Fiche de Biostatistique

Analyse de données spatialisées

D. Chessel & J. Thioulouse

Résumé

La fiche regroupe quelques éléments de base dans l'analyse des données multivariées spatialisées. On aborde les tests élémentaires (Gery, Moran, Mantel) et des pratiques courantes (arbres de longueur minimale, courbes de niveaux). Les liens entre matrices de coordonnées, matrices de distances et graphes de voisinage sont envisagés. L'analyse en coordonnées principales est définie.

Plan

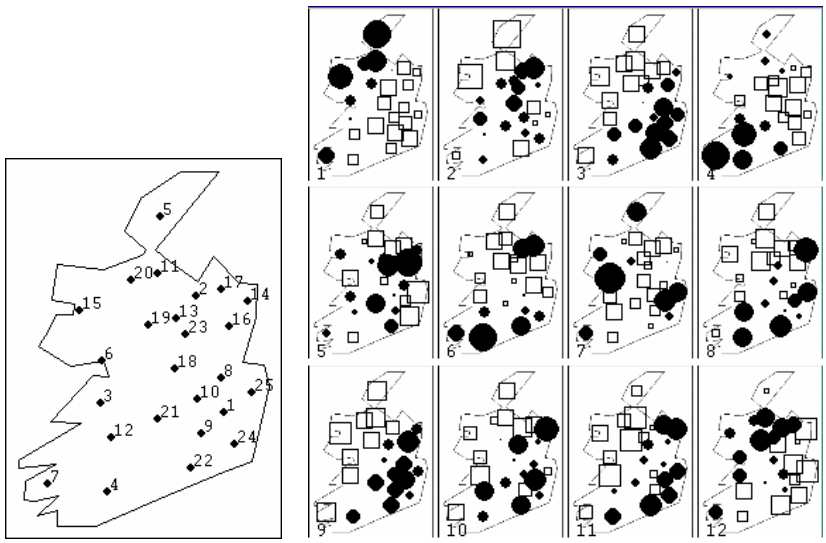
1. Introduction

Une grande partie des données acquises, en génétique, en écologie ou en biologie des populations est spatialisée. Une mesure se réfère à un endroit de mesure. Relier l'espace et l'observation est un problème tellement général qu'une partie des statistiques y est consacrée. On parle de statistiques spatiales. Le chapitre réunit quelques éléments de base dans ce domaine. Les enregistrements de l'espace lui-même sont multiples.

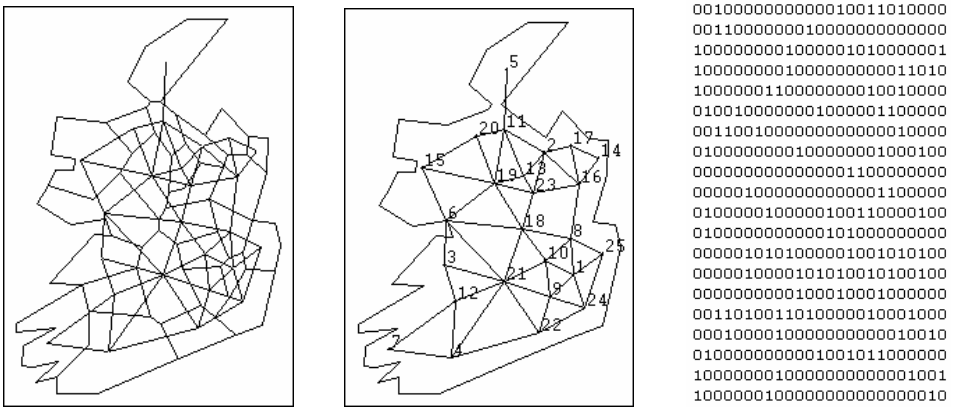
Enregistrements surfaciques : la mesure porte sur une surface bornée par une frontière. C'est le support des données socio-économiques. Un des articles fondateurs de ce domaine [1] traite des comptés d'Irlande :

Cartographie par niveau de gris sur des unités surfaciques. 25 districts d'Irlande. Le district de Dublin est extrait du jeu de données. Carte des districts avec numérotation naturelle. Tableau de 12 variables mesurées sur les 24 districts. Données célèbres reprises dans [2] p. 53. Code des variables : 1-2-3 répartition (en 1 pour 1000) des propriétés agricoles en 3 groupes d'imposition (<10 £, 10-50 £, >50 £). 4-5-6-7 Nombres moyens d'animaux pour 1000 acres de prairies et cultures respectivement 4- vaches laitières, 5- autres bestiaux, 6- cochons, 7- moutons. 8- Pourcentage de population urbanisée (villes et villages) en 1 pour 1000 9- Nombre de voitures pour 1000 habitants 10- Nombre de licences de radio pour 1000 habitants 11- Ventes de détail moyenne par habitant en £ 12- Pourcentage de célibataires parmi les hommes de 30-34 ans en 1 pour 1000. Données normalisées.

Enregistrements ponctuels : la mesure se réfère à deux coordonnées (x, y). On passe des données surfaciques aux données ponctuelles en choisissant un point particulier par unités :



Enregistrement du voisinage : l'espace est défini par une relation de voisinage, donc une matrice qui a autant de lignes et de colonnes qu'il y a de points de mesures. Cette matrice contient à la ligne *i* et à la colonne *j* la valeur 1 si les points sont voisins 0 sinon. Par exemple, deux unités surfaciques sont voisines si elles ont une frontière commune :



```
000000001110000000000000011
000000000001010011000001000
000001100001000000000100000
000000100001000000000110000
000000000001000000000000000
001000000000000100110100000
001100000000100000000000000
100000000100000010100000001
100000000100000000000110100
100000011000000000100100000
010010000000010000011000000
001100100000000000000100000
010000000010000000010001000
000000000000000011000000000
000001000000000000011000000
010000010000010011000010000
010000000000001010000000000
000001010100000010010101000
000001000001010100101001000
000000000001000100010000000
001101001101000001000010000
000100001000000000000100100
010000000000010010110000000
100000001000000000000001001
100000010000000000000000010
```

Les points sont les sommets du graphes, les paires de points sont les arêtes du graphe. On peut utiliser un graphe de voisinages pour exprimer la forme d'espaces particuliers comme les réseaux hydrographiques, les frontières infranchissables, …

Enregistrement des distances : le cas le plus simple est celui de la distance euclidienne canonique : $d\left(A,B\right)=\sqrt{\left(x_A-x_B\right)^2+\left(y_A-y_B\right)^2}$. On obtient ainsi une matrice de distances.

