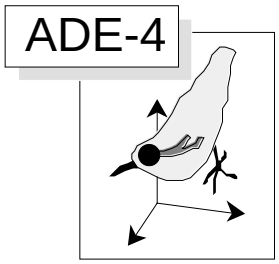


Autocorrélation et composantes cartographiables

Résumé

La fiche donne quelques illustrations des notions élémentaires de variance locale et d'autocorrélation spatiale à la base de l'analyse des structures spatiales. Ces deux notions utilisent des graphes de voisinage. Le test de Geary, l'AFC des matrices de graphe, la diagonalisation des opérateurs de Moran et l'analyse multivariée basée sur la maximisation de l'autocorrélation sont des éléments préparatoires aux méthodes multimétriques.

Plan

0 — Expérience préliminaire 2

1 — Les comtés d'Irlande..... 4

2 — Variance locale et test de Geary..... 7

3 — Autocorrélation et opérateur de Moran..... 11
- 3.1 — L'indice de Moran 11
- 3.2 — Opérateurs de voisinages..... 12
- 3.4 — AFC et graphe de voisinage 13
- 3.5 — Vecteurs propres de voisinage 16

4 — L'analyse globale..... 18

5 — Conclusion 22

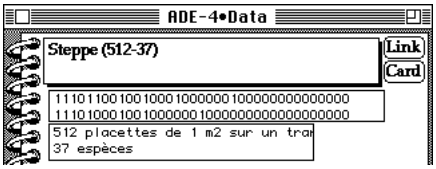
Références 24

D. Chessel, J. Thioulouse et S. Champely

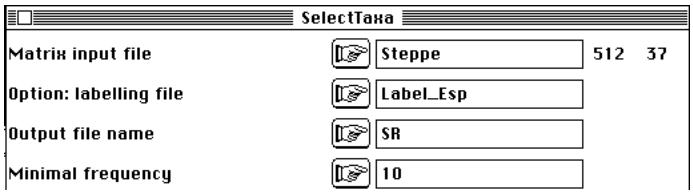
0 — Expérience préliminaire

L'ordination sous contrainte spatiale est un sujet un peu paradoxal. La majorité des observations écologiques sont référencées au temps et à l'espace. L'information spatiale (mode de dispersion dans l'espace concret des points de mesure) entre cependant rarement, de façon explicite, dans le traitement des données, bien qu'elle apparaisse dans nombre d'études au moment de l'interprétation. Le premier article sur l'ACP en écologie (Goodall 1954 [1]) comme l'un des premiers articles sur l'AFC en écologie (Hatheway 1971 [2]) cartographie des coordonnées factorielles et notent l'efficacité de cette pratique. Hill 1974 [3] comme Estève 1978 [4] représente les coordonnées factorielles le long d'un transect tout comme Dessier et Laurec 1978 [5] 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.

Aller à la carte Steppe de la pile ADE-4•Data :



Faire avec le champ de gauche un fichier Steppe.Car. Le passer en binaire (Steppe 512-37). Utiliser EcolTools : SelectTaxa pour éliminer les espèces rares. Le programme édite par ordre croissant le nombre de présences par espèce. Si on ne garde que les espèces présentes au moins 10 fois, il ne conserve que 24 colonnes dans le fichier de sortie (SR 512-24).

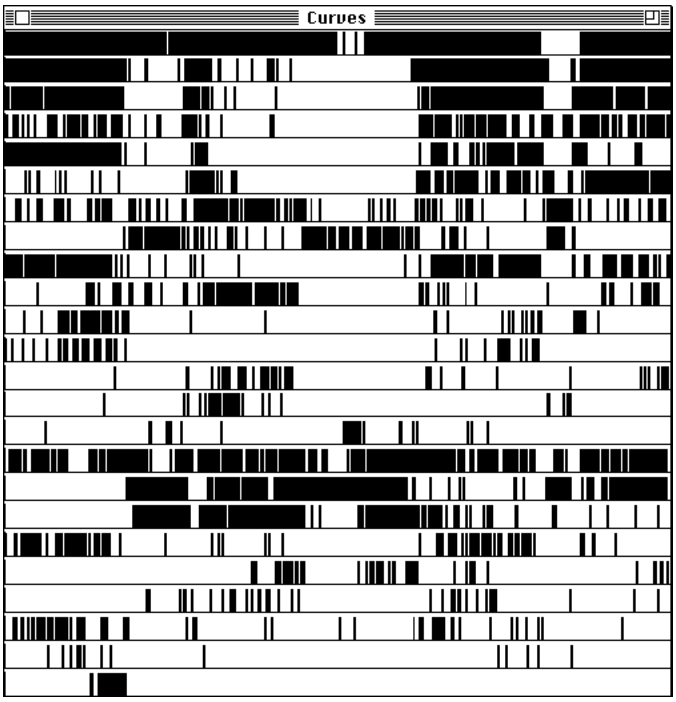



```
Input data file: Steppe  Rows: 512   Col: 37
Input label file: Label_Esp

Output data file: SR  Rows: 512   Col: 24
Output label file: SR_label
Selected taxa
-------------------------------------------------------------
|New=Former |New=Former |New=Former |New=Former |New=Former |
-------------------------------------------------------------
|   1 =   1 |   2 =   2 |   3 =   3 |   4 =   4 |   5 =   5 |
|   6 =   6 |   7 =   7 |   8 =   8 |   9 =   9 |  10 =  10 |
|  11 =  11 |  12 =  12 |  13 =  13 |  14 =  14 |  15 =  15 |
|  16 =  16 |  17 =  17 |  18 =  18 |  19 =  19 |  20 =  20 |
|  21 =  21 |  22 =  22 |  23 =  23 |  24 =  24 |
-------------------------------------------------------------
```

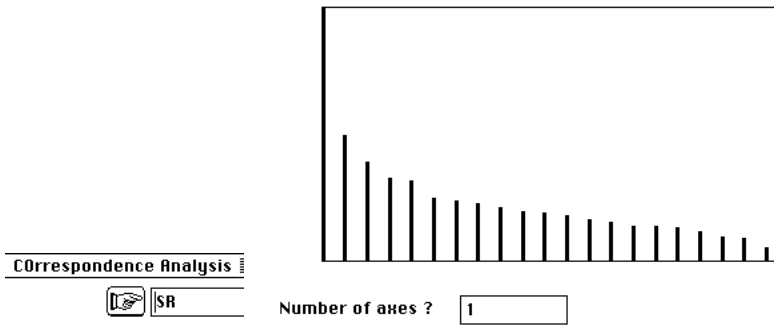
```
1- Noaea mucronata        13- Fagonia kaherina
2- Plantogo albicans         14- Lygeum spartum
3- Hernaria fontananensii       15- Peganum harmala
4- Stipa parviflora          16- Koeleria pubescens
5- Helianthemum hirtum       17- Stipa retorta
6- Poa bulbosa            18- Anacyclus clavatus
7- Anabasis Oropediorum      19- Stipa barbata
8- Salsola vermiculata       20- Schismus barbatus
9- Atractylis serratuloïdes     21- Bromus rubens
10- Artemisia Herba Alba     22- Echium humile
11- Pithuranthos scoparium      23- Hypocrepis scabra
```

12- Teucrium polium 24- Hedysorum spinosissimum

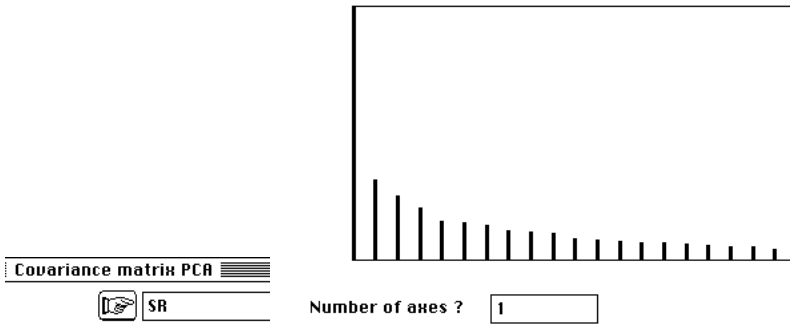
Éditer le tableau SR (512 placettes alignées sur un transect de 5 km, présence-absence de 24 espèces végétales d'une steppe semi-aride) avec Curves : Bars**.**



Faire l'AFC de ce tableau :

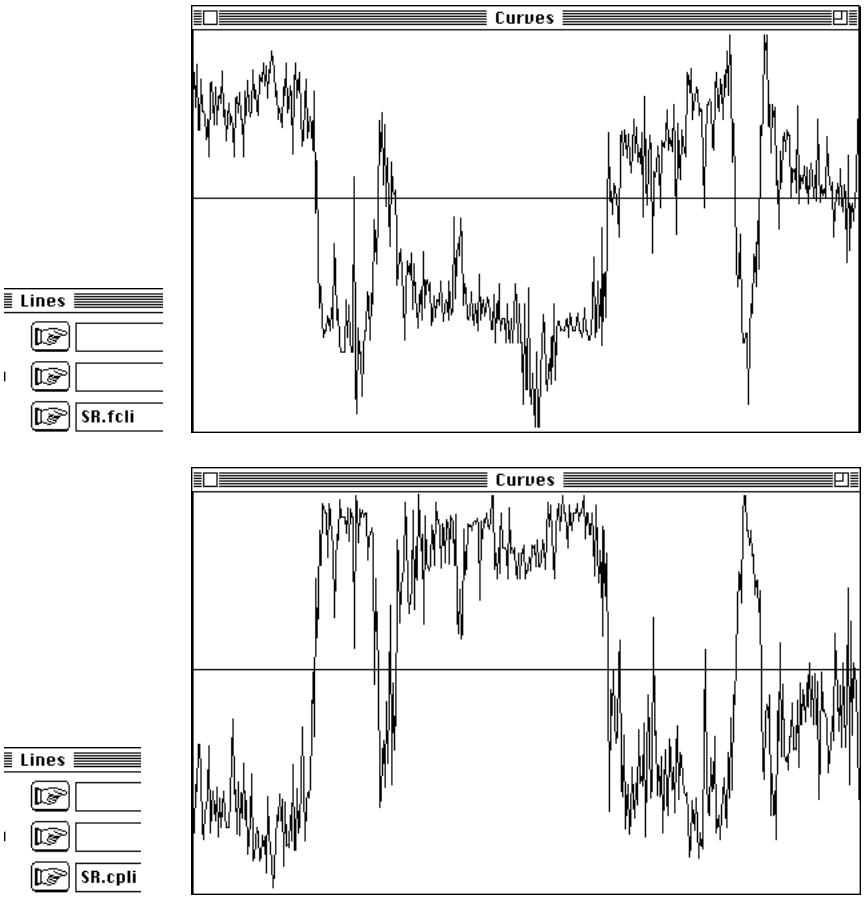


Faire ensuite l'ACP centrée de ce même tableau :



Utiliser Curves : Lines pour restituer l'évolution, le long du transect, de la première coordonnée de chaque analyse. On notera l'étroite similitude des deux résultat 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

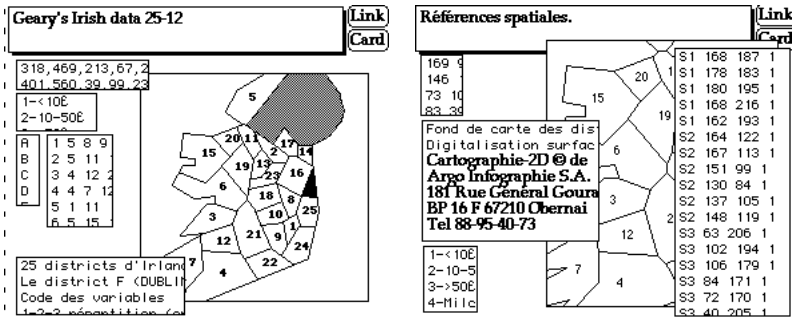
spatiale sous-jacente intervenait directement, alors qu'il n'en est rien. Cette intervention active, qui semble dans le cas présent inutile, est le fait des méthodes d'ordination locale et globale.



