouvelle potentiellement les démarches de la pensée interdisciplinaire, il ne faut pas minimiser la nécessité du traitement réel des données. En effet, « *la confrontation aux données réelles est autant une vérification du réalisme des constructions mathématiques ou informatiques, qu'une provocation à la prise en compte de problèmes statistiques nouveaux* (Escoufier, 1983) ». La distance qui sépare les analyses locales des objectifs cartographiques (chapitre 1) est une illustration de la première

assertion. Par ailleurs, la consultation statistique présentée dans le chapitre 2, en posant explicitement les problèmes de typologie de structures multiéchelles a permis d'initier un problème statistique nouveau, ce qui confirme la deuxième partie de la citation. Ces deux premiers chapitres confortent ainsi l'idée que « *les échanges interdisciplinaires ont une dynamique originale, parce que non entièrement dépendants du processus bibliographique* (Chessel, 1992) ».

Le dernier chapitre montre comment « *une discipline expérimentale recrée, à partir de ses besoins, de la statistique ad hoc dont une bonne partie est implicitement en connexion avec des modèles centraux* (Chessel, 1992) ». La pratique empirique d'Abouheif s'est avérée être implicitement reliée au calcul des indices de Moran et Geary. En transportant par hasard d'un point à un autre une idée, on a trouvé un endroit où les concepts et les usages propres à la statistique spatiale, s'expriment selon une articulation beaucoup plus fine et mieux adaptée. En effet, la proposition méthodologique de Pace et Le Sage (2002), d'utiliser une matrice bistochastique ne trouve pas son plein usage en statistique spatiale alors qu'elle s'impose naturellement pour l'intégration des proximités évolutives par l'intermédiaire de la matrice d'Abouheif.

Par ailleurs, en faisant d'une observation numérique organisée (un trait en face d'une phylogénie), le représentant d'une classe de problèmes (l'analyse statistique de la structure), on a pu faire appel à une pratique (transformée orthonormale) faisant référence à un théorème universel (théorème de Parseval) afin de définir une méthode canonique. Encore un exemple où une pratique canonique attachée à un théorème universel est transférée dans le champ expérimental par l'intermédiaire du dialogue interdisciplinaire.

Nous n'avons considéré dans tout ce travail que des analyses multidimensionnelles à une échelle (chapitre 1) et des analyses multiéchelles d'une variable (chapitre 2 et 3) : le nombre de méthodes proposées est déjà considérable. Les analyses multidimensionnelles multiéchelles, dont l'archétype est le cube de données relevés-espaces-échelles (spatiales, temporelles, ou évolutives), a donné lieu à plusieurs développements statistiques connus sous le nom d'ordination multiéchelle (« *multiscale ordination* », (Couteron & Ollier, sous presse; Di Bella & Jona-Lasinio, 1996; Noy-Meir & Anderson, 1971; Ver Hoef & Glenn-Lewin, 1989; Wagner, 2004)). La prise en compte au niveau des analyses multidimensionnelles multiéchelles, de la diversité des analyses de base, tant du point de vue de l'analyse multidimensionnelle que du point de vue de l'analyse multiéchelle, constitue sûrement un enjeu considérable pour les chercheurs en écologie. Ces méthodes, bien que linéaires,

constituent toutefois un ensemble d'une complexité déjà non négligeable pour le thématicien. Avant de pouvoir passer au modèle mathématique supérieur, qui s'exprime par analogie aux méthodes K-tableaux par l'analyse des triplets $(\mathbf{X}, \mathbf{Q}, \mathbf{A}_k)_{1 \leq k \leq K}$, il semblait donc indispensable de commencer par une réflexion de fond s'adressant aux méthodes les plus simples. Maintenant que les programmes d'analyse multiéchelle d'une variable sont en place, dans le même environnement que ceux d'analyse multidimensionnelle, on peut commencer notre réflexion de fond sur les analyses multidimensionnelles multiéchelles associées.

BIBLIOGRAPHIE

Chessel, D. (1992) Echanges interdisciplinaires en analyse de données écologiques. Mémoire d'habilitation. Université Lyon 1.

Couteron, P. & Ollier, S. (sous presse) A generalized variogram-based framework for multiscale ordination. *Ecology*.

Di Bella, G. & Jona-Lasinio, G. (1996) Including spatial contiguity information in the analysis of multispecific patterns. *Environmental and Ecological Statistics*, 3, 269-280.

Escoufier, Y. (1983) Réflexions sur les activités du statisticien universitaire. *Statistique et Analyse des Données*, 8, 76-82.

Noy-Meir, I. & Anderson, D.J. (1971). Multivariate pattern analysis, or multiscale ordination: towards a vegetation hologram ? In *Statistical Ecology, III Many species populations ecosystems and systems analysis* (eds G.P. Patil, E.C. Pielou & W.E. Waters), pp. 208-231. Pennsylvania State University Press.

Pace, R.K. & LeSage, J.P. (2002) Semiparametric maximum likelihood estimates of spatial dependence. *Geographical Analysis*, 34, 76-90.

Ver Hoef, J.M. & Glenn-Lewin, C.G. (1989) Multiscale ordination: a method for detecting pattern at several scales. *Vegetatio*, 82, 59-67.

Wagner, H.H. (2004) Direct multi-scale ordination with canonical correspondence analysis. *Ecology*, 85, 342-351.

Wartenberg, D.E. (1985) Multivariate spatial correlations: a method for exploratory geographical analysis. *Geographical Analysis*, 17, 263-283.

BIBLIOGRAPHIE

Abouheif, E. (1999) A method for testing the assumption of phylogenetic independence in comparative data. *Evolutionary Ecology Research*, 1, 895-909.

Abramovich, F., Bailey, T.C., & Sapatinas, T. (2003). Wavelet analysis and its applications.