En mathématiques, on appelle *distance* définie sur un ensemble E une fonction d de ExE dans R qui vérifie pour tout x, y et z éléments de E :

$$(1)\ d(x,y)\geq 0$$
$$(2)\ d(x,y)=0 \Leftrightarrow x=y$$
$$(3)\ d(x,y)=d(y,x)$$

$$(4)\ d(x,y) \le d(x,z) + d(z,y)$$

En statistiques, on appelle *dissimilarité* définie sur un ensemble fini I à *n* éléments (numérotés 1, 2, ..., *i*, ..., *n*) une fonction de IxI dans R qui vérifie pour tout *i* et *j* :

$$(1)\ \partial_{ij} \ge 0$$
$$(2)\ \partial_{ii} = 0$$
$$(3)\ \partial_{ij} = \partial_{ji}$$

En biologie, on utilise le terme de distance pour désigner la différence mesurée entre deux individus, deux populations, deux sites, ..., sans se préoccuper de définition. Pour suivre la coutume on appellera *matrice de distances* une matrice contenant une *dissimilarité observée*. Les matrices de distances sont donc des matrices carrées (*n* lignes et *n* colonnes), contenant des nombres positifs (1), symétriques (3), ayant des éléments nuls sur la diagonale (2). Sur les comtés d'Irlande, on obtient :

```
max value =  2.18763e+02
Content as 1000*x/max
------------------------------------------------------------
[  1]    0
[  2]  413    0
[  3]  428  493    0
[  4]  487  740  307    0
[  5]  712  301  677  969    0
[  6]  458  390  148  454  536    0
[  7]  653  820  330  204 1000  464    0
[  8]  120  295  426  556  596  415  698    0
[  9]  107  475  365  382  764  427  555  204    0
[ 10]  104  352  336  449  641  354  592  108  123    0
[ 11]  531  151  489  773  198  355  816  421  573  452    0
…
[ 24]  115  528  483  467  827  540  654  234  118  203  645
       423  482  500  706  407  541  333  508  671  276  171
       416    0
[ 25]  118  381  522  605  681  525  767  114  225  186  519
       507  365  316  655  238  372  276  424  567  335  337
       302  192    0
------------------------------------------------------------
```

Un tableau de données spatialisées peut donc se trouver en face d'un tableau de coordonnées spatiales, un graphe de voisinage ou une matrice de distances. Il existe une multitude de pratiques potentielles soit pour amener l'espace dans le mode de perception des données (tableau contre tableau) soit amener les données dans le mode de perception de l'espace (matrice de distances contre matrice de distances).

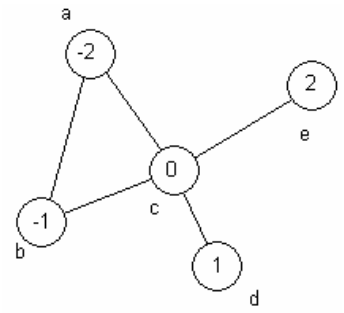
2. Tests élémentaires

2.1. Variance locale et test de Geary

L'ouvrage classique de Cliff & Ord *op. cit.* 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 [3] et le procédé a été utilisé indépendamment

par Light & Margolin [4] 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

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

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

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

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 :

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

On retiendra la relation fondamentale :

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

La variance est *la moitié* de la moyenne des carrés des différences élémentaires. 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. S'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 $c = \frac{\hat{V}_{loc}}{\hat{V}}$ de la variance estimée localement sur la variance estimée totalement soit égal à 1 ou encore que $I_G = \frac{1-c}{\sqrt{\text{var}\left(c\right)}}$ ne soit pas significativement différent 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 échantillon 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.

Exécuter le test de Geary sur les données d'Irlande (NGStat: Geary Test) :


```
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|
l---l-----------l-----------l----------l-----------l----------l
```

Le listing donne dans l'ordre

— le numéro de la variable,

— la quantité c observée (rapport de la variance locale ou variance mesurée sur les couples de voisins seulement à la variance totale ou variance mesurée sur l'ensemble des couples),

— l'approximation normale associée sous l'hypothèse de normalité et de non corrélation spatiale (test paramétrique),

— la probabilité de dépasser l'observation dans le test précédent,

— l'approximation normale associée sous l'hypothèse de loi quelconque unique et de non corrélation spatiale (test non paramétrique *distribution free*),

— la probabilité de dépasser l'observation dans le test précédent.

Le lissage des cartes par courbes de niveaux est légitime pour la plupart des variables. On a encadré les résultats de la variable 4, qui sont conformes à ceux de Cliff & Ord (1973 op. cit. page 57). La confrontation de ces statistiques aux 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 [5] que nous appelons aussi opérateurs de Moran.

2.2. L'indice de Moran

La notion d'autocorrélation spatiale mesure essentiellement la ressemblance entre voisins. L'idée est initialement celle de Moran (1948)[6]. 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) [7], 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)[8] et pour approfondir l'usage des coordonnées concrètes dans l'espace comme données numériques (Wartenberg 1985a)[9].

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

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

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 :

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

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 :

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

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). On peut rendre ces deux notions cohérentes (§ 4).

2.3. Test de Mantel

Il est utilisé [10] si l'espace est introduit par une matrice de distances spatiales. On trouve une présentation détaillée dans [11] (p.70-75). L'espace est connu par une matrice **S** de distances spatiales. Les données forment un tableau duquel on déduit une distance entre les individus consignée dans une matrice de distances **D**. La corrélation entre les deux est mesurée directement par $\sum_{i=1}^{n}\sum_{i=1}^{n} s_{ij}d_{ij}$. Les couples *ii* ne jouent aucun rôle puisque les distances sont nulles. Peu importe également que l'on compte une fois ou deux fois les couple *ij* et *ji* . Seul importe le type de permutations utilisées. Une des matrices est laissée en place et dans l'autre lignes et colonnes sont permutées à l'identique, par exemple :

$$25134 \Rightarrow \begin{bmatrix} 11 & 12 & 13 & 14 & 15 \\ 21 & 22 & 23 & 24 & 25 \\ 31 & 32 & 33 & 34 & 35 \\ 41 & 42 & 43 & 44 & 45 \\ 51 & 52 & 53 & 54 & 55 \end{bmatrix} \rightarrow \begin{bmatrix} 22 & 25 & 21 & 23 & 24 \\ 52 & 55 & 51 & 53 & 54 \\ 12 & 15 & 11 & 13 & 14 \\ 32 & 35 & 31 & 33 & 34 \\ 42 & 45 & 41 & 43 & 44 \end{bmatrix}$$

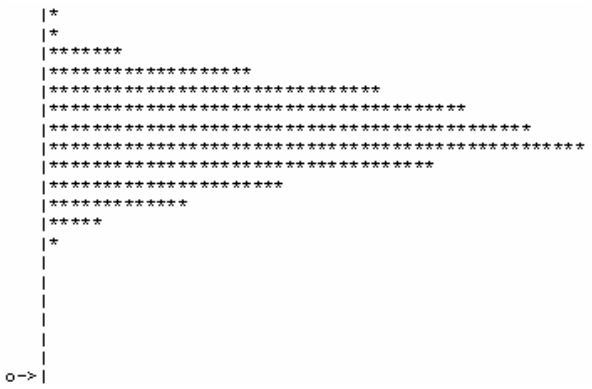
Pour chacune de *m* permutations de ce type, on calcule la statistique $\sum_{i=1}^{n}\sum_{i=1}^{n} s_{ij}d_{ij}$ et on compare la valeur observée à l'ensemble des permutations. L'habitude veut que l'on corrige par les moyennes et les écarts-types pour faire apparaître exactement la corrélation entre les deux statistiques :

$$\begin{bmatrix} d_{21} & d_{31} & d_{32} & d_{41} & d_{42} & d_{43} & \dots & d_{n1} & d_{n2} & \dots & d_{n(n-1)} \\ s_{21} & s_{31} & s_{32} & s_{41} & s_{42} & s_{43} & \dots & s_{n1} & s_{n2} & \dots & s_{n(n-1)} \end{bmatrix}$$