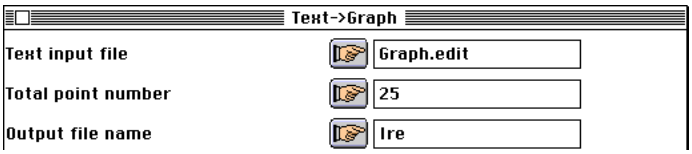
Si les points de mesure se suivent le long d'un transect le seul numéro d'ordre des lignes des tableaux de données contient toute l'information de proximité entre points, qu'on s'en serve ou non. Dans tous les autres cas cette information doit être intégrée explicitement.

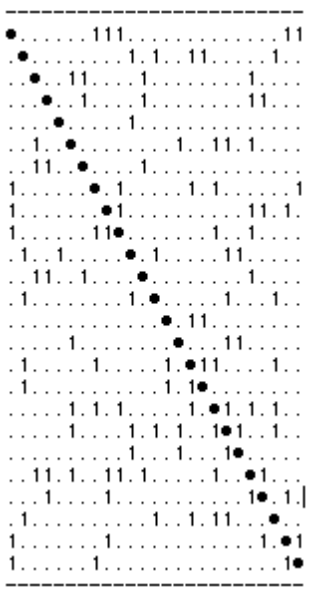
1 — Les comtés d'Irlande

On utilisera pour les illustrations un des jeux de données les plus célèbres de la statistique spatiale, celui des comtés d'Irlande de Geary (1954)[6]. Le matériel nécessaire est donné dans les cartes Irlande et Irlande+1 de la pile ADE-4•Data.

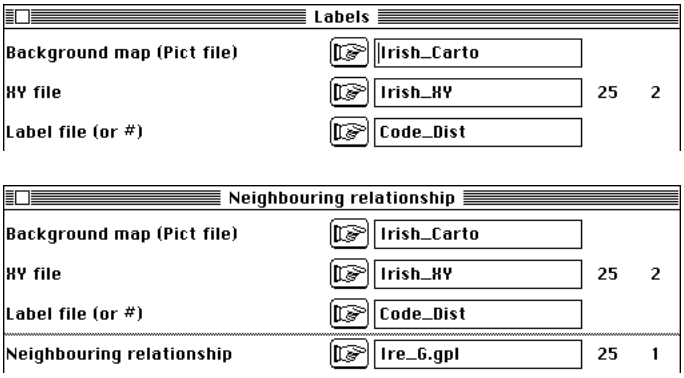


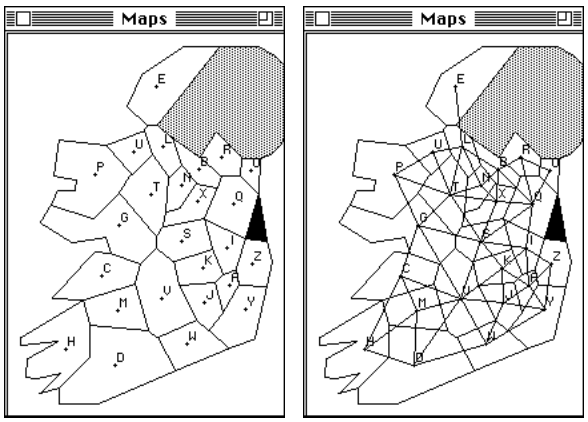
Récupérer tous les fichiers dans le dossier de travail. Lire le graphe de voisinage (NGUtil: Text->Graph) :



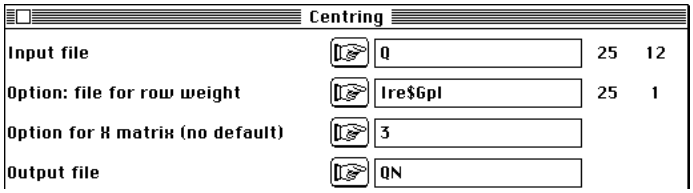


La matrice de voisinage **M** indique à la ligne *i* et à la colonne *j* si le comté *i* est voisin (1) ou non voisin (0) du comté *j*. Labelliser le fond de carte et vérifier la cohérence de l'information en plaçant la relation de voisinages sur la carte (Maps) :

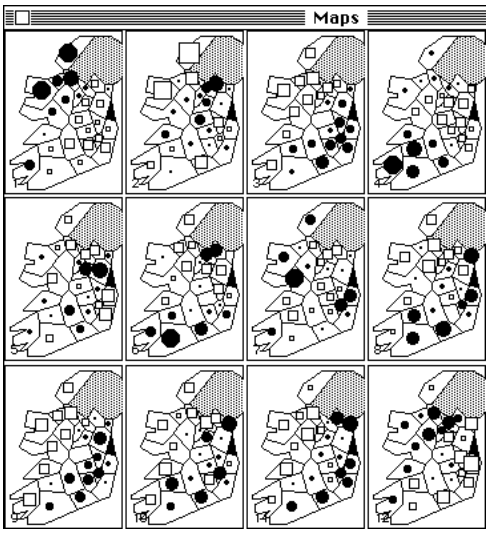


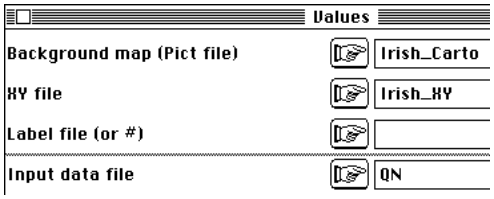


Normaliser les données avec Bin->Bin : Centring :

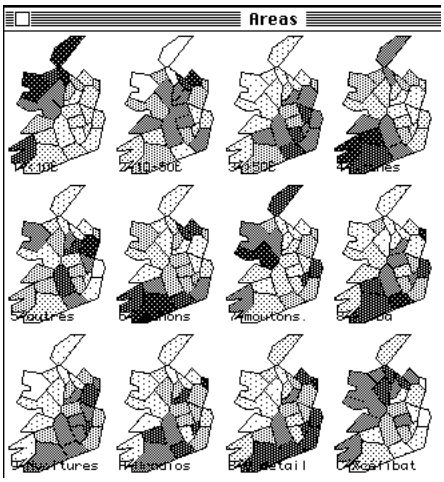


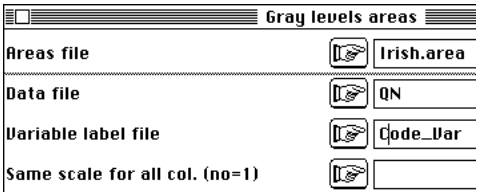
Cartographier par valeurs ponctuelles :



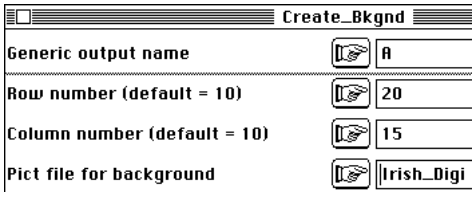


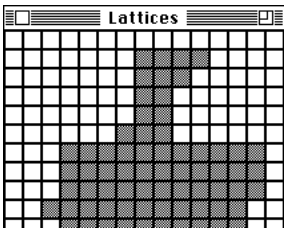
Cartographier encore par unités surfaciques :

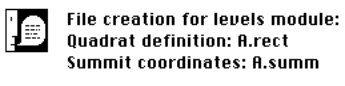


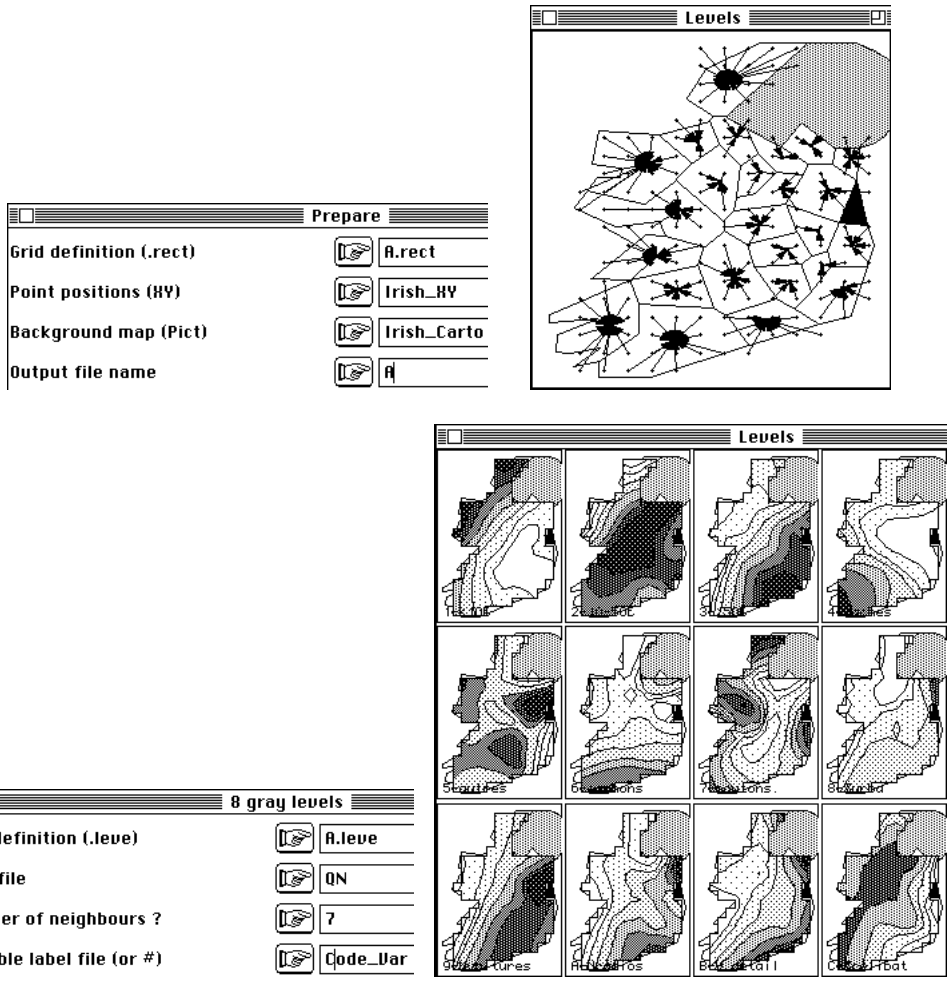


Implanter une grille pour les courbes de niveaux :