Ackerly, D.D. (1997) Plant life histories: a meeting of phylogeny and ecology. *Tree Physiology*, 12, 7-9.

Ackerly, D.D. & Reich, P.B. (1999) Convergence and correlations among leaf size and function in seed plants: a comparative test using independent contrasts. *American Journal of Botany*, 86, 1272-1281.

Afriat, S.N. (1957) Orthogonal and oblique projectors and the characteristics of pairs of vector spaces. *Proceedings of the Cambridge Philosophical Society, Mathematical and Physical Sciences*, 53, 800-816.

Allain, C. & Cloitre, M. (1991) Characterizing the lacunarity of random and deterministic fractal sets. *Physical Review A*, 44, 3552-3557.

Anselin, L. (1996). The Moran scatterplot as an ESDA tool to assess local instability in spatial association. In *Spatial analytical perspectives on GIS* (eds M.M. Fischer, H.J. Scholten & D. Unwin), pp. 111-125. Taylor and Francis, London.

Anselin, L. & Hudak, S. (1992) Spatial econometrics in practice: A review of software options. *Regional Science and Urban Economics*, 22, 509-536.

Anselin, L., Syabri, I., & Smirnov, O. (2002) Visualizing multivariate spatial correlation with dynamically linked windows. In *CSISS Specialist Meeting on New Tools in Spatial Data Analysis* (eds L. Anselin & S.J. Rey), Santa Barbara, CA.

Anselin, L. (1995) Local indicators of spatial association-LISA. *Geographical Analysis*, 27, 93-115.

Aubry, P. (2000) Le traitement des variables régionalisées en écologie. Apports de la géomatique et de la géostatistique. Thèse de doctorat, Université Claude Bernard.

Auda, Y. (1983) Rôle des méthodes graphiques en analyse des données : application au dépouillement des enquêtes écologiques. Thèse de 3^o cycle, Université Lyon 1.

Aufaure, M.A., Yeh, L., & Zeitouni, K. (2000). Fouille de données spatiales. Ecole Thématique "Nouveaux défis en Sciences de l'Information : Documents & Evolution", Faculté des Sciences de Saint-Jérôme, Marseille.

Banet, T.A. & Lebart, L. (1984). Local and Partial Principal Component Analysis (PCA) and Correspondence Analysis (CA). In COMPSTAT 84 (ed I.A.f.S. Computing.), pp. 113-123. Physica-Verlag, Vienna.

Barbault, R. (1992) *Ecologie des peuplements. Structure, dynamique et évolution* Masson, Paris.

Barracough, T.G., Vogler, A.P., & Harvey, P.H. (1998) Revealing the factors that promote speciation. *Philosophical Transactions of the Royal Society London B*, 353, 241-249.

Bavaud, F. (1998) Models for spatial weights: a systematic look. *Geographical Analysis*, 50, 155-171.

Bauwens, D. & Díaz-Uriarte, R. (1997) Covariation of life-history traits in lacertid lizards: a comparative study. *The American Naturalist*, 149, 91-111.

Benali, H. & Escofier, B. (1990) Analyse factorielle lissée et analyse factorielle des différences locales. *Revue de Statistique Appliquée*, 38, 55-76.

Berge, C. (1967) *Théorie des graphes et ses applications* Dunod, paris.

Besse, P. (1979) *Etude descriptive d'un processus ; approximation, interpolation*. Thèse de 3^{ème} cycle, Université Paul Sabatier, Toulouse.

Bivand, R. (1980) A Monte Carlo study of correlation estimation with spatially autocorrelated observations. *Quaestiones Geographicae*, 6, 5-10.

Blondel, J. (1985) *Biogéographie évolutive* Masson, Paris.

Blomberg, S.P., Garland, T., & Ives, A.R. (2003) Testing for phylogenetic signal in comparative data. *Evolution*, 57, 717-745.

Bohte, Z., Cepar, D., Kosmelj, K., & Ljubljana, Y.U. (1980) Clustering of time series. COMPSTAT.

Borcard, D. & Legendre, P. (1994) Environmental control and spatial structure in ecological communities: an example using oribatid mites (Acari, Oribatei). *Environmental and Ecological Statistics*, 1, 37-61.

Borcard, D., Legendre, P., & Drapeau, P. (1992) Partialling out the spatial component of ecological variation. *Ecology*, 73, 1045-1055.

Borcard, D., Legendre, P., Avois-Jacquet, C., & Tuomisto, H. (2004) Dissecting the spatial structure of ecological data at multiple scales. *Ecology*, 85, 1826-1832.

Boyé, M., Cabaussel, G., & Perrot, Y. (1979). *Climatologie*. In *Atlas des départements français d'Outre Mer*, 4: la Guyane Française (ed C.a. ORSTOM), pp. 1-4.

Bradshaw, G.A. & Spies, T.A. (1992) Characterizing canopy gap structure in forests using wavelet analysis. *Journal of Ecology*, 80, 205-215.

Brillinger, D.R., Guttorp, P.M., & Schoenberg, F.P. (2002). Point processes, temporal. In *Encyclopedia of Environmetrics* (eds A.H. El-Shaarawi & W.W. Piegorsch), Vol. 3, pp. 1577–1581. John Wiley & Sons, Ltd, Chichester.

Brown, J.H. & Maurer, B.A. (1989) Macroecology: the division of food and space among species on continents. *Science*, 243, 1145-1150.

Buskirk, J.V. (1997) Independent evolution of song structure and note structure in American wood warblers. *Proceedings of the Royal Society B*, 264, 755-761.

Champely, S. & Chessel, D. (2002) Measuring biological diversity using Euclidean metrics. *Environmental and Ecological Statistics*, in press.