Par exemple, pour les données d'Irlande on peut comparer les distances :

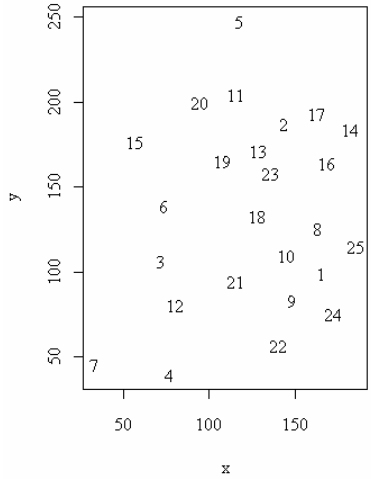
$$s_{ij} = \sqrt{\left(x_i - x_j\right)^2 + \left(y_i - y_j\right)^2} \leftrightarrow d_{ij} = \sqrt{\sum_{k=1}^{p} \frac{\left(x_{ik} - x_{jk}\right)^2}{Var_j}}$$

```
r index :  5.114e-01
number of random matching: 1000   Observed: 0.511390
Histogramm:  minimum = -0.261035, maximum = 0.511390
number of simulation X<Obs: 1000 (frequency: 1.000000)
number of simulation X>=Obs: 0 (frequency: 0.000000)
```



Le test de Mantel utilise une distance spatiale. Le test de Geary utilise un graphe de voisinage. On peut aussi tester le rôle de l'espace par une simple régression multiple sur les polynômes des coordonnées cartésiennes.

```
> plot(x,y,type="n")
> text(x,y,as.character(1:25))
```



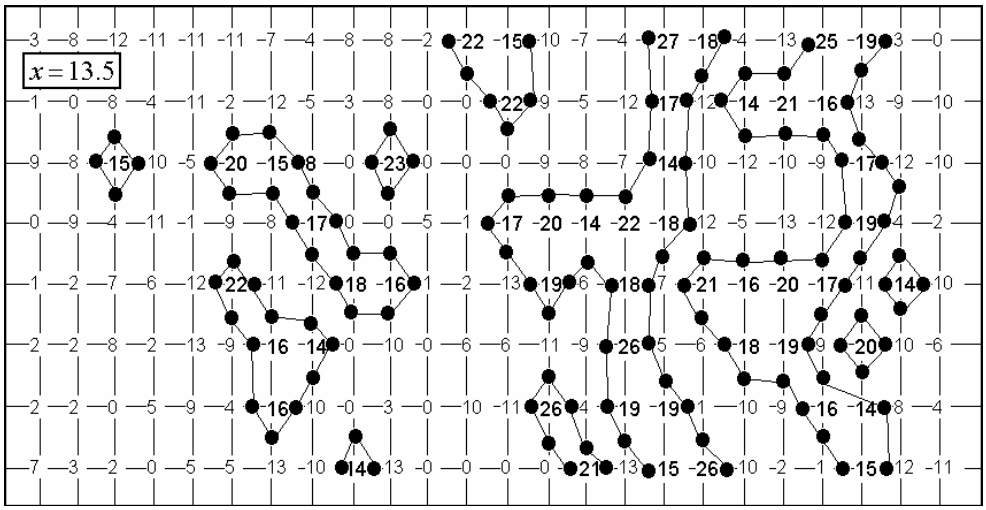

```
> qnor <- scale(read.table("Q.txt"))
> z <- qnor[,1]
> anova(lm(z~x+y+I(x*y)+I(x^2)+I(y^2)))
Analysis of Variance Table

Response: z
          Df Sum Sq Mean Sq F value  Pr(>F)
x          1   6.59    6.59   48.47 1.2e-06 ***
y          1   9.63    9.63   70.87 7.8e-08 ***
I(x * y)   1   0.35    0.35    2.54 0.12746
I(x^2)     1   1.96    1.96   14.42 0.00122 **
```

```
I(y^2)     1   2.90    2.90   21.31 0.00019 ***
Residuals 19   2.58    0.14
---
Signif. codes:  0  `***'  0.001  `**'  0.01  `*'  0.05  `.'  0.1  ` '  1
```

Le lien avec l'espace peut donc passer par un tableau (polynôme des coordonnées), une matrice de distances (Mantel), un graphe de voisinage (Geary-Moran). Divers procédés permettent de passer d'une forme à l'autre et la statistique spatiale est souvent faite de pratiques empiriques.

3. Courbes de niveaux

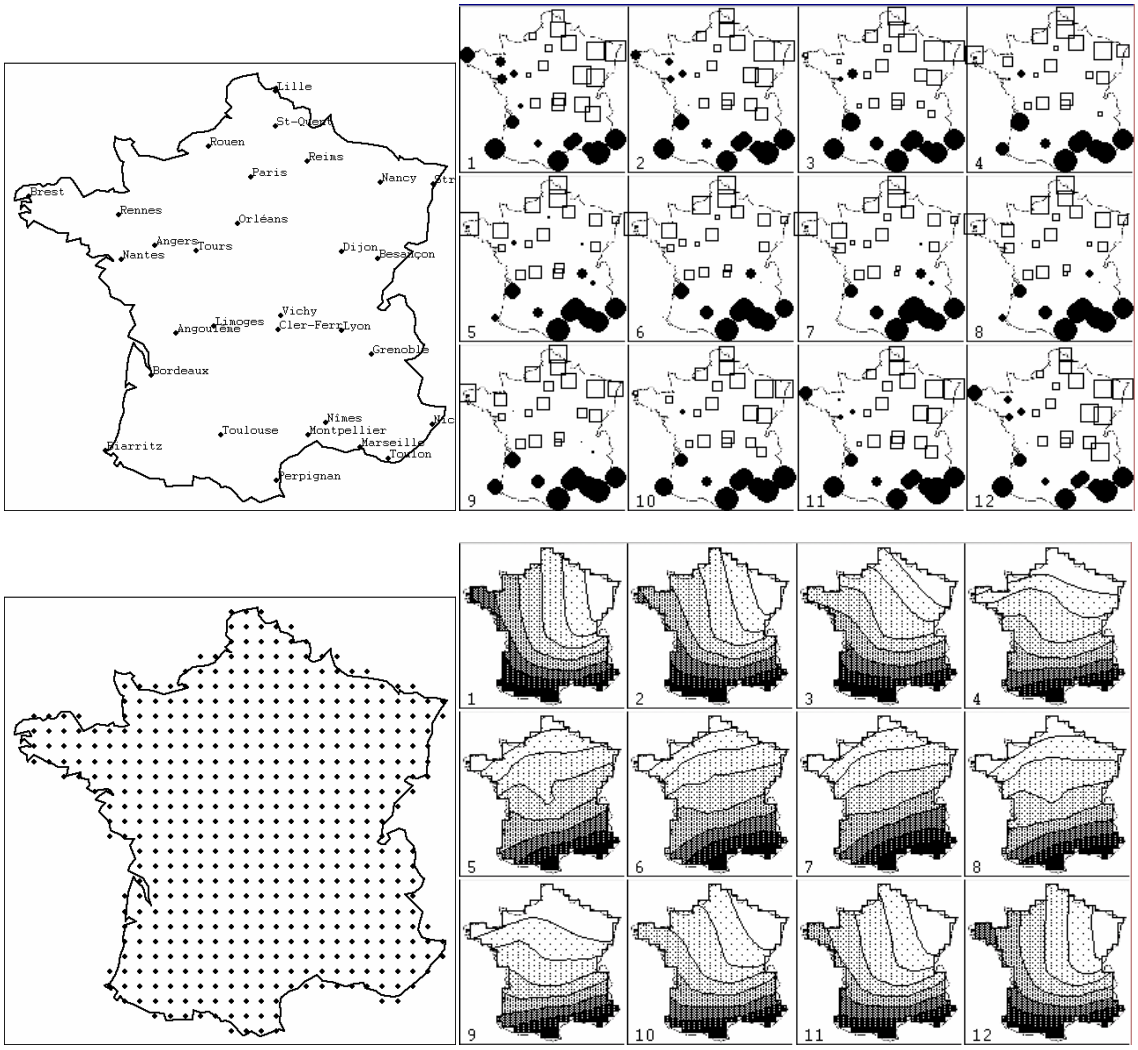
	3	8	12	11	11	11	7	4	8	8	2	22	15	10	7	4	27	18	4	13	25	19	3	0
	1	0	8	4	11	2	12	5	3	8	0	0	22	9	5	12	17	12	14	21	16	13	9	10
	9	8	15	10	5	20	15	8	0	23	0	0	0	9	8	7	14	10	12	10	9	17	12	10
	0	9	4	11	1	9	8	17	0	0	5	1	17	20	14	22	18	12	5	13	12	19	4	2
	1	2	7	6	12	22	11	12	18	16	1	2	13	19	6	18	7	21	16	20	17	11	14	10
	2	2	8	2	13	9	16	14	0	10	0	6	6	11	9	26	5	6	18	19	9	20	10	6
	2	2	0	5	9	4	16	10	0	3	0	10	11	26	4	19	19	1	10	9	16	14	8	4
	7	3	2	0	5	5	13	10	14	13	0	0	0	0	21	13	15	26	10	2	1	15	12	11