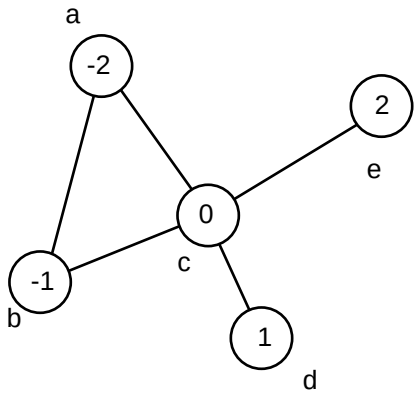


L'impact visuel des courbes de niveau étant très supérieur à celui des autres représentations, on doit veiller à un usage contrôlé. L'option GraphUtil : 2DLowess Residuals de permet de connaître les résidus de prédiction en chaque point (rapportés à l'écart-type de départ). On noterait en l'utilisant que l'erreur de prédiction est systématiquement forte, pour toutes les variables, sur le district isolé du nord.

2 — Variance locale et test de Geary

L'ouvrage classique de Cliff & Ord [7] présente deux tests de signification de la structure spatiale d'une variable. Le premier est celui de l'indice de Geary. Il utilise la notion de graphe de voisinage.

Pour comprendre la signification de cet indice une réécriture de la notion de variance est indispensable. Elle a été faite par Lebart [8] et le procédé a été utilisé indépendamment par Light & Margolin [9] 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 figurent 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

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

En général pour n observation $x_1, \cdots, x_i \cdots x_n$ de poids $p_1, \cdots, p_i \cdots p_n$ la variance est définie par :

$$\bar{x} = \sum_{i=1}^{n} p_i x_i \text{ et } Var = \sum_{i=1}^{n} p_i \left(x_i - \bar{x}\right)^2$$

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

	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 :

$$Var = (1/2) \sum_{i=1}^{n} \sum_{j=1}^{n} p_i p_j \left(x_i - x_j\right)^2$$

On retiendra la relation fondamentale :

$$\sum_{i=1}^{n} \sum_{j=1}^{n} p_i p_j \left(x_i - x_j\right)^2 = 2 \sum_{i=1}^{n} p_i \left(x_i - \bar{x}\right)^2 \text{ [R1]}$$

L'intérêt de cette observation est de séparer les couples de points en deux catégories, les couples de voisins d'une part, les couples de non voisins de l'autre.

	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 somme des carrés des différences (100) se décompose en somme sur les couples de voisins (22) et somme sur les couples de non voisins (78). La variance (100/50) se décompose en deux parties (22/50 et 78/50) appelées respectivement variance locale (entre voisins) et variance globale (entre non voisins). En général :

$$Var = (1/2)\sum_{i,j} p_i p_j \left(x_i - x_j\right)^2 = (1/2) \sum_{i \text{ voisin } j} p_i p_j \left(x_i - x_j\right)^2 + (1/2) \sum_{i \text{ non voisin} j} p_i p_j \left(x_i - x_j\right)^2$$

$$Var = Var_{loc} + Var_{glo}$$

Ce point de vue a l'avantage de la simplicité et un inconvénient issu du fait que dans la plupart des cas une écrasante majorité de couples sont des couples de non voisins. La variance locale représente alors une toute petite partie de la variance totale.

Il y a plusieurs manières de se servir de cette observation. La première en date sert à tester la signification de cette variance locale pour une variable donnée. C'est l'indice de Geary. On note sur l'exemple, que, puisque la variance totale est la moyenne pour les 25 couples des carrés des différences, mais que seulement 20 couples sont utiles (les 5 autres valeurs sont forcément nulles). Il vaut donc mieux considérer que la variance est la moyenne sur les couples utiles. Ici la pondération est uniforme ($p_i = 1/n$) :

$$\hat{V} = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 = \frac{1}{2n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(x_i - x_j\right)^2$$

Dans l'exemple, on obtient 100/40, soit 2.5. On retrouve l'estimateur habituel d'une variance. On peut se demander si la moyenne des carrés des différences sur l'ensemble des couples voisins seulement n'est pas un autre estimateur de cette variance :

$$\hat{V}_{loc} = \frac{1}{2m}\sum_{i \text{ voisin } j}^{n}\left(x_i - x_j\right)^2$$

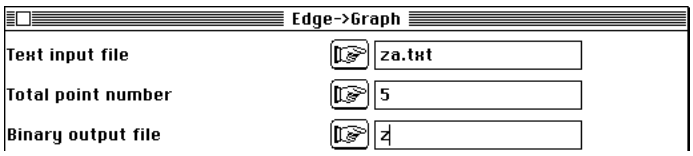
où m désigne le nombre de couples de voisins (chaque paire est comptée deux fois, un point n'étant jamais voisin de lui même). Dans l'exemple m vaut 10 et la quantité 22/20. Si il n'y a pas de structure spatiale les valeurs des carrés des différences entre voisins sont en moyenne les mêmes que sur l'ensemble des couples. On s'attend à ce que le rapport de la variance estimée localement sur la variance estimée totalement soit égal à 1 ou encore que :

$$c = \frac{\hat{V}_{loc}}{\hat{V}} \qquad I_G = \frac{1-c}{Ect(c)}$$

ne soit pas significativement de 0. La quantité c est le coefficient de contiguïté de Geary et I_G est la valeur normalisée de $-c$. Le dénominateur est l'écart-type du rapport des

variances estimées. Il est connu dans deux modèles, respectivement N (les observations sont un échantillons d'une loi normale indépendante de la structure spatiale) et R (les observations sont un cas arbitraire parmi les *n*! possibilités de placer les valeurs observées sur les *n* points de la structure spatiale). Dans l'exemple le rapport des variances estimées encore noté *c* vaut 1.1/2.5 soit 0.44. Implanter les valeurs -2, -1, 0, 1, 2 dans un fichier binaire ZQ (5-1). Implanter la liste des arêtes (1 2, 1 3, 2 3, 3 4, 3 5) dans un fichier texte.

Exécuter NGUtil : Edge->Graph :




```
Neighboorhood relation reading
Input text file: za.txt
Neighboorhood graph in binary file: z$G
It contains graph matrix (LEBART's M) with 5 rows and columns
Neighboorhood weights in binary file: z$Gpl
It contains  5 rows and 1 column
Symetric neighboorhood matrix (. for 0):
-----
●11..
1●1..
11●11
..1●.
..1.●
-----
```

Lancer NGStat : Geary Test :


```
Geary Autocorrelation test
Neighboorhood graph: z$G
Data matrix: zq
Point number: 5
|---|-----------|-----------|----------|-----------|----------|
| N°| c observ. | Test N    |  Proba   | Test R    |  Proba   |
|---|-----------|-----------|----------|-----------|----------|
|  1|  4.400e-01|  1.462e+00| 7.184e-02|  1.689e+00| 4.556e-02|
|---|-----------|-----------|----------|-----------|----------|
```

On retrouve la valeur de l'indice c (.44) et son niveau de signification dans les deux modèles. On ne peut accorder de signification à ces tests que pour des valeurs suffisantes du nombre de points ($n = 20$).

Exécuter le test de Geary sur les données d'Irlande :


```
Geary Autocorrelation test
Neighboorhood graph: Irish$G
Data matrix: Q
Point number: 25
l---l-----------l-----------l----------l-----------l----------l
l N°l c observ. l Test N    l  Proba   l Test R    l  Proba   l
l---l-----------l-----------l----------l-----------l----------l
|  1|  3.477e-01|  4.314e+00| 8.014e-06|  3.902e+00| 4.779e-05|
|  2|  5.840e-01|  2.751e+00| 2.970e-03|  2.376e+00| 8.750e-03|
|  3|  3.925e-01|  4.018e+00| 2.939e-05|  4.719e+00| 1.184e-06|
```

4	3.418e-01	4.353e+00	6.716e-06	3.771e+00	8.135e-05
5	1.026e+00	-1.707e-01	5.678e-01	-1.668e-01	5.662e-01
6	6.533e-01	2.293e+00	1.093e-02	2.080e+00	1.875e-02
7	8.686e-01	8.689e-01	1.925e-01	7.387e-01	2.300e-01
8	6.148e-01	2.547e+00	5.425e-03	2.590e+00	4.796e-03
9	5.124e-01	3.225e+00	6.306e-04	3.599e+00	1.597e-04
10	8.141e-01	1.229e+00	1.095e-01	1.235e+00	1.085e-01
11	5.267e-01	3.130e+00	8.733e-04	3.350e+00	4.045e-04
12	6.465e-01	2.338e+00	9.689e-03	2.552e+00	5.357e-03

On a encadré les résultats de la variables 4, qui sont conformes à ceux de Cliff & Ord (1973 op. cit. page 57). La confrontation de ces statistiques au cartes des variables s'impose. La question sera alors clairement posée : devant une série de cartes plus ou moins simples : comment faire leur lecture simultanée, leur synthèse, voir leur ordination ou leur classification en plusieurs types ? On utilise pour répondre à cette question les opérateurs de voisinages [10]que nous appelons aussi opérateurs de Moran.