Chatelin, F. (1988) Valeurs propres de matrices Masson, Paris.

Chessel, D. (1992) Echanges interdisciplinaires en analyse de données écologiques. Mémoire d'habilitation. Université Lyon 1.

Chessel, D., Dufour, A.-B., & Thioulouse, J. (Submitted) The ade4 package. *R News*.

Chessel, D. & Mercier, P. (1993). Couplage de triplets statistiques et liaisons espèces-environnement. In *Biométrie et Environnement* (eds J.D. Lebreton & B. Asselain), pp. 15-44. Masson, Paris.

Chevenet, F., Dolédec, S., & Chessel, D. (1994) A fuzzy coding approach for the analysis of long-term ecological data. *Freshwater Biology*, 31, 295-309.

Cheverud, J. & Dow, M.M. (1985) An autocorrelation analysis of genetic variation due to lineal fission in social groups of rhesus macaques. *American Journal of Physical Anthropology*, 67, 113-122.

Clarke, K.R. & Warwick, R.M. (1999) The taxonomic distinctness measure of biodiversity: weighting of step lengths between hierarchical levels. *Marine Ecology - Progress Series*, 184, 21-29.

Cleveland, W.S. (1994) *The elements of graphing data* AT&T Bell Laboratories, Murray Hill, New Jersey.

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

Clobert, J., Garland, T., & Barbault, R. (1998) The evolution of demographic tactics in lizards: a test of some hypotheses concerning life history evolution. *Journal of Evolutionary Biology*, 11, 329-364.

Conradsen, K., Nielsen, B.K., & Thyrted, T.A. (1985) Comparison of min/max autocorrelation factor analysis and ordinary factor analysis. In *Proceedings from Symposium in Applied Statistics*, Vol. 47-56. Technical University of Denmark, Lyngby, Denmark.

Cornillon, P.-A., Amenta, P., & Sabatier, R. (1999). Three-way data arrays with double neighbourhood relations as a tool to analyze a contiguity structure. In *Classification and data analysis. Theory and Application* (eds M. Vichi & O. Opitz), pp. 263-270. Springer-Verlag, Berlin.

Cornillon, P.-A. (1998) *Prise en compte de proximités en analyse factorielle et comparative*. Thèse, Ecole Nationale Supérieure Agronomique, Montpellier.

Cornillon, P.-A. & Sabatier, D. (1998). Local multivariate analysis. In *Advances in data science and classification* (eds A. Rizzi, M. Vichi & H.H. Bock). Springer.

Cornillon, P.-A., Pontier, D., & Rochet, M.J. (2000) Autoregressive models for estimating phylogenetic and environmental effects: accounting for within-species variations. *Journal of Theoretical Biology*, 202, 247-256.

Couteron, P. & Ollier, S. (sous presse) A generalized variogram-based framework for multiscale ordination. *Ecology*.

Couteron, P. (2001) Using spectral analysis to confront distributions of individual species with an overall periodic pattern in semi-arid vegetation. *Plant Ecology*, 156, 229-243.

Couteron, P. (2002) Quantifying change in patterned semi-arid vegetation by Fourier analysis of digitised aerial photographs. *International Journal of Remote Sensing*, 23, 3407-3425.

Couteron, P., Mahamane, A., & Ouedraogo, P. (1996) Analyse de la structure de peuplements ligneux dans un "fourré tigré" au nord Yatenga (Burkina Faso). Etat actuel et conséquences évolutives. *Annales des Sciences Forestières*, 53, 867-884.

Couteron, P., Pélissier, R., Mapaga, D., Molino, J.F., & Teillier, L. (2002) Ecological valorisation of a management-oriented forest inventory in French Guiana. *Forest Ecology and Management*.

Cox, D.R. & Lewis, P.A.W. (1969) *L'analyse statistique des séries d'évènements* Traduction de Larrieu (J.) Dunod, Paris.

Dale, M.R.T. (1999) *Spatial pattern analysis in plant ecology* Cambridge University Press.

Dale, M.R.T., Dixon, P., Fortin, M.J., Legendre, P., Myers, D., & Rosenberg, M. (2002) Conceptual and mathematical relationships among methods for spatial analysis. *ecography*, 25, 558-577.

Dale, M.R.T. & Mah, M. (1998) The use of wavelets for spatial pattern analysis in ecology. *Journal of Vegetation Science*, 9, 805-814.

Darwin, C. (1859) *L'origine des espèces* Flammarion.

Daubechies, I. (1992) *Ten Lectures on Wavelets* SIAM, Philadelphia.