Les courbes de niveaux sont un jeu algorithmique sur un ensemble de valeurs placées sur un réseau. Ci-dessus : nombre de cannes de Framboisier par placettes de 1 m de côtés considéré comme mesure ponctuelle d'abondance[12]. Les points sont placés entre une mesure supérieure au seuil et une mesure inférieure au seuil par interpolation linéaire. Quelles que soient les valeurs, on a une solution unique.

Les mesures ne sont pas en général sur un réseau. On estime alors à partir des données en (x,y) des valeurs sur un réseau qui recouvre la zone d'étude. Il existe des méthodes

simples (régression polynomiale) ou sophistiquées (krigeage). Une des plus efficaces est la régression locale étendue de une à deux dimensions.



Température (12 mois, moyennes pluriannuelles) en 30 villes. Données normalisées. Illustrations de développements mathématiques dans **13** *et* **14** *qui abordent la cartographie d'une évolution temporelle (évolution d'une courbe de 12 points en 30 sites) alors qu'ici on a bordé l'évolution temporelle d'une carte.*

4. 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 positif. 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 $\alpha = \overline{x}$ 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 :

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

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

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

où m_v 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 :

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

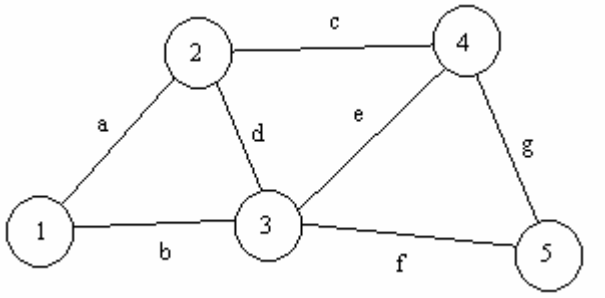
et si on réécrit l'indice de Geary sous la forme :

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

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

4.1. Décomposition matricielle

On peut donc redéfinir ces indices de manière plus efficace. La remarque fondamentale de départ est dans [15]. Soit un graphe de voisinage entre *n* points comportant *m* arêtes.



Soit **L** 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 $\mathbf{L}_{ik} = 1$, $\mathbf{L}_{il} = -1$ et $\mathbf{L}_{ij} = 0$ ailleurs. L'écriture est unique dès que la numérotation des sommets est donnée. Soit **M** la matrice de voisinage (*n* lignes et *n* colonnes) et **N** la matrice diagonale des degrés des sommets (nombre de voisins). Dans l'exemple :

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

On a : $\boxed{\mathbf{L}^t\mathbf{L} = \mathbf{N} - \mathbf{M}}$

$\mathbf{L}^t\mathbf{L}$ est une matrice symétrique et non négative ($\mathbf{x}^t\mathbf{L}^t\mathbf{L}\mathbf{x} \geq 0$). Les poids de voisinage sont sur la diagonale de $\mathbf{D} = \frac{1}{2m}\mathbf{N}$ (les arêtes sont comptées deux fois). $\mathbf{P} = \frac{1}{2m}\mathbf{M}$ est la table de contingence associée à la matrice du graphe.

Dans les indices de Moran et de Geary, seuls les numérateurs mesurent le lien de la variable avec l'espace. Les dénominateurs ne servent que de facteur d'échelle et sont des constantes dans l'espace des *n*! permutations des données. Considérons alors une statistique brute $(y_1, ..., y_n)$ sur les *n* sommets et $(x_1, ..., x_n)$ la variable centrée *avec la moyenne de voisinage.* Le numérateur de l'indice de Geary est la variance locale :

$$\hat{V}_{loc} = \frac{1}{4m}\sum_{i}^{n}\sum_{\text{voisin } j}\left(x_i - x_j\right)^2 = \frac{1}{4m}\sum_{i}^{n}\sum_{\text{voisin } j}\left(y_i - y_j\right)^2 = \frac{1}{2m}\sum_{\substack{i \\ i<j}}^{n}\sum_{\text{voisin } j}\left(y_i - y_j\right)^2 = \frac{1}{2m}\mathbf{y}^t\mathbf{L}^t\mathbf{L}\mathbf{y}$$

A noter que *m* est ici le nombre de paires de voisin et 2*m* est le nombre de couples. Le numérateur de l'indice de Moran est la covariance locale :

$$\frac{1}{2m}\sum_{i}\sum_{[\mathbf{M}]\text{voisin } j}\left(x_i - \bar{x}\right)\left(x_j - \bar{x}\right) = \frac{1}{2m}\sum_{i}\sum_{[\mathbf{M}]\text{voisin } j} y_i y_j = \frac{1}{2m}\mathbf{y}^t\mathbf{M}\mathbf{y} = \mathbf{y}^t\mathbf{P}\mathbf{y}$$

Enfin la variance totale est :

$$\sum_{i=1}^{n} p_i\left(x_i - \bar{x}\right)^2 = \sum_{i=1}^{n} p_i y_i^2 = \mathbf{y}^t\mathbf{D}\mathbf{y} = \frac{1}{2m}\mathbf{y}^t\mathbf{N}\mathbf{y}$$

La relation $\frac{1}{2m}\mathbf{L}^t\mathbf{L} = \frac{1}{2m}\mathbf{N} - \frac{1}{2m}\mathbf{M}$ assure la décomposition $\frac{1}{2m}\mathbf{M} = \frac{1}{2m}\mathbf{N} + \frac{1}{2m}\mathbf{L}^t\mathbf{L}$

d'où Variance totale = Variance locale + Covariance locale.