3 — Autocorrélation et opérateur de Moran

3.1 — L'indice de Moran

La notion d'autocorrélation spatiale mesure essentiellement la ressemblance entre voisins. L'idée est initialement celle de Moran (1948)**[11]**. L'indice d'autocorrélation spatiale de Moran est décrit dans l'ouvrage de base de Cliff & Ord, en parallèle avec l'indice de Geary qui a une fonction voisine. Utiliser les tests de Moran ou ceux de Geary donne des résultats voisins. Si un test d'autocorrélation est nécessaire on utilisera donc celui de Geary. Mais la différence des principes de base est sensible.

L'indice de Geary dit si la variabilité entre points voisins est plus petite, significativement, qu'attendue d'un modèle aléatoire. L'indice de Moran dit si la ressemblance entre points voisins est plus grande, significativement, qu'attendue d'un modèle aléatoire. On comprend bien que la nuance n'est pas fondamentale. Par contre, les analyses locales, basées sur l'indice de Geary, cherchent la structure de la variance entre points voisins. Les analyses basées sur l'indice de Moran cherchent, à l'inverse, la structure de la ressemblance entre voisins. La nuance s'apparente à une antinomie complète d'objectifs.

La difficulté vient de ce que la variance de voisinage est une forme quadratique et a été intégrée naturellement en analyse des données. La notion d'autocorrélation spatiale ne l'est pas. Son intégration en analyse multivariée n'est pas naturelle. Tentée par Wartenberg (1985c) **[12]**, cette insertion n'est pas optimum du point de vue mathématique, tout en étant très légitime du point de vue expérimental. On rapprochera cette tentative des travaux du même auteur pour utiliser l'autocorrélation spatiale dans l'interprétation d'une analyse ordinaire (Wartenberg 1985b)**[13]** et pour approfondir l'usage des coordonnées concrètes dans l'espace comme données numériques (Wartenberg 1985a)**[14]**.

L'indice de Moran est défini, dans les notations de paragraphe 2 par :

$$I = \frac{n \sum_{i \text{ voisin} j} (x_i - \bar{x})(x_j - \bar{x})}{m \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

On reconnaît la moyenne pour les couples de voisins des quantités $(x_i - \bar{x})(x_j - \bar{x})$ rapportée à la moyenne des quantités $(x_i - \bar{x})^2$. La variance totale qui intervient dans le *c* de Geary est donc la variance estimée (calculée avec *n* -1) et celle qui intervient dans le *I* de Moran est la variance descriptive (calculée avec *n*). Il ne s'agit pas d'une imprécision, bien au contraire. Les deux indices ont la même logique dans deux cadres complémentaires. On notera toujours **M** la matrice à *n* lignes et *n* colonnes dite matrice de voisinage où $m_{ij} = 1$ si *i* et *j* sont voisins, $m_{ij} = 0$ dans le cas contraire.

$$\mathbf{M} = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

Cette structure de voisinage en appelle deux qui servent de références, respectivement :

$$\mathbf{U}_n - \mathbf{I}_n = \begin{bmatrix} 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 \end{bmatrix} \qquad \mathbf{I}_n = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

On notera toujours $\mathbf{U}_n$ la matrice à *n* lignes et *n* colonnes dont tous les éléments sont égaux à 1. Dans la première un point est voisin de tous les autres sauf de lui-même, dans la seconde un point n'est voisin que de lui-même. Alors :

$$c = \frac{\frac{1}{2m}\sum_{(i,j)\text{voisins}[\mathbf{M}]}(x_i - x_j)^2}{\frac{1}{2n(n-1)}\sum_{(i,j)\text{voisins}[\mathbf{U}_n - \mathbf{I}_n]}(x_i - x_j)^2} \qquad I = \frac{\frac{1}{m}\sum_{(i,j)\text{voisins}[\mathbf{M}]}(x_i - \bar{x})(x_j - \bar{x})}{\frac{1}{n}\sum_{(i,j)\text{voisins}[\mathbf{I}_n]}(x_i - \bar{x})(x_j - \bar{x})}$$

Dans le premier cas, la variance est la variabilité moyenne entre deux points (référence pour la variabilité de voisinage), dans le second cas, c'est la covariance de la variable avec elle-même (référence pour la covariance de voisinage). Comment rendre totalement compatibles ces deux notions si voisines ?

3.2 — Opérateurs de voisinages

L'indice de Geary, contrairement à l'indice de Moran, semble supprimer toute notion de moyenne. En outre il est, comme rapport de somme de carrés, toujours positifs. C'est pourquoi, il conduira Lebart à l'introduction des métriques de voisinage. 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 :

$$\frac{1}{m}\sum_{(i,j)\text{voisins}[\mathbf{M}]}(x_i-\alpha)(x_j-\alpha)$$

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

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

où *mv* désigne la moyenne de voisinage de la variable **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 :

$$I^*=\frac{\frac{1}{m}\sum_{(i,j)\text{voisins}[\mathbf{M}]}(x_i-mv(\mathbf{x}))(x_j-mv(\mathbf{x}))}{\frac{1}{n}\sum_{(i,j)\text{voisins}[\mathbf{I}_n]}(x_i-mv(\mathbf{x}))(x_j-mv(\mathbf{x}))}$$

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

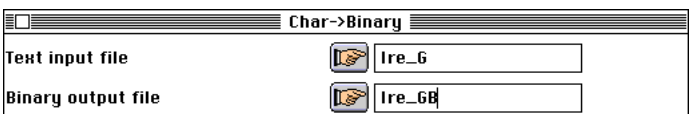
$$c^*=\frac{\frac{1}{2m}\sum_{(i,j)\text{voisins}[\mathbf{M}]}(x_i-x_j)^2}{\frac{1}{2n^2}\sum_{(i,j)\text{voisins}[\mathbf{U}_n]}(x_i-x_j)^2}$$

alors on a simplement $I^*+c^*=1$.

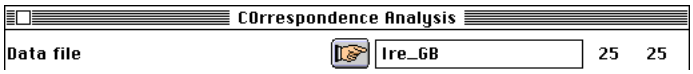
3.4 — AFC et graphe de voisinage

Lebart (in Benzecri 1973 **[15]** p. 244-261 et 1984 **[16]**) discute des propriétés de l'analyse des correspondances de la matrice du graphe de voisinage. Elles sont aisées d'accès.

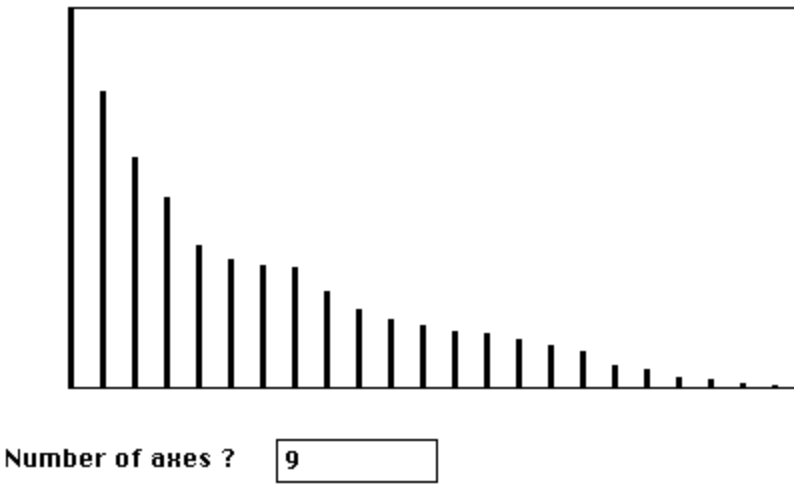
Passer d'abord la matrice du graphe de voisinages en binaire :



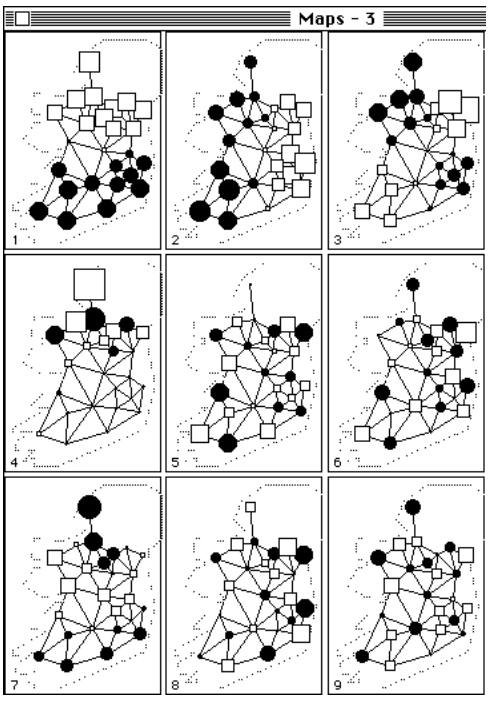
Exécuter l'analyse avec le module COA :

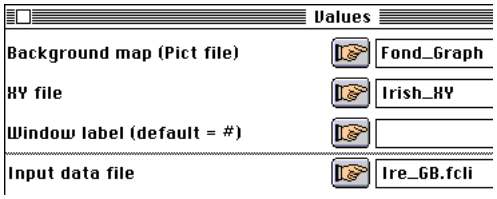


On note que la somme totale du tableau vaut 106, ce qui signifie qu'il y a 53 arêtes dans le graphe.



On garde, sans plus de raisons, les 9 premiers facteurs. Depuis les travaux de Williams (1952)[17] , on sait qu'on vient de trouver des codes numériques des lignes et des colonnes qui maximisent sous contrainte d'orthogonalité la corrélation associée à la table de contingence dérivée du graphe de voisinage. On peut cartographier ces coordonnées factorielles :





Or les lignes et les colonnes de la matrice M sont les 25 districts. Les coordonnées lignes et colonnes sont-elles égales ?