- de Belair, G.** (1981) Biogéographie et aménagement : la plaine de La Mafragh (Annaba, Algérie). Thèse de 3^o cycle. Université Paul Valéry, Montpellier.
- Delattre, P.** (1995) Interdisciplinaires (recherches). Encyclopaedia Universalis, 12 (version CD-ROM 5.0 1999).
- Delor, C., Perrin, J., Truffert, C., Asfirane, F., & Rossi, P.** (1998) Images géophysiques dans le socle guyanais. *Géochronique*, 67, 7-12.
- Dessier, A. & Laurec, A.** (1978) Le cycle annuel du zooplancton à Pointe-Noire (RP Congo). *Description mathématique. Oceanologica acta*, 1, 285-304.
- Di Bella, G. & Jona-Lasinio, G.** (1996) Including spatial contiguity information in the analysis of multispecific patterns. *Environmental and Ecological Statistics*, 3, 269-280.
- Diggle, P.J.** (1990) *Time Series: a biostatistical introduction* Clarendon Press, Oxford.
- Dolédec, S., Chessel, D., & Olivier, J.M.** (1995) L'analyse des correspondances décentrée: application aux peuplements ichtyologiques du haut-Rhône. *Bulletin Français de la Pêche et de la Pisciculture*, 336, 29-40.
- Drake, J.B. & Weishampel, J.F.** (2000) Multifractal analysis of canopy height measures in a longleaf pine savanna. *Forest Ecology and Management*, 128, 121-127.
- Dungan, J.L., Perry, J., Dale, M.R.T., Citron-Pousty, S., Fortin, M.J., Jakomulska, A., Legendre, A., Miriti, M., & Rosenberg, M.S.** (2002) A balanced view of scaling in spatial statistical analysis. *Ecography*, 25, 626-640.
- Durand, J.-D., Guinand, B., & Bouvet, Y.** (1999) Local and global multivariate analysis of geographical mitochondrial DNA variation in *Leuciscus cephalus* L. 1758 (Pisces: Cyprinidae) in the Balkan Peninsula. *Biological Journal of the Linnean Society*, 67, 19-42.
- Ersbll, B.K. (1989) Transformations and classifications of remotely sensed data. Ph.D. thesis, University of Denmark, Lyngby.
- Escoufier, Y.** (1987). The duality diagramm : a means of better practical applications. In *Development in numerical ecology* (eds P. Legendre & L. Legendre), pp. 139-156. NATO advanced Institute , Serie G .Springer Verlag, Berlin.
- Escoufier, Y.** (1983) Réflexions sur les activités du statisticien universitaire. *Statistique et Analyse des Données*, 8, 76-82.
- Estève, J.** (1978). Les méthodes d'ordination : éléments pour une discussion. In *Biométrie et Ecologie* (eds J.M. Legay & R. Tomassone), pp. 223-250. Société Française de Biométrie, Paris.
- Faraj, A. & Cailly, F.** (2001) Spatial contiguity analysis: a method for describing spatial structures of seismic data. *Journal of Petroleum Science and Engineering*, 31, 93-111.
- Farrell, B.D., Dussourd, D.E., & Mitter, C.** (1991) Escalation of plant defense : do latex and resin canals spur plant diversification?. *The American Naturalist*, 138, 881-900.

Felsenstein, J. (1985) Phylogenies and the comparative method. *The American Naturalist*, 125, 1-15.

Felsenstein, J. (2004) *Infering phylogenies* Sinauer, Sunderland.

Fievet, E., Eppe, F., & Dolédec, S. (2001). Etude de la variabilité morphométrique et génétique des populations de Cacadors (*Atya innocous* et *Atya scabra*) de l'île de Basse-Terre. Direction Régionale de L'Environnement Guadeloupe, Laboratoire des hydrosystèmes fluviaux, Université Lyon 1, 43 Bd du 11 Novembre 1918, 69622, Villeurbanne cedex, France.

Fisher, N.I. (1993) *Statistical Analysis of Circular Data* Cambridge University Press.

Fortin, M.-J. & Jacquez, G.M. (2000) Randomization tests and spatially autocorrelated data. *Bulletin of the Ecological Society of America*, 81, 201-205.

Fortin, M.-J., Dale, M.R.T., & Ver Hoef, J.M. (2002). Spatial analysis in ecology. In *Encyclopedia of Environmetrics* (eds A.H. El-Shaarawi & W.W. Piegorsch), Vol. 2, pp. 2051-2058. John Wiley & Sons, Chichester.

Frontier, S. & Pichod-Viale, D. (1990) *Ecosystèmes. Structure, fonctionnement, evolution*, Second edn. Dunod.

Gabriel, K.R. & Sokal, R.R. (1969) A new statistical approach to geographic variation analysis. *Systematic Zoology*, 18, 259-278.

Garland, T.J. & Ives, A.R. (2000) Using the past to predict the present: confidence intervals for regression equations in phylogenetic comparative methods. *The American Naturalist*, 155, 346-364.

Garland, T.J. & Janis, C.M. (1993) Does metatarsal/femur ratio predict maximal running speed in cursorial mammals? *Journal of Zoology*, 229, 133-151.

Geary, R.C. (1954) The contiguity ratio and statistical mapping. *The Incorporated Statistician*, 5, 115-145.

Ghertsos, K., Luczak, C., & Dauvin, J.-C. (2001) Identification of global and local components of spatial structure of marine benthic communities: example from the Bay of Seine (Eastern English Channel). *Journal of Sea Research*, 45, 63-77.

Gimaret-Carpentier, C. (1999) Analyse de la biodiversité à partir d'une liste d'occurrences d'espèces : nouvelles méthodes d'ordination appliquées à l'étude de l'endémisme dans les Ghâts occidentaux. Thèse de doctorat, Université Lyon 1.

Gittleman, J.L. & Kot, M. (1990) Adaptation: statistics and a null model for estimating phylogenetic effects. *Systematic Zoology*, 39, 227-241.

Gittleman, J.L., Anderson, C.G., Kot, M., & Luh, H.K. (1996). Phylogenetic lability and rates of evolution: A comparison of behavioral, morphological and life history traits. In

Phylogenies and the Comparative Method in Animal Behaviour (ed E.P. Martins), pp. 166-205. Oxford University Press, Oxford.

Goodall, D.W. (1954) Objective methods for the classification of vegetation III. An essay in the use of factor analysis. *Australian Journal of Botany*, 2, 304-324.

Goodall, D.W. (1974) A new method for the analysis of spatial pattern by random pairing of quadrats. *Vegetatio*, 53, 153-160.

Goovaerts, P. (1992) Multivariate geostatistical tools for studying scale-dependent correlation structures and describing space-time variation, Thèse de doctorat, Université Catholique de Louvain, Louvain la Neuve.

Goulard, M., Voltz, M., & Monestiez, P. (1987) Comparaison d'approches multivariées pour l'étude de la variabilité spatiale des sols. *Agronomie*, 7, 657-665.