Cette décomposition est curieuse en ce que deux termes seulement sur les trois sont toujours positifs. 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, autocorrélé négativement la variance locale est plus forte que la variance et l'autocovariance est négative. Les deux statistiques disent la même chose tandis que leur somme est constante.

On peut résumer le total simplement. Un graphe de voisinage **M** définit des poids de voisinage (nombre d'arêtes arrivant en un point sur nombre d'arêtes total). Les données sont centrées (et si nécessaire normées) en utilisant cette pondération. Les poids sont écrits dans la matrice diagonale **D** et $\mathbf{y}^t\mathbf{D}\mathbf{y}$ est la variance globale de la variable.

$$\mathbf{y}^t\mathbf{D}\mathbf{y} = \mathbf{y}^t\left(\mathbf{D}-\mathbf{P}\right)\mathbf{y} + \mathbf{y}^t\mathbf{P}\mathbf{y} = \frac{1}{2}\left(\frac{1}{2m}\sum_{i,j\ \text{voisins}}^{n}\left(y_i - y_j\right)^2\right) + \frac{1}{2m}\sum_{i,j\ \text{voisins}}^{n} y_i y_j$$

La variance se décompose en variance locale et autocovariance.

La version purement informatique (test de Monte-Carlo[16]) et de plus multivariée du test de Geary utilise ces remarques. Soient *n* points de mesure reliés entre eux par un graphe de voisinage et **X** un tableau à *n* lignes et *p* variables quantitatives centrées pour la pondération de voisinage. La covariance spatiale de la colonne *j* de **X** est la quantité $\mathbf{x}_j^t\mathbf{P}\mathbf{x}_j$. **P** est la table de contingence associée au graphe et **D** est la pondération de voisinage. Le test compte pour chaque variable la fréquence des permutations aléatoires des *n* valeurs pour lesquelles on dépasse la covariance spatiale observée. Elle fait de même pour la quantité :

$$\sum_{j=1}^{p}\mathbf{x}_j^t\mathbf{P}\mathbf{x}_j$$

On obtient un test multivarié non paramétrique du lien entre le tableau et le graphe de voisinage, extension du test de Geary ou de Moran. Noter que les variables de **X** sont normalisées par la pondération de voisinage à chaque permutation. On peut comparer et noter la cohérence des p-values (en 1/10000) :

Test N	Test R	Random.
0	0	0
30	88	1140
0	0	0
0	1	0
5678	5662	3040
109	188	80
1925	2300	1380
54	48	140
6	2	10
1095	1085	3540
9	4	10
97	54	70

Le test global donne :

```
X = Total spatial covariance
number of random permutations: 1000   Observed:  3.579e+00
Histogram:  minimum = -2.232e+00, maximum =  3.579e+00
number of simulations X<Obs: 1000 (frequency:  1.000e+00)
number of simulations X>=Obs: 0 (frequency:  0.000e+00)
```

Le test variable par variable est la version test de permutations de l'indice de Moran dans sa version échantillonnage dans l'ensemble des permutations. Les tests de Geary équivalents sont des approximations (théorèmes de convergence). Il vaut mieux, vu le coût des calculs actuel, accorder plus de confiance au premier qu'aux seconds. Le test global porte exactement sur la trace de l'analyse globale. Il serait pris en défaut vraisemblablement par un mélange de variables (dans le même tableau) respectivement à variance locale forte et à autocorrélation spatiale forte.

4.2. Composantes cartographiables

On a alors simplement de nouvelles analyses multivariées sous contraintes spatiales. La première est l'analyse en composantes principales locales de Lebart, qui utilise le schéma :

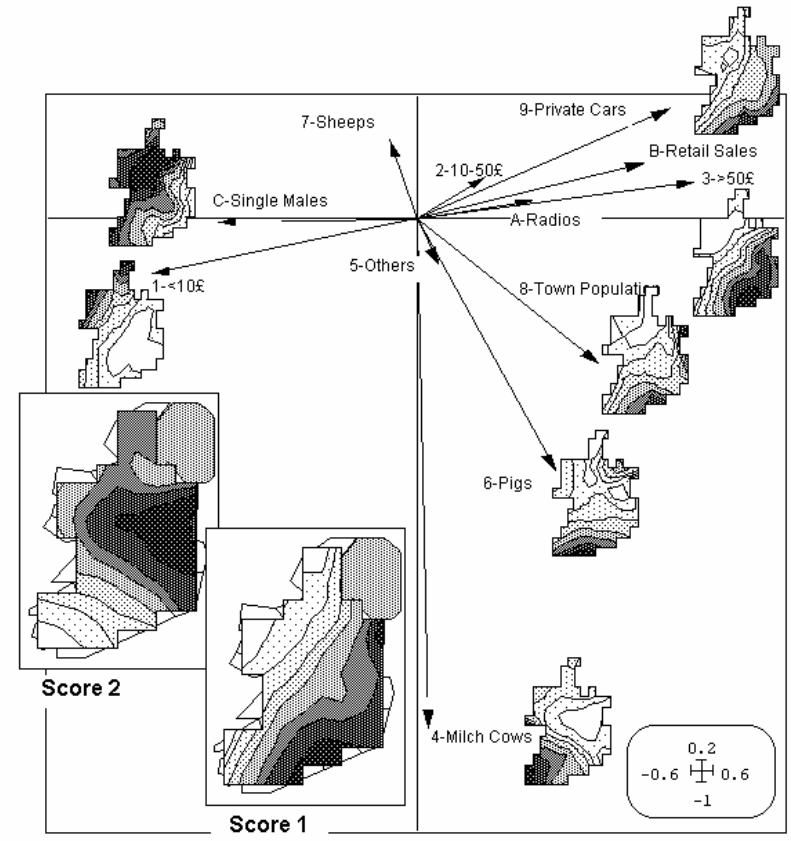
$$\begin{array}{ccc} & \mathbf{I}_p & \\ \boxed{p} & \rightarrow & \boxed{p} \\ \mathbf{X}^t \uparrow & & \downarrow \mathbf{X} \\ \boxed{n} & \leftarrow & \boxed{n} \\ & \mathbf{D}-\mathbf{P} & \end{array}$$

Elle donne directement l'axe principal **a** qui maximise $\mathbf{a}^t\mathbf{X}^t\left(\mathbf{D}-\mathbf{P}\right)\mathbf{Xa}$ donc la variance locale de la coordonnée $\mathbf{Xa}$. La seconde fait l'inverse et on peut l'appeler analyse en composantes principales cartographiables. Il suffit d'étendre les propriétés du schéma de dualité et appliquer l'extension à :

$$\begin{array}{ccc} & \mathbf{I}_p & \\ \boxed{p} & \rightarrow & \boxed{p} \\ \mathbf{X}^t \uparrow & & \downarrow \mathbf{X} \\ \boxed{n} & \leftarrow & \boxed{n} \\ & \mathbf{P} & \end{array}$$

Elle donne directement l'axe principal **a** qui maximise $\mathbf{a}^t\mathbf{X}^t\mathbf{PXa}$ donc l'autocovariance de la coordonnée $\mathbf{Xa}$. Noter qu'ici **P** n'est pas un produit scalaire, qu'il n'y a pas de nuages de points donc pas d'analyse d'inertie mais simplement l'utilisation de la décomposition spectrale de l'opérateur $\mathbf{X}^t\mathbf{PX}$.