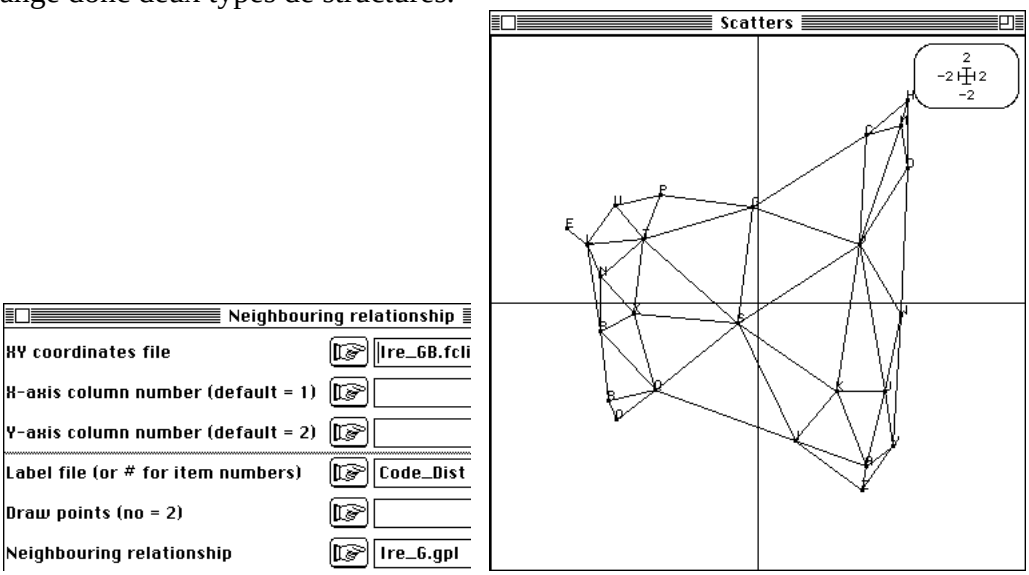
Ire_GB.fcli

	1	2	3	4	5	6	7	8	9
1	0.8145	-1.2255	0.5420	0.0048	-0.1968	-0.0720	0.0793	0.2636	0.4915
2	-1.1753	-0.2145	-0.5803	-0.8460	0.5954	-0.5281	0.5132	0.5923	-0.0619
3	0.8179	1.2587	-0.4760	0.1643	1.1076	0.8164	-0.2264	0.0705	0.0968
4	1.1284	1.0032	-0.5584	0.0634	1.1368	0.7131	0.4990	-0.5418	-0.4283
5	-1.4210	0.5526	1.0899	-2.9420	0.0792	0.5856	1.9540	-0.3940	0.8719
6	-0.0319	0.7127	0.3350	-0.3019	-0.7672	-0.1912	-0.8927	-0.3333	-0.9115

Ire_GB.fcco

	1	2	3	4	5	6	7	8	9
1	0.8145	-1.2255	0.5420	-0.0048	0.1968	0.0720	0.0793	-0.2636	-0.4915
2	-1.1753	-0.2145	-0.5803	0.8460	-0.5954	0.5281	0.5132	-0.5923	0.0619
3	0.8179	1.2587	-0.4760	-0.1643	-1.1076	-0.8164	-0.2264	-0.0705	-0.0968
4	1.1284	1.0032	-0.5584	-0.0634	-1.1368	-0.7131	0.4990	0.5418	0.4283
5	-1.4210	0.5526	1.0899	2.9420	-0.0792	-0.5856	1.9540	0.3940	-0.8719
6	-0.0319	0.7127	0.3350	0.3019	0.7672	0.1912	-0.8927	0.3333	0.9115

Il semble que la réponse est oui pour les facteurs 1, 2 et 3, mais qu'elle est négative pour les suivants. En fait on ne peut que trouver deux cas : les coordonnées sont exactement les mêmes (facteur direct) pour un même axe ou exactement de signe opposé (facteur inverse). Un facteur direct représente une autocorrélation positive, donc une structure cartographiable. Un facteur inverse représente une autocorrélation négative, donc une structure qui relève de la variance locale. Cette analyse du graphe mélange donc deux types de structures.



Observer sur la figure de la page précédente que les cartes 1 à 3 sont de "bonnes cartes", que les cartes 4 à 6 sont complètement déstructurées. On peut représenter également le graphe de voisinage dans le plan 1-2 (ci-dessus). Les deux premiers facteurs, comme dans les exemples de Lebart (op. cit.) sur les régions ou les départements de France, reconstitue l'espace concret. Il convient de faire apparaître cette dualité de résultat dans l'AFC du graphe en séparant ce qui a produit l'un ou l'autre des deux types de structure.

Vue d'une manière ou de l'autre cette analyse des correspondances fournit des premières coordonnées donnant les cartes les plus simples possibles ou des coordonnées

reconstituant l'espace concret à partir de la notion de voisinage. Cela vient des propriétés de l'analyse des correspondances comme analyse canonique (Williams 1952 op. cit.). Il apparaît alors que cette approche réunit la notion de maximiser la corrélation entre voisins ou minimiser la variance entre voisins sous couvert d'une variance totale unité. Cela ne peut fonctionner que si on place sur les points des poids proportionnels à leur nombre de voisins.

C'est exactement l'inverse qu'on a, jusqu'à présent, chercher à faire. On veut imposer aux arêtes du graphe des poids dérivés des poids issus du tableau de données, alors qu'il est plus naturel d'imposer dans l'analyse du tableau des poids de voisinage. Il est logique, dans l'analyse d'une structure spatiale de considérer qu'un point a d'autant plus d'importance qu'il est muni de plus de voisins.

Il faut enfin résoudre la question du mélange des types d'autocorrélations (positive et négative) qui intervient dans l'AFC du graphe de voisinage, tout en gardant la notion de covariance de voisinage, ce que ne font pas les opérateurs de Méot & Coll. [18]. L'introduction des poids de voisinage permet de répondre à ces exigences.

3.5 — Vecteurs propres de voisinage

On note *n* le nombre total de points (ici 25). Considérons la distribution marginale de la table de contingence définie par la matrice du graphe. Si **M** est cette matrice, *m* est le nombre total de couples de voisins (chaque arête est comptée deux fois) la table de contingence est $\mathbf{P}=(1/m)\mathbf{M}$. Si le point *i* a m_i voisins, son poids est m_i/m, la fréquence des arêtes qui ont l'origine ou l'extrémité au point *i*. On note **D** la diagonale des poids de voisinage.

On note $\mathbf{P}_0$ la matrice $\mathbf{D}^{-1}\mathbf{P}\mathbf{D}^{-1} - \mathbf{1}_{nn}$ du schéma de dualité de l'AFC du graphe de voisinage ($\mathbf{P}_0$, **D**, **D**). $\mathbf{D}^{-1}\mathbf{P}$ associe à un vecteur **x** à *n* composantes (x_1, x_2, ..., x_n) un vecteur **y** à *n* composantes (y_1, y_2, ..., y_n) où y_i est la moyenne des valeurs de **x** pour les voisins du point *i*.

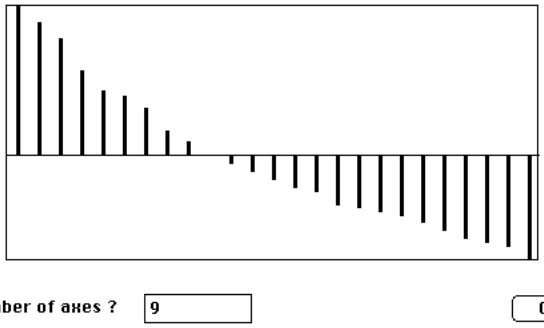
Si **x** est **D** centré, **y** l'est aussi. L'intérêt de la pondération **D** est que si **x** ne l'est pas **y** a encore la même moyenne que **x**. La moyenne des moyennes sur les voisins est la moyenne initiale. Centrer **x** ou **y**, c'est la même chose. L'opérateur $\mathbf{D}^{-1}\mathbf{P} - \mathbf{1}_{nn}\mathbf{D}$ est **D** symétrique, mais non positif (ce qui crée les problèmes dans l'AFC du graphe et pose des questions à Wartenberg, op. cit., en plus de la question des poids). Nous l'appellerons l'opérateur de Moran, en souvenir de celui qui a ouvert le voie.

Les valeurs propres de l'opérateur de Moran sont en valeur absolue les radicaux des valeurs propres de l'AFC de Lebart, mais sont positives si l'autocorrélation est positive et négatives si l'autocorrélation est négative. Le tri est automatique. Les cartes des vecteurs propres sont les mêmes mais automatiquement rangées par valeur décroissante d'autocorrélation. Pour s'en convaincre détruire les fichiers propres à l'AFC du graphe et exécuter l'option Moran Eigenvectors du module Distances :

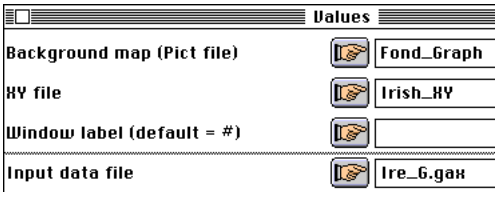
qui affiche :

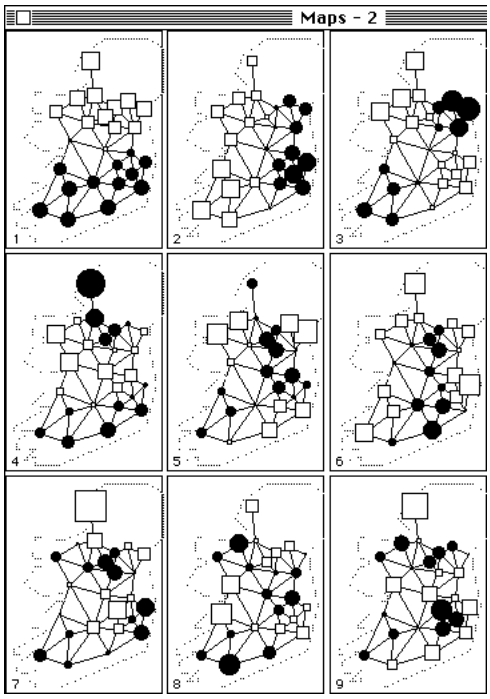
```
Moran operator diagonalization
Neighboorhood graph: Irish_G
-----------------------------------------
Num  Eigenval.|Num. Eigenval.|Num. Eigenval.|Num. Eigenval.|
001  8.948e-01|002  7.912e-01|003  6.979e-01|004  5.082e-01|
005  3.798e-01|006  3.489e-01|007  2.766e-01|008  1.423e-01|
009  8.101e-02|010  1.863e-09|011 -5.100e-02|012 -1.013e-01|
```

```
013 -1.499e-01|014 -1.962e-01|015 -2.210e-01|016 -3.041e-01|
017 -3.184e-01|018 -3.413e-01|019 -3.628e-01|020 -4.072e-01|
021 -4.543e-01|022 -5.055e-01|023 -5.226e-01|024 -5.507e-01|
025 -6.344e-01|

File Irish_Gvp contains the eigenvalues
--- It has 25 rows and 1 column
File Irish_Gax contains eigenvectors (norm = 1 for neighboorhood weights)
--- It has 25 rows and 9 columns
```