Greenacre, M.J. (1984) Theory and applications of correspondence analysis Academic Press, London.

Greig-Smith, P. (1952) The use of random and contiguous quadrats in the study of the structure of plant communities. *Annals of Botany*, London, 16, 293-316.

Greig-Smith, P. (1961) Data on pattern within plant communities. I The analysis of pattern. *Journal of Ecology*, 49, 695-702.

Greig-Smith, P. & Chadwick, M.J. (1965) Data on pattern within plant communities. III. *Acacia-Capparis* semi-desert scrub in the Sudan. *Journal of Ecology*, 53, 465-474.

Griffith, D.A. (2000) Eigenfunction properties and approximations of selected incidence matrices employed in spatial analyses. *Linear Algebra and its Applications*, 321, 95-112.

Grunsky, E.C. (2002) R: a data analysis and statistical programming environment—an emerging tool for the geosciences. *Computers & Geosciences*, 28, 1219-1222.

Grunsky, E.C. & Agterberg, F.P. (1988) Spatial and multivariate analysis of geochemical data from metavolcanic rocks in the Ben Nevis area, Ontario. *Mathematical Geology*, 20, 825-861.

Grunsky, E.C. & Agterberg, F.P. (1989) The application of spatial factor analysis to unconditional simulations with implications for mineral exploration. In *Proceedings, 21st International Symposium on Computers in the Mineral Industry*, pp. 194-208. Society of Mining Engineers of AIME, Littleton, Colorado, Las Vegas, Nevada, March 1989.

Grunsky, E.C. & Agterberg, F.P. (1991) SPFA: a FORTRAN-77 program for spatial factor analysis of multivariate data. *Computers & Geosciences*, 17, 133-160.

Grunsky, E.C., Chen, Q., & Agterberg, F.P. (1996). Applications of spatial factor analysis to multivariate data. In *Geologic Modeling and Mapping* (eds A. Foerster & D.F. Merriams), pp. 229-261. Plenum, New York.

Guttorp, P.M., Brillinger, D.R., & Schoenberg, F.P. (2002). Point processes, spatial. In *Encyclopedia of Environmetrics* (eds A.H. El-Shaarawi & W.W. Piegorsch), Vol. 3, pp. 1571-1573. John Wiley & Sons, Ltd, Chichester.

Haar, A. (1910) Zur Theorie der Orthogonalen Funktionensysteme. *Mathematische Annalen*, 69, 331–371.

Hanafi, M. (1997) Structure de l'ensemble des analyses multivariées des tableaux de données à trois entrées : éléments théoriques et appliqués. Thèse de doctorat, Université Lyon 1.

Hansen, T.F. & Martins, E.P. (1996) Translating between microevolutionary process and macroevolutionary patterns: the correlation structure of interspecific data. *Evolution*, 50, 1404-1417.

Hanski, I. (1994) Spatial scale, patchiness and population dynamics on land. *Phil. Trans. R. Soc. London*, 343B, 19-25.

Harvey, P.H., Leigh Brown, A.J., Maynard Smith, J., & Nee, S., eds. (1996) *New Uses for New Phylogenies*. Oxford University Press, Oxford.

Harvey, P.H. & Pagel, M. (1991) *The Comparative Method in Evolutionary Biology* Oxford University Press.

Harville, D.A. (1997) *Matrix algebra from a statistician's perspective* Springer, New York.

Hatheway, W.H. (1971). Contingency table analysis of rain forest vegetation. In *Statistical Ecology. III Many species populations ecosystems and systems analysis* (eds G.P. Patil, E.C. Pielou & W.E. Waters), pp. 271-314. Pennsylvania State University Press.

Hérissé, C. (2001). Influences des environnements locaux et régionaux sur l'ichtyofaune: structure en réseau et relation de voisinage. Approche exploratoire. Application au Bassin de la Haute-Saône. DEA analyse et modélisation des systèmes biologiques, Université Claude Bernard, Lyon.

Hill, M.O. (1973) The intensity of spatial pattern in plant communities. *Journal of Ecology*, 61, 225-235.

Hill, M.O. (1974) Correspondence analysis : A neglected multivariate method. *Journal of the Royal Statistical Society, C*, 23, 340-354.

Hill, M.O. & Smith, A.J.E. (1976) Principal component analysis of taxonomic data with multi-state discrete characters. *Taxon*, 25, 249-255.

- Hutter, S.** (2001). Etude géomorphologique du massif forestier de Counami. CIRAD.
- Ihaka, R. & Gentleman, R.** (1996) R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5, 299-314.
- Jayet, H.** (1999) *Analyse spatiale quantitative, une introduction* Hermes.
- Jenkins, G.M. & Watts, D.G.** (1968) *Spectral analysis and Its Applications* Holden-Day: San Francisco.
- Kaiser, H.F.** (1958) The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23, 187-200.
- Keitt, T.H., Bjørnstad, O.N., Dixon, P., & Citron-Pousty, S.** (2002) Accounting for spatial pattern when modeling organism-environment interactions. *Ecography*, 25, 616–625.
- Kiers, H.A.L.** (1994) Simple structure in component analysis techniques for mixtures of qualitative and quantitative variables. *Psychometrika*, 56, 197-212.
- Koenig, W.D. & Knops, J.M.H.** (1998) Testing for spatial autocorrelation in ecological studies. *Ecography*, 21.
- Kroonenberg, P.M. & Lombardo, R.** (1999) Nonsymmetric correspondence analysis: a tool for analysing contingency tables with a dependence structure. *Multivariate Behavioral Research*, 34, 367-396.
- Lark, R.M. & Webster, R.** (1999) Analysis and elucidation of soil variation using wavelets. *European Journal of Soil Science*, 50, 185-206.
- Lark, R.M. & Webster, R.** (2001) Changes in variance and correlation of soil properties with scale and location: analysis using an adapted maximal overlap discrete wavelet transform. *European Journal of Soil Science*, 52, 547-562.
- Lavit, C.** (1988) *Analyse conjointe de tableaux quantitatifs* Masson, Paris.
- Le Foll, Y.** (1982) Pondération des distances en analyse factorielle. *Statistique et Analyse des données*, 7, 13-31.
- Le Guyader, H.** (2003) *Classification et évolution* Le Pommier.
- Lebart, L.** (1969) *Analyse statistique de la contiguïté*. Publication de l'Institut de Statistiques de l'Université de Paris, 28, 81-112.
- Legay, J.M.** (1976) Pour une Biométrie. *Statistique et Analyse des Données*, 1, 5-11.
- Legay, J.M. & Barbault, R.** (1995). Une révolution silencieuse dans les sciences de la Nature. In *La révolution technologique en écologie* (eds J.M. Legay & R. Barbault). Masson.