On fait une typologie de cartes en associant les variables qui ont même structure spatiale pour créer une carte de synthèse. La proposition de faire l'analyse ordinaire puis de conserver les coordonnées qui ont une bonne structure spatiale faite dans [8] est une introduction à cette pratique.



On peut remarquer la solution d'une difficulté associée à l'écriture :

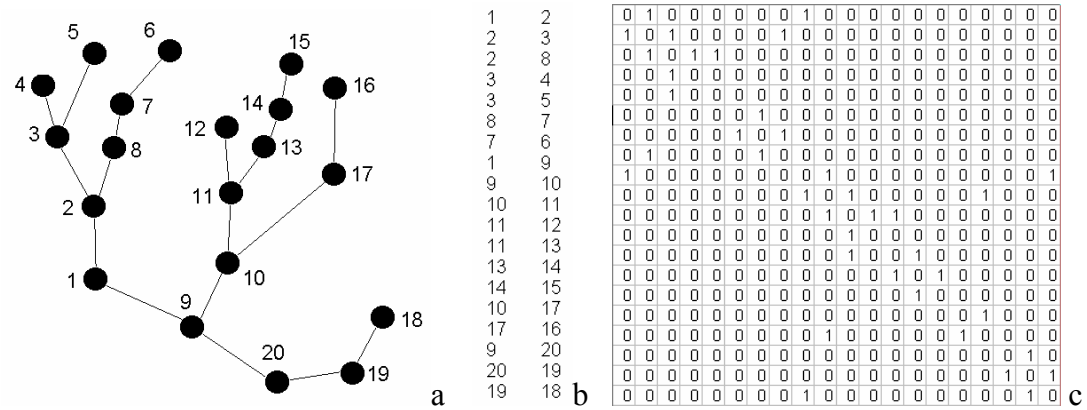
$$\mathbf{y}^t\mathbf{D}\mathbf{y} = \mathbf{y}^t\left(\mathbf{D}-\mathbf{P}\right)\mathbf{y} + \mathbf{y}^t\mathbf{P}\mathbf{y} \Rightarrow \mathbf{a}^t\mathbf{X}^t\mathbf{D}\mathbf{X}\mathbf{a} = \mathbf{a}^t\mathbf{X}^t\left(\mathbf{D}-\mathbf{P}\right)\mathbf{X}\mathbf{a} + \mathbf{a}^t\mathbf{X}^t\mathbf{P}\mathbf{X}\mathbf{a}$$

On pourrait penser que la décomposition implique que les dernières composantes de l'analyse locale sont les premières de l'analyse locale. Il n'en est rien car la somme n'est pas constante. Pour minimiser la variance locale, l'analyse locale minimise la variance simple sans augmenter de ce fait la covariance locale. L'analyse simple donne des composantes de variance maximale, l'analyse locale trouve les combinaisons les plus variables d'un voisin à l'autre et la troisième donne les composantes les plus lisses. Ce sont trois objectifs très dissemblables. 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. Dans ADE-4, on appelle l'analyse locale *analyse de Geary* et l'analyse en composantes cartographiables *analyse de Moran*.

4.3. Vecteurs propres de voisinages

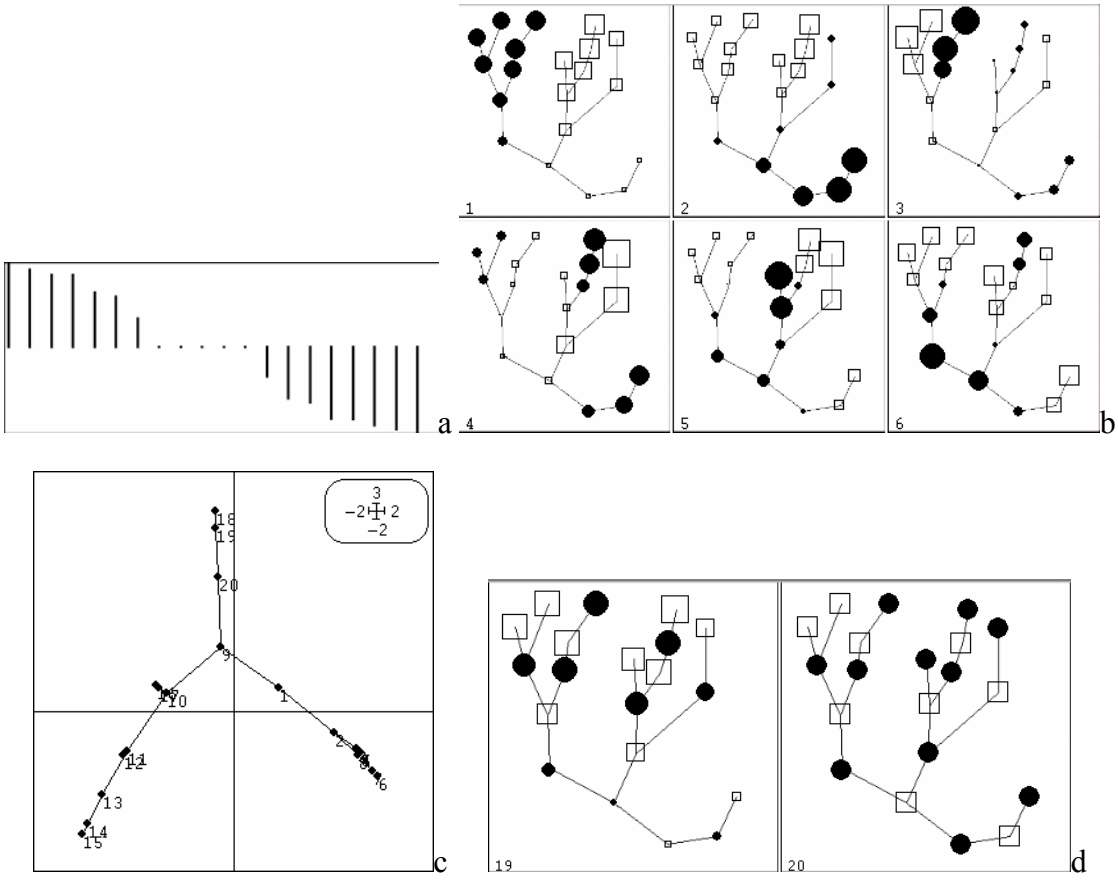
Les opérateurs de voisinage [5] donnent enfin une solution à une question complexe. Certains systèmes présentent une structure de voisinage très particulière qui ne supporte pas l'utilisation de coordonnées (x, y). La proximité spatiale de deux populations de poissons ne se mesure pas à vol d'oiseaux, mais le long du réseau avec éventuellement

un passage dans l'océan. Si on veut utiliser la stratégie de représentation de l'espace par un tableau, on a besoin de codes numériques qui ne sont pas des polynômes en x et y mais qui rendent comptent des déplacements potentiels dans l'espace étudié. Une solution générale est donnée par les vecteurs propres de **P**.



a - graphe de voisinage initial, b - liste des arêtes, c - matrice de voisinage.