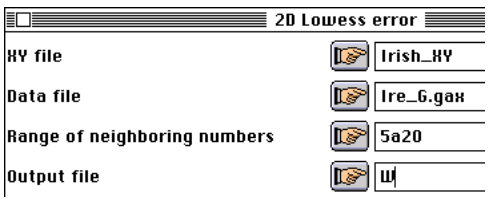


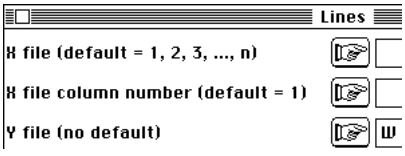
Au milieu de la liste des valeurs propres, observer la valeur 1.863 10^{-9}, c'est-à-dire 0. Celles qui précèdent sont des corrélations positives et celles qui suivent des corrélations négatives. On garde alors les 9 vecteurs propres et on cartographie ces 9 premiers associés aux 9 valeurs positives. On retrouve ainsi les seules cartes artificielles qui ont une signification spatiale :

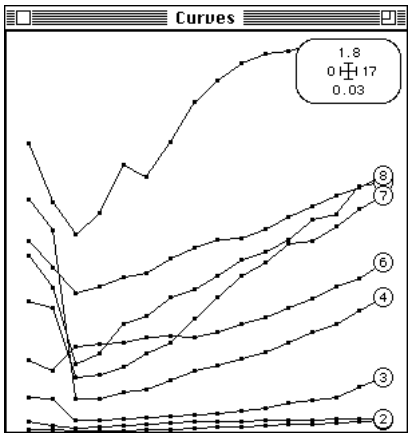




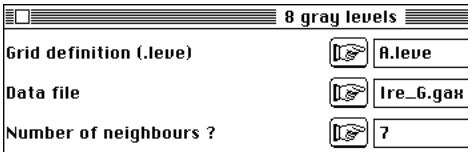
Le caractère particulier de ces codes numériques renvoie à la cartographie par Levels. Pour choisir le nombre de voisins utiliser encore MapUtil : 2D Lowess error et visualiser dans Curves :

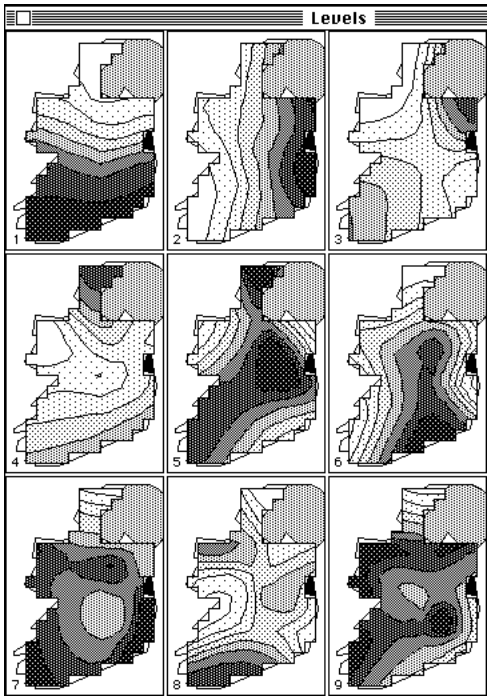






Observer que l'erreur croît avec le rang du vecteur, ce qui est normal puisqu'ils sont de corrélation spatiale décroissante. De plus, il convient de ne pas utiliser un trop grand nombre de voisins, ce qui introduirait de l'erreur artificielle de modélisation. On peut s'en tenir à 7 :



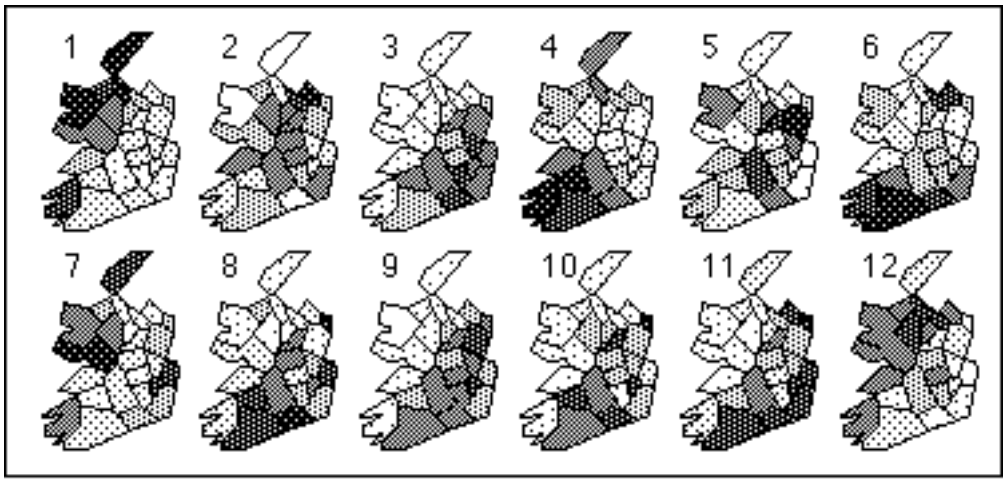


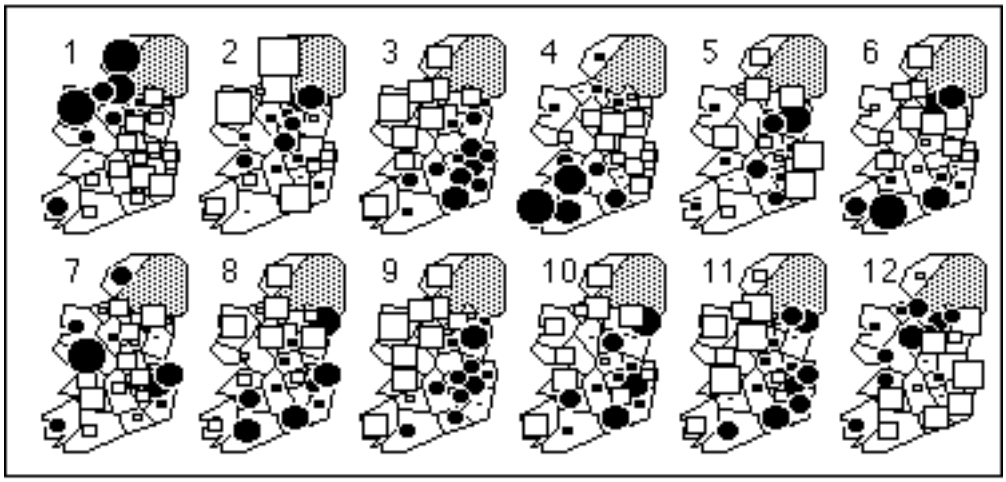
On reconnaît les fonctions x, y et xy sur les trois premières cartes. Les cartes les plus simples sont proches des fonctions polynomiales des coordonnées.

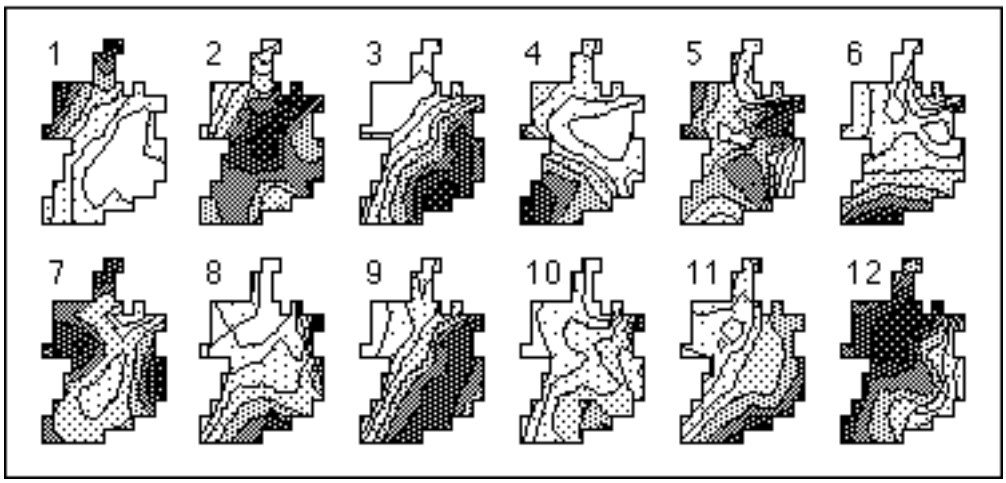
La technique employée s'appliquant à toute forme de voisinages, on a ici l'indication très précieuse que les vecteurs propres de l'opérateur de Moran sont la solution des problèmes demandant d'introduire des covariables exprimant l'espace, en place et lieu des polynômes des coordonnées préconisées par les méthodes dérivées de la *Trend Surface Analysis*, par exemple par Borcard et Coll.**[19]**.

4 — L'analyse globale

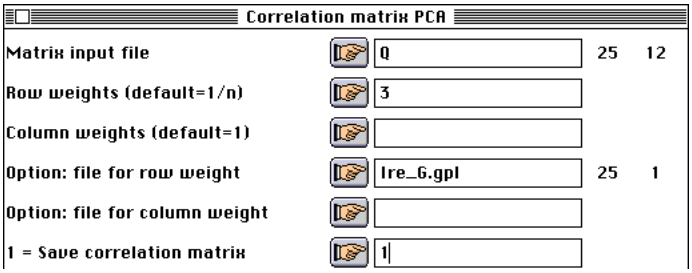
La figure qui suit repose la même question. De quelque manière qu'on s'y prenne, les cartes par variables se ressemblent suffisemment pour appeler une analyse multivariée qui intègre ce phénomène :



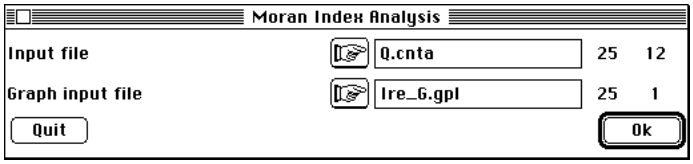




Pour atteindre l'objectif fixé, faire d'abord l'ACP normée du tableau, puis utiliser l'option Moran Index Analysis du module NGStat :



Noter qu'il est impératif d'introduire à ce niveau les poids de voisinage.