Legendre, P., Dale, M.R.T., Fortin, M.J., Gurevitch, J., Hohn, M., & Myers, D. (2002) The consequences of spatial structure for the design and analysis of ecological surveys. *Ecography*, 25, 601-615.

Legendre, P. & Legendre, L. (1998) *Numerical ecology*, 2nd English edition edn. Elsevier Science BV, Amsterdam.

Leps, J. (1990). Comparison of transect methods for the analysis of spatial pattern. In *Spatial Processes in plant Communities* (eds F. Krahulec, A.D.Q. Agniew, S. Agniew & H.J. Willems), pp. 71-81. SPB Academic Publishing bv, The Hague, Liblice, Tchechoslovaquie.

Levin, S.A. (1992) The problem of pattern and scale in ecology. *Ecology*, 73, 1943-1967.

Liebhold, A.M. & Gurevitch, J. (2002) Integrating the statistical analysis of spatial data in ecology. *ecography*, 25, 553-557.

Light, R.J. & Margolin, B.H. (1971) An analysis of variance for categorical data. *Journal of the American Statistical Association*, 66, 534-544.

Losos, J.B., Jackman, T.R., Larson, A., Queiroz, d., & L., R.-S.K. (1998) Contingency and determinism in replicated adaptive radiations of island lizards. *Science*, 279, 2115-2118.

Malhotra, A., Thorpe, R.S., Black, H., Daltry, J.C., & W., W. (1996). Relating geographic patterns to phylogenetic process. In *New Uses for New Phylogenies* (eds P.H. Harvey, A.J. Leigh Brown, J. Maynard Smith & N. S.). Oxford University Press, Oxford.

Martins, E.P. (1996) Phylogenies, spatial autoregression, and the comparative method: a computer simulation test. *Evolution*, 50, 1750-1765.

Martins, E.P. (2000) Adaptation and the comparative method. *Tree*, 15, 296-299.

Martins, E.P. & Hansen, T.F. (1997) Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. *The American Naturalist*, 149, 646-667.

Méot, A. (1992) Explication de contraintes de voisinage en analyse multivariée. Application dans le cadre de problématiques agronomiques. Thèse de 3^e cycle, Université Claude Bernard (Lyon I).

Méot, A., Chessel, D., & Sabatier, R. (1993). Opérateurs de voisinage et analyse des données spatio-temporelles. In *Biométrie et Environnement* (eds J.D. Lebreton & B. Asselain), pp. 45-72. Masson, Paris.

Mercier, P. (1991) Analyses des relations espèces-environnement et étude de la co-structure d'un couple de tableaux. Thèse de doctorat, Université Lyon 1.

Milési, J.P., Egal, E., & Ledru, P. (1995) Les minéralisations du nord de la Guyane Française dans leur cadre géologique. *Chronique de la recherche minière*, 518, 5-58.

Mitter, C., Farrell, B., & Wiegmann, B. (1998) The phylogenetic study of adaptive zones: has phytophagy promoted insect diversification? *The American Naturalist*, 132, 107-128.

Mom, A. (1998) Eigenstructure of distance matrices with an equal distance subset. *Linear Algebra and its Applications*, 280, 245-251.

Monestiez, P. (1978). Méthodes de classification automatique sous contraintes spatiales. In *Biométrie et Ecologie* (eds J.M. Legay & R. Tomassone), pp. 367-379. Société Française de Biométrie, Paris.

Monestiez, P., Goulard, M., & Charmet, G. (1994) Geostatistics for spatial genetic structures: study of wild populations of perennial ryegrass. *Theoretical and applied genetics*, 88, 33-41.

Moran, P.A.P. (1948) The interpretation of statistical maps. *Journal of the Royal Statistical Society, B*, 10, 243-251.

Moran, P.A.P. (1950) Notes on continuous stochastic phenomena. *Biometrika*, 37, 17-23.

Mugglestone, M.A. & Renshaw, E. (1996) A practical guide to the spectral analysis of spatial point processes. *Computational Statistics & Data Analysis*, 21, 43-65.

Nelson, R. (1988) Using airborne laser data to estimate forest canopy and stand characteristics. *Journal of Forestry*, 86, 31-38.

Nielsen, A.A. (1995a) Change detection in multi-spectral, bi-temporal spatial data using orthogonal transformations. In <http://citeseer.nj.nec.com/63505.html>.

Nielsen, A.A. (1995b) Multi-channel remote sensing data and orthogonal transformations for change detection. In <http://citeseer.nj.nec.com/56095.html>.

Nielsen, A.A. (1999) C04351 Statistical Image Analysis, Spring 1999 Orthogonal Transformations. In <http://citeseer.nj.nec.com/428248.html>.