La matrice **M** (20-20) définit la matrice **D** (poids de voisinages) et la matrice **P**. La diagonalisation de $\mathbf{D}^{-1}\mathbf{P}$ fournit une base de vecteurs propres **D**-orthogonaux, c'est à dire *n* (ici 20) codes numériques de moyennes nulles, de variances (pondération de voisinage) égales à 1, de covariances nulles et rangées par autocovariances (Moran) décroissantes. Les valeurs propres sont positives (scores lisses) puis nulles puis négatives. On utilise les premières pour numériser l'espace :



a - les valeurs propres , b - les 6 premiers vecteurs propres de voisinage, c - positions des points à l'aide des deux premiers vecteurs propres de voisinage, d - les 2 derniers vecteurs propres de voisinage sont d'indice de Geary maximum.

5. Tableaux, graphes et distances

Nous venons de voir une technique pour passer d'un graphe de voisinage à un tableau. Quand les données forment plusieurs tableaux appariés, l'espace y prendra place avec un tableau de coordonnées ou de vecteurs propres de voisinage. Les opérations inverses de tableaux à distances ou de distances à graphes de voisinage sont également possibles et utiles.

5.1. Les matrices de distances

Les données en présence-absence peuvent être transformées en matrices de distance. Deux objets (en écologie, lignes ou colonnes d'un tableau floro-faunistiques) sont comparés sur une liste de valeurs. Ces valeurs sont réduites en 0-1 (1 si la valeur est strictement positive, 0 sinon). Deux relevés sont ainsi comparés par la liste des espèces présentes, deux espèces sont comparées par la liste des relevés dans lesquels elles sont présentes. Ces listes ont la forme :

01100001010010...
01010001100010...

n est le nombre d'enregistrements, *a* est le nombre de concordances 11, *b* le nombre de concordances 10, *c* le nombre de concordances 01 et *d* le nombre de concordances 00. Ainsi deux espèces sont présentes ensemble dans un même relevé *a* fois, deux relevés possèdent *a* espèces en commun. Les deux objets définissent donc la table de contingence 2-2 :

	1	0	Tot
1	a	b	$a+b$
0	c	d	$c+d$
Tot	$a+c$	$b+d$	n

Les quatre nombres de la table définissent une similarité entre les deux objets. On peut utiliser :

$S_1 = \frac{a}{a+b+c}$ Indice de communauté de Jaccard

$S_2 = \frac{a+d}{n}$ Indice de Sokal & Michener

$S_3 = \frac{a}{a+2(b+c)}$ Indice de Sokal & Sneath

$S_4 = \frac{a+d}{a+2(b+c)+d}$ Indice de Rogers et Tanimoto

$S_5 = \frac{2a}{2a+b+c}$ Indice de Sorensen

$S_6 = \frac{a-(b+c)+d}{n}$ Indice de Gower & Legendre

$S_7 = \frac{a}{\sqrt{(a+b)(a+c)}}$ Indice de Ochiai

$S_8 = \frac{ad}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$ Indice de Sockal & Sneath

$S_8 = \frac{ad-bc}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$ Phi de Pearson

$S_{10} = \frac{a}{n}$ avec l'unité si les deux objets sont identiques

On trouvera les références d'origine dans **17**. Ces indices sont tous inférieurs ou égaux à 1 et la distance associée est définie par :

$$D_k = \sqrt{1 - S_k}$$

Les données quantitatives peuvent également être transformées en matrice de distances :

Distance de Manhattan, ou city block, ou de Gower 1971a (référence p. 20 dans **18**), ou D3 de Gower & Legendre **17** :

$$d_1(i,j) = \frac{1}{p}\sum_{k=1}^{p}\frac{|x_{ik} - x_{jk}|}{r_k} \text{ avec } r_k = \max_{i=1}^{n}(x_{ik}) - \min_{i=1}^{n}(x_{ik})$$

Distance de Manhattan, ou de Cain & Harrison (référence p. 20 dans **18**), ou D3 de Gower & Legendre **17** :

$$d_2(i,j) = \frac{1}{p}\sum_{k=1}^{p}\frac{|x_{ik} - x_{jk}|}{r_k} \text{ avec } r_k = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ik} - m_k)^2} \text{ et } m_k = \frac{1}{n}\sum_{i=1}^{n} x_{ik}$$

Distance de Canberra, ou de Lance & Williams (référence p. 20 dans **18**), ou D7 de Gower & Legendre **17** :

$$d_3(i,j) = \frac{1}{p}\sum_{k=1}^{p}\frac{|x_{ik} - x_{jk}|}{|x_{ik}| + |x_{jk}|}$$

Distance de Bray-Curtis ou de Odum (références p. 20 dans **18**), ou D8 de Gower & Legendre **17** :

$$d_4(i,j) = \frac{\frac{1}{p}\sum_{k=1}^{p}|x_{ik} - x_{jk}|}{\sum_{k=1}^{p} x_{ik} + x_{jk}}$$

Distance D5 de Gower & Legendre **17** pour données positives seulement :

$$d_5(i,j) = \frac{1}{p}\sum_{k=1}^{p}\frac{(x_{ik}-x_{jk})^2}{(x_{ik}+x_{jk})^2}$$

Distance D9 de Gower & Legendre [17] pour données positives seulement :

$$d_6(i,j) = \frac{\sum_{k=1}^{p}|x_{ik}-x_{jk}|}{\sum_{k=1}^{p}\max(x_{ik},x_{jk})}$$

Distance D10 de Gower & Legendre [17] pour données positives seulement :

$$d_7(i,j) = \frac{1}{p}\sum_{k=1}^{p}\left(1-\frac{\min(x_{ik},x_{jk})}{\max(x_{ik},x_{jk})}\right)$$

Les données en proportion ont également leurs distances spécifiques, parmi lesquelles on trouve les distances génétiques sur un locus :

Distance d1 (voir [11] formule (5.7) p. 68).

$$d_1 = \frac{1}{2}\sum_i |p_i - q_i|$$

Distance d2 (indice de chevauchement de niche, voir [11] formule (5.8) p. 68).

$$d_2 = 1 - \frac{\sum_i p_i q_i}{\sqrt{\sum_i p_i^2}\sqrt{\sum_i q_i^2}}$$

Distance de Rogers :

$$d_3 = \frac{1}{2}\sqrt{\sum_i (p_i - q_i)^2}$$

Distance de Nei :

$$d_4 = -\ln\left(\frac{\sum_i p_i q_i}{\sqrt{\sum_i p_i^2}\sqrt{\sum_i q_i^2}}\right)$$