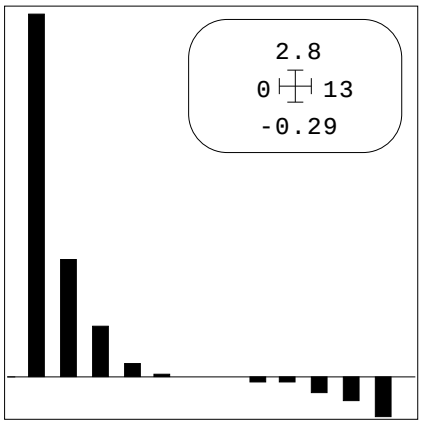
On obtient :

```
Moran Index linear analysis
Access to neighbouring relationship: Ire_G.gpl
Preliminary analysis: Q.cnta
----------------------------------------
Trace =  3.579e+00
```

Le nom Q_MA a été défini par défaut.

```
File Q_MA.xmcs contains spatial covariances
--- It has 12 rows and 12 columns

Num  Eigenval.|Num. Eigenval.|Num. Eigenval.|Num. Eigenval.|
001  2.741e+00|002  8.975e-01|003  3.801e-01|004  1.030e-01|
005  2.213e-02|006  8.262e-08|007 -2.164e-03|008 -7.901e-03|
009 -2.406e-02|010 -9.193e-02|011 -1.577e-01|012 -2.814e-01|

File Q_MA.xmvp contains the eigenvalues
--- It has 12 rows and 1 column
```

Les valeurs propres (Curves sur Q_MA.xmvp) ne sont pas toutes positives :



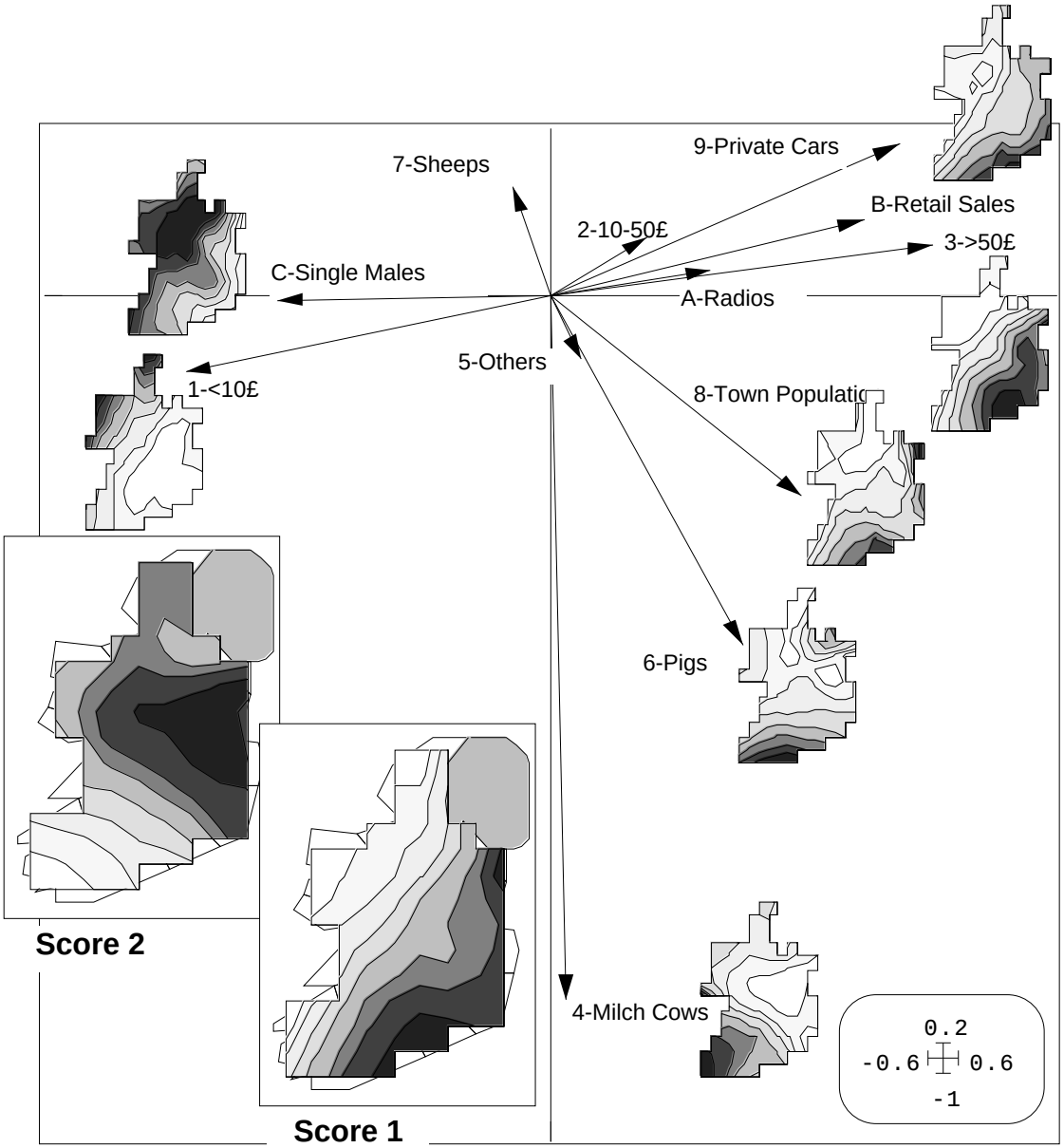
Le fichier .xmw1 contient les coefficients des combinaisons linéaires que sont les variables de synthèse. x désigne une analyse à un tableau, m rappelle qu'on maximise le coefficient de Moran et w indique des poids (weights).

```
File Q_MA.xmw1 contains weights of columns for row scores
--- It has 12 rows and 2 columns
File :Q_MA.xmw1
|Col.|   Mini   |   Maxi   |
|----|----------|----------|
|   1|-4.317e-01| 4.539e-01|
|   2|-8.328e-01| 1.799e-01|
|----|----------|----------|

File Q_MA.xmli contains row scores (linear combination having maximal Moran
Index)
--- It has 25 rows and 2 columns
```

Le fichier .xmli contient les variables de synthèse. x désigne une analyse à un tableau, m rappelle qu'on maximise le coefficient de Moran et li indique des codes

lignes. Ces codes de synthèse donne des cartes optimales par combinaisons des variables de départ.

```
File :Q_MA.xmli
|Col.|   Mini   |   Maxi   |
|----|----------|----------|
|   1|-4.305e+00| 3.476e+00|
|   2|-3.413e+00| 1.777e+00|
|----|----------|----------|

Axis:   1  Spatial correlation: 5.592e-01
Axis:   2  Spatial correlation: 5.425e-01
```



Ces indications sont fondamentales. Les valeurs propres sont les produits des variances des codes de synthèse par la corrélations spatiales associée. Les variances sont directement comparables aux inerties des facteurs de l'analyse de base. Les corrélations spatiales indique la qualité des cartes. Comme dans nombre d'analyses, en maximisant la covariance spatiale, on maximise la variance et la corrélation spatiale. Le produit des deux est la valeur propre. On trace pour interpréter deux facteurs de corrélation équivalente les cartes des codes de synthèse (Levels sur Q_MA.xmli) :

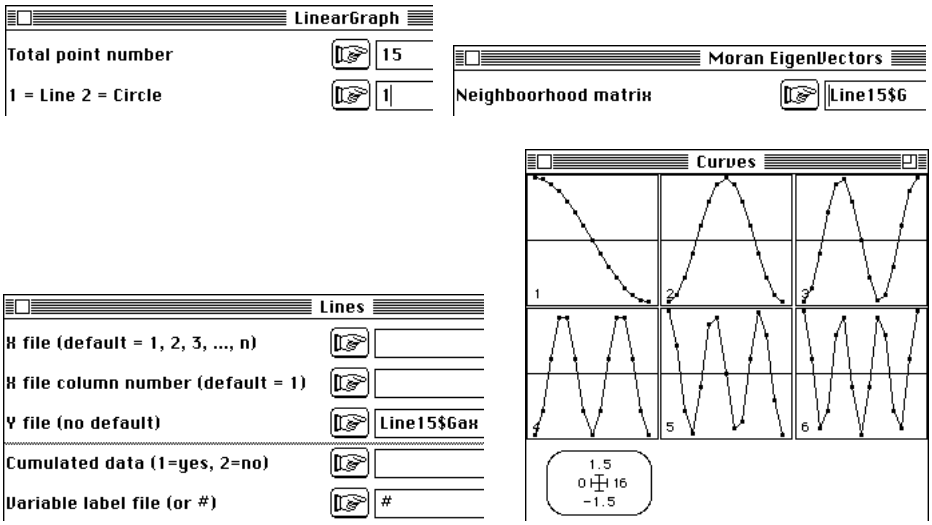
L’interprétation des résultats ne pose aucun problème : il existe deux structures cartographiables, celle de la richesse totale (nord-est/sud-ouest) liée à la plus grande partie des variables et celle de l’élevage bovin largement indépendant de la précédente. Des exemples de taille beaucoup plus importante devront montrer si cette procédure est capable de détecter des structures inaccessibles par les analyses classiques.

Les poids des variables (Scatters sur Q_MA.xmw1) permet d’exprimer clairement l’essentiel des résultats.

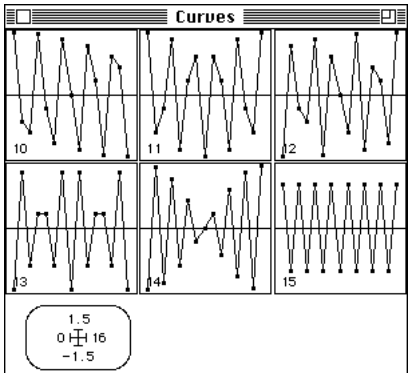
5 — Conclusion

Pour résumer cette fiche sur les structures spatiales, nous retiendrons les éléments suivants. Une des approches les plus simples d’introduction de la structure spatiale dans l’analyse des données est celle des graphes de voisinages.