Nielsen, A.A. & Conradsen, K. (1997) Multivariate alteration detection (MAD) in multispectral, bi-temporal image data: a new approach to change detection studies. In <http://www.imm.dtu.dk/~aa/tech-rep-1997-11/>. Tech. rep. 199711, Department of Mathematical Modelling, Technical University of Denmark.

Nielsen, A.A., Conradsen, K., Pedersen, J.L., & Steinfeld, A. (1997) Spatial factor analysis of stream sediment geochemistry data from South Greenland. In *Proceedings of the Third Annual Conference of the International Association for Mathematical Geology* (ed V. Pawlowsky-Glahn), pp. 955-960, Barcelona, Spain.

Nielsen, A.A., Conradsen, K., & Simpson, J.J. (1998) Multivariate alteration detection (MAD) and MAF post-processing in multispectral, bi-temporal image data: new approaches to change detection studies. *Remote Sensing of Environment*, 64, 1-19.

Nielsen, A.A. & Larsen, R. (1994) Restoration of Geris data using the maximum noise fractions transform. In *First International Airborne Remote Sensing Conference and Exhibition*, Strasbourg, France, 11–15 September 1994.

Noy-Meir, I. & Anderson, D.J. (1971). Multivariate pattern analysis, or multiscale ordination: towards a vegetation hologram ? In *Statistical Ecology, III Many species populations ecosystems and systems analysis* (eds G.P. Patil, E.C. Pielou & W.E. Waters), pp. 208-231. Pennsylvania State University Press.

Ollier, S., Chessel, D., Couteron, P., Péliissier, R., & Thioulouse, J. (2003) Comparing and classifying one-dimensional spatial patterns: an application to laser altimeter profiles. *Remote Sensing of Environment*, 85, 453-462.

Ollier, S., Couteron, P., & Chessel, D. (sous presse) Orthonormal transforms to describe and test the phylogenetic signal. *Biometrics*.

Ollier, S., Dray, S., & Chessel, D. (soumis) Taking into account spatial dependence in multivariate analysis: a generalization of Wartenberg's multivariate spatial correlation. *Geographical Analysis*.

Orr, M.R. & Smith, T.B. (1998) Ecology and speciation. *Trends in Ecology and Evolution*, 13, 502-506.

Pace, R.K. & Barry, R. (1997) Sparse spatial autoregressions. *Statistics and Probability Letters*, 33, 291-297.

Pace, R.K. & LeSage, J.P. (2002) Semiparametric maximum likelihood estimates of spatial dependence. *Geographical Analysis*, 34, 76-90.

Pace, R.K. & LeSage, J.P. (2003) Conditional autoregressions with doubly stochastic weight matrices.

Pace, R.K. & Zou, D. (2000) Closed-form maximum likelihood estimates of nearest neighbor spatial dependence. *Geographical Analysis*, 32, 154-172.

Palmeira, L. (2003-2004). Influence des substitutions dépendantes du voisinage sur les méthodes reconstruction phylogénétique, Lyon.

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

Percival, D. (1993) An introduction to spectral analysis and wavelets, *International workshop on advanced mathematical tools in metrology*.

Percival, D. (2003). Wavelets. In *Encyclopedia of Environmetrics* (eds A.H. El-Shaarawi & W.W. Piegorisch). John Wiley & Sons, Ltd, Chichester.

- Percival, D.B.** (1995). On Estimation of the Wavelet Variance.
- Percival, D.B. & Walden, A.T.** (2000) Wavelet Methods for Time Series Analysis Cambridge University Press.
- Perrière, G. & Gouy, M.** (1996) WWW-Query: An on-line retrieval system for biological sequence banks. *Biochimie*, 78, 364-369.
- Perry, J.N., Liebhold, A.M., Rosenberg, M.S., Dungan, J., Miriti, M., Jakomulska, A., & Citron-Pousty, S.** (2002) Illustrations and guidelines for selecting statistical methods for quantifying spatial patterns in ecological data. *Ecography*, 25, 578-600.
- Podos, J.** (2001) Correlated evolution of morphology and vocal signal structure in Darwin's finches. *Nature*, 409, 185-188.
- Poizat, G. & Pont, D.** (1996) Multi-scale approach to species-habitat relationships: juvenile fish in a large river section. *Freshwater Biology*, 36, 611-622.
- Priestley, M.B.** (1981) Spectral analysis and time series Academic Press, London.
- Qu, Y., Adam, B., Thornquist, M., Potter, J.D., Thompson, M.L., Yasui, Y., Davis, J., Schellhammer, P.F., Cazares, L., Clements, M.A., Wright, G.L., & Feng, Z.** (2003) Data reduction using a discrete wavelet transform in discriminant analysis of very high dimensionality data. *Biometrics*, 59, 143-151.
- Rao, C.R.** (1982) Diversity and dissimilarity coefficients: a unified approach. *Theoretical Population Biology*, 21, 24-43.
- Renshaw, E.** (1997) Spectral techniques in spatial analysis. *Forest Ecology and Management*, 94, 165-174.
- Renshaw, E.** (2002) Two-dimensional spectral analysis of marked point processes. *Biometrical Journal*, 44, 718-745.
- Ripley, B.D.** (1978) Spectral analysis and the analysis of pattern in plant communities. *Journal of Ecology*, 66, 965-981.
- Ritchie, J.C., Evans, D.L., Jacobs, D., Everitt, J.H., & Wertz, M.A.** (1993) Measuring canopy structure with an airborne laser altimeter. *Transaction of the ASAE*, 36, 1235-1238.
- Rochet, M.J., Cornillon, P.-A., Sabatier, R., & Pontier, D.** (2000) Comparative analysis of phylogenetic and fishing effects in life history patterns of teleost fishes. *Oikos*, 91, 255-270.
- Rohlf, F.J.** (2001) Comparative methods for the analysis of continuous variables: geometric interpretations. *Evolution*, 55, 2143-2160.