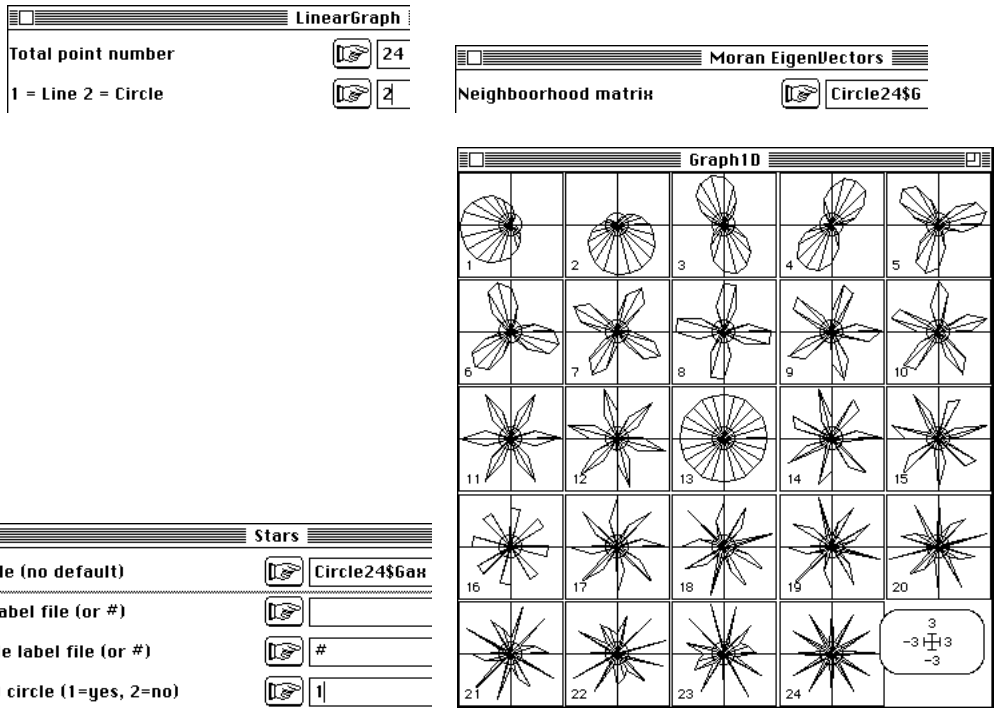
Cette notion est d’une grande plasticité. Elle intervient pour les mesures simplement ordonnées sur un axe (gradient temporel, échantillonnage sur un transect , NGUtil: LinearGraph) :



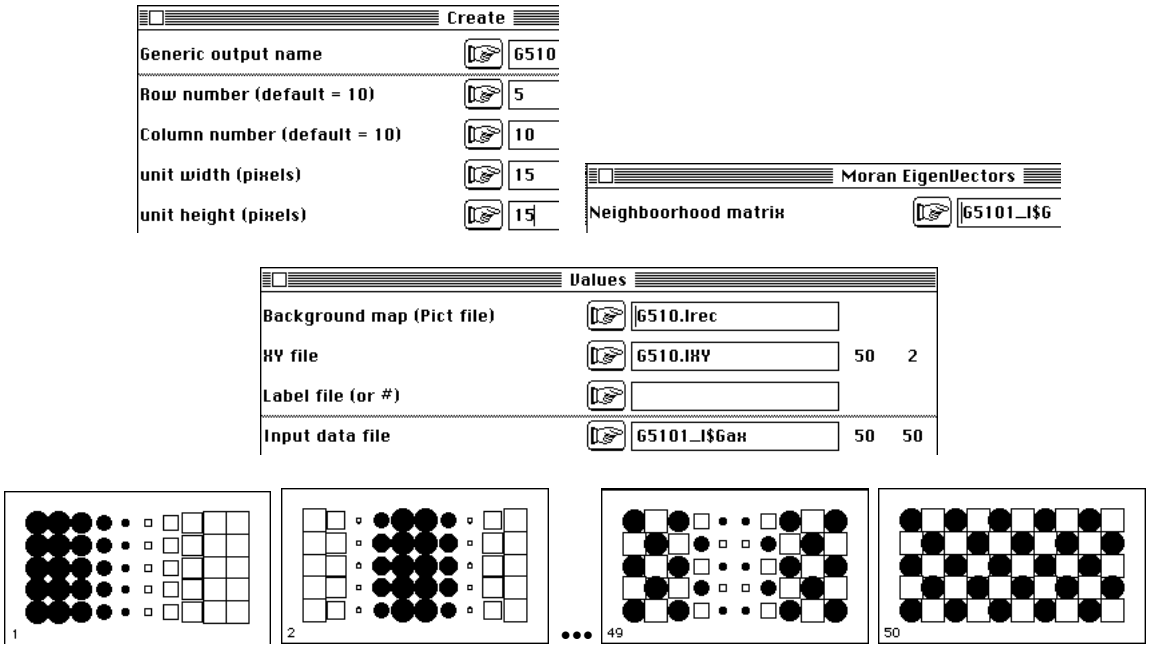
Deux points sont voisins s’ils se suivent. Le premier vecteur propre de l’opérateur associé modélise les gradients. Le dernier approche les processus à forte autocorrélation négative :



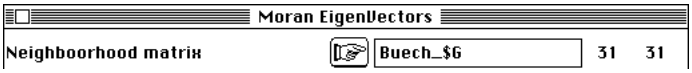
Elle intervient pour les mesures périodiques (chronobiologie) :

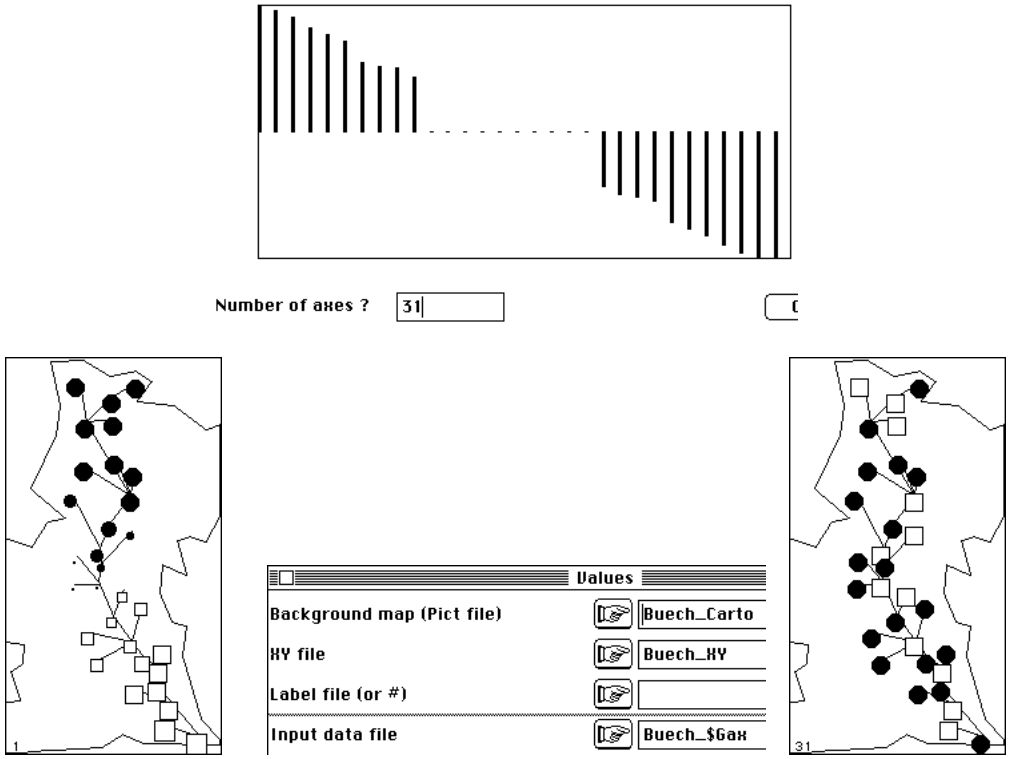


Elle concerne facilement les mesures sur grilles complètes :



Elle aborde les unités surfaciques (ci-devant) comme les structures arborescentes :





Quand on possède un graphe de voisinage, un opérateur symétrique définit conjointement la variance locale et l'autocorrélation spatiale comme deux composantes complémentaires de la variance initiale. Le test de Geary garantit la signification statistique des ces mesures. Les vecteurs propres de ces opérateurs donnent des codes numériques qui ont du sens et recouvrent l'éventail du plus autocorrélé au plus localement variable.

Devenant un produit scalaire, la variance locale ou l'autocorrélation permettent de faire de l'analyse des données sous contrainte spatiale.

Reste alors la question de fond. Comment mesurer une structure spatiale en utilisant plusieurs échelles ? Qu'est-ce qu'une échelle spatiale ? Qu'est-ce qu'une mesure multi-échelles de la covariation entre deux variables ? Les grilles de dénombrement montrent clairement qu'un seul graphe de voisinage ne suffit pas à rendre compte de la complexité multi-échelle (programme GNDEN). Déjà difficile pour les grilles complètes, la question empire avec des structures de voisinage quelconque. La notion de graphe de voisinage s'étend elle aux multi-échelle ? Peut-on accéder à des analyses multivariées et multi-échelles ? Cette question sera ouverte dans la suite.

Références

[1] 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.

[2] Hathaway, W.H. (1971) Contingency table analysis of rain forest vegetation. In : Statistical Ecology. III Many species populations ecosystems and systems analysis. Patil, G.P., Pielou, E.C. & Waters, W.E. (Eds.) Pennsylvania State University Press. 271-314.

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

[4] Esteve, J. (1978) Les méthodes d'ordination : éléments pour une discussion. In : Biométrie et Ecologie. Legay, J.M. & Tomassone, R. (Eds.) Société Française de Biométrie, Paris. 223-250.

[5] Dessier, A. & Laurec, A. (1978) Le cycle annuel du zooplancton à Pointe-Noire (RP Congo). Description mathématique. Oceanologica acta : 1, 3, 285-304.

[6] Geary, R.C. (1954) The contiguity ratio and statistical mapping. *The incorporated Statistician* : 5, 3, 115-145.

[7] Cliff, A.D. & Ord, J.K. (1973) *Spatial autocorrelation*. Pion, London. 1-178.

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

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

[10] 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.

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

[12] Wartenberg, D. (1985c) Multivariate spatial correlations: a method for exploratory geographical analysis. *Geographical Analysis* : 17, 4, 263-283.

[13] Wartenberg, D.E. (1985b) Spatial autocorrelation as a criterion for retaining factors in ordinations of geographic data. *Mathematical Geology* : 17, 665-682.

[14] Wartenberg, D.E. (1985a) Canonical trend surface analysis: a method for descibing geographic pattern. *Systematic Zoology* : 34(3), 259-279.

[15] Benzecri, J.P. & Coll. (1973) *L'analyse des données. II L'analyse des correspondances*. Bordas, Paris. 1-620.

[16] Lebart, L. (1984) Correspondence analysis of graph structure. *Bulletin technique du CESIA*, Paris : 2, 1-2, 5-19.

[17] Williams, E.J. (1952) Use of scores for the analysis of association in contingency tables. *Biometrika* : 39, 274-289.

[18] Méot, A., Chessel, D. & Sabatier, R. (1993) Opérateurs de voisinage et analyse des données spatio-temporelles. In : *Biométrie et Environment*. Lebreton, J.D. & Asselain, B. (Eds.) Masson, Paris. 45-72.

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

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.