Royer, J.J. (1984) Proximity analysis: a method for multivariate geodata processing. Application to geochemical processing. *Sciences de la Terre, Série informatique* 20, 585-591.

Sanderson, M.J., Baldwin, B.G., Bharatan, G., Campbell, C.S., Ferguson, D., Porter, C., Von Dohlen, C., Wojciechowski, M.F., & Donoghue, M.J. (1993) The growth of phylogenetic information and the need for a phylogenetic database. *Systematic Biology*, 42, 562-568.

Sanderson, M.J. & Donoghue, M.J. (1996) Reconstructing shifts in diversification rates on phylogenetic trees. *Trends in Ecology and Evolution*, 11, 15-20.

Sanderson, M.J. & Donoghue, M.J. (1998) Phylogenetic supertrees: assembling the trees of life. *Trends in Ecology and Evolution*, 13, 105-109.

Sandjivy, L. & Galli, A. (1984) Analyse krigéante et analyse spectrale. *Science de la Terre, Série Informatique*, 21, 115-124.

Schuster, A. (1898) On the Investigation of Hidden Periodicities with Application to a Supposed 26 Day Period of Meteorological Phenomena. *Terrestrial Magnetism*, 3, 13-41.

Smith, A.B., Littlewood, D.T.J., & Wray, G.A. (1996). Comparative evolution of larval and adult life-history stages and small subunit ribosomal RNA amongst post-Palaeozoic echinoids. In *New Uses for New Phylogenies* (eds P.H. Harvey, A.J. Leigh Brown, J. Maynard Smith & S. Nee). Oxford University Press, Oxford.

Smouse, P. & Peakall, R. (1999) Spatial autocorrelation analysis of individual multiallele and multilocus genetic structure. *Heredity*, 82, 561-573.

Sokal, R.R. (1979). Ecological parameters inferred from spatial correlograms. In *Contemporary quantitative ecology and related econometrics* (eds G.P. Patil & M. Rosenzweig), pp. 167-196. International Co-operative Publishing House, Fairland.

Sokal, R.R. & Rohlf, F.J. (1969) *Biometry* Third edition. W.H. Freeman and Company, New-York.

Solow, A.R. (1994) Detecting change in the composition of a multispecies community. *Biometrics*, 50, 556-565.

Statzner, B., Hoppenhaus, K., Arens, M.-F., & Richoux, P. (1997) Reproductive traits, habitat use and templet theory: a synthesis of world-wide data on aquatic insects. *Freshwater Biology*, 38, 109-135.

Stokes, G.G. (1879) Note on Searching for Periodicities. *Proceedings of the Royal Society for Industrial and Applied Mathematics*, 29, 122.

St-Onge, B. (1999) Estimating individual tree heights of the boreal forest using airborne laser altimetry and digital videography. In *Workshop on mapping surface structure and topography by airborne and spaceborne lasers*, Vol. reference 28. ISPRS, Lajolla, Californie.

St-Onge, B.A., Couture, M., & Alleaume, S. (1998) Forest stand structure mapping using a species-controlled textural approach. In International Forum on Automated Interpretation of High Spatial Resolution Digital Imagery for Forestry. in press, Victoria.

Switzer, P. & Green, A.A. (1984). Min/max autocorrelation factors for multivariate spatial imagery. Tech. rep. 6, Stanford University.

Tenenhaus, M. & Young, F.W. (1985) An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical multivariate data. *Psychometrika*, 50, 91-119.

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

Tiefelsdorf, M., Griffith, D.A., & Boots, B. (1999) A variance-stabilizing coding scheme for spatial link matrices. *Environment and Planning A*, 31, 165-180.

Tisné-Agostini, D. (1988) Description par analyse en composantes principales de l'évolution de la production du clémentinier en association avec 12 types de porte-greffe. Rapport technique, DEA Analyse et modélisation des systèmes biologiques, Université Lyon 1.

Upton, G. & Fingleton, B. (1985) Spatial data analysis by example. Vol. 1: Point pattern and quantitative data John Wiley & Sons, Chichester.

Vaidyanathan, P.P. (1993) Multirate Systems and Filter Banks Prentice-Hall, New Jersey.

Ver Hoef, J.M. & Glenn-Lewin, C.G. (1989) Multiscale ordination: a method for detecting pattern at several scales. *Vegetatio*, 82, 59-67.

Ver Hoef, J.M., Cressie, N.A.C., & Glenn-Lewin, D.C. (1993) Spatial models for spatial statistics: some unification. *Journal of Vegetation Science*, 4, 441-452.

Vitt, L.J., Zani, P.A., & Esposito, M.C. (1999) Historical ecology of Amazonian lizards : implications for community ecology. *Oikos*, 87, 286-294.

Wackernagel, H. (2003) Multivariate geostatistics. An introduction with applications, Third edition edn. Springer.

Wagner, H.H. (2003) Spatial covariance in plant communities: integrating ordination, geostatistics, and variance testing. *Ecology*, 84, 1045-1057.

Wagner, H.H. (2004) Direct multi-scale ordination with canonical correspondence analysis. *Ecology*, 85, 342-351.

Wartenberg, D.E. (1985) Multivariate spatial correlations: a method for exploratory geographical analysis. *Geographical Analysis*, 17, 263-283.

Watt, A.S. (1947) Pattern and process in plant community. *Journal of Ecology*, 35, 1-22.

Weishampel, J., Sun, G., & Harding, D.J. (1996) Remote sensing of forest canopies. *Selbyana*, 17, 6-14.

Yoccoz, N. (1988) Le rôle du modèle euclidien d'analyse des données en biologie évolutive. Thèse de doctorat, Université Lyon 1.