Distance de Edwards :

$$d_5 = \sqrt{1-\sum_i \sqrt{p_i q_i}}$$

Les distances génétiques s'étendent aux données multilocus. Soit **A** un tableau de fréquences alléliques avec t lignes (populations) et m colonnes (allèles). Soit v le nombre de loci. Le locus j a $m(j)$ allèles.

$$m = \sum_{j=1}^{v} m(j)$$

Pour la $i^{\text{ème}}$ ligne et la $k^{\text{ème}}$ modalité de la variable j, on note la valeur a_{ij}^k ($1 \le i \le t$, $1 \le j \le v$, et $1 \le k \le m(j)$), la valeur du tableau des données brutes. Soit :

$$a_{ij}^{+} = \sum_{k=1}^{m(j)} a_{ij}^k \quad \text{et} \quad p_{ij}^k = \frac{a_{ij}^k}{a_{ij}^{+}}$$

Soit le tableau $\mathbf{P} = \left[p_{ij}^k\right]$ et les paramètres :

$$p_{ij}^{+} = \sum_{k=1}^{m(j)} p_{ij}^k = 1, p_{i+}^{+} = \sum_{j=1}^{v} p_{ij}^{+} = v, p_{++}^{+} = \sum_{j=1}^{v} p_{i+}^{+} = tv$$

Les plus connues sont :

Distance de Rogers [19] (Voir [20]) :

$$D_1(a,b) = \frac{1}{v}\sum_{k=1}^{v}\sqrt{\frac{1}{2}\sum_{j=1}^{m(k)}\left(p_{aj}^k - p_{bj}^k\right)^2}$$

Distance de Nei [21] (Voir [20]) :

$$D_2(a,b) = -Ln\left(\frac{\sum_{k=1}^{v}\sum_{j=1}^{m(k)} p_{aj}^k p_{bj}^k}{\sqrt{\sum_{k=1}^{v}\sum_{j=1}^{m(k)}\left(p_{aj}^k\right)^2}\sqrt{\sum_{k=1}^{v}\sum_{j=1}^{m(k)}\left(p_{bj}^k\right)^2}}\right)$$

Distance de Edwards [22] (Voir [23]) :

$$D_3(a,b) = \sqrt{\frac{1}{v}\sum_{k=1}^{v}\left(1 - \sum_{j=1}^{m(k)}\sqrt{p_{aj}^k p_{bj}^k}\right)}$$

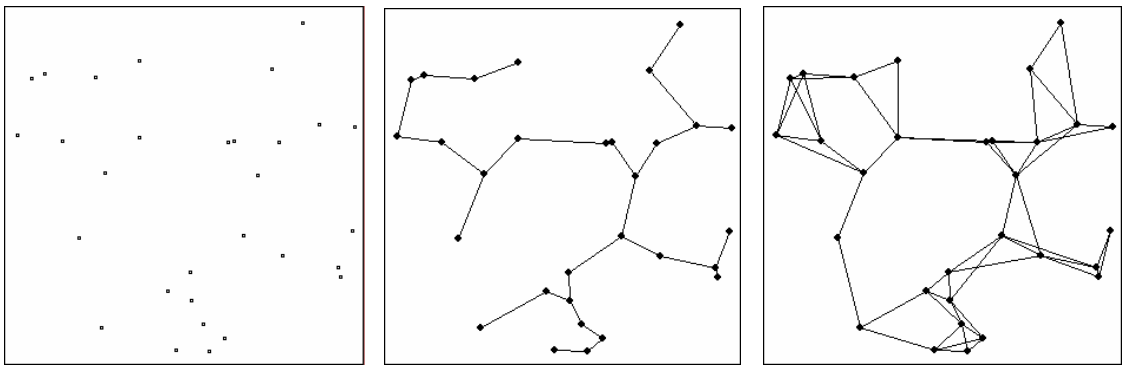
Enfin, toute méthode d'analyse de données basée sur un schéma du type :

$$\begin{array}{ccc} \boxed{p} & \overset{\mathbf{Q}}{\rightarrow} & \boxed{p} \\ \mathbf{X}^t \uparrow & & \downarrow \mathbf{X} \\ \boxed{n} & \underset{\mathbf{D}}{\leftarrow} & \boxed{n} \end{array}$$

définit une matrice de distances par la relation fondamentale $d(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|_{\mathbf{Q}}$. La liste n'est en rien exhaustive.

5.2. L'arbre de longueur minimale

On utilise en général les matrices de distances dans les classifications. Une pratique intéressante consiste à les transformer en graphe de voisinage par l'observation suivante.

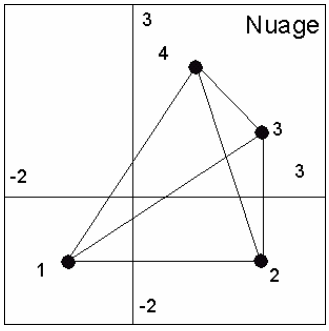


Si *n* points donnent une matrice de distances (*n*-*n*) $\mathbf{D} = \left[d_{ij} \right]$, on peut associer à chaque graphe sur cet ensemble sa longueur, c'est à dire la somme des longueurs de ses arêtes en attribuant à la longueur de l'arête qui relie *i* à *j* la distance d_{ij}. Parmi les graphes connexes (ceux formés d'une seule composante connexe qui permettent de passer d'un point quelconque à un autre par une suite d'arêtes), il en existe un ou plusieurs qui présentent une longueur minimale. Ce ou ces graphes optimaux sont sans cycles (suite d'arêtes qui boucle) car on pourrait enlever une arête et diminuer la longueur. Un graphe connexe et sans cycle est appelé un arbre et on obtient le ou les arbres de longueur minimale (si toutes les distances sont différentes, la solution est unique). Voir dans [24] (p. 163-166) la présentation de trois algorithmes et leurs justifications. On peut rechercher un nouvel arbre de longueur minimale qui n'utilise aucune des arêtes du précédent et ainsi de suite. On obtient des graphes de voisinage qui rendent compte de la matrice de distances. On se sert des arbres de longueur minimale pour introduire la statistique non paramétrique sur des données multivariées [25]. On représente aussi les arbres de longueur minimale sur les cartes factorielles pour compenser les déformations de la projection.

5.3. Représentations euclidiennes

Les tableaux de données peuvent être transformés en matrices de distances. L'inverse est il possible ? Pas toujours, mais l'opération, quand elle est possible est fort utile. La question se pose ainsi : quand on a un tableau de données à 2 colonnes on peut représenter un nuage de point et définir une matrice de distances :

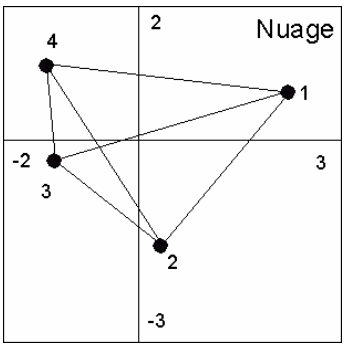
-1	-1
2	-1
2	1
1	2



0	3	3.6056	3.6056
3	0	2	3.1623
3.6056	2	0	1.4142
3.6056	3.1623	1.4142	0

Quand on a la matrice de distance, comment savoir si il existe un nuage de points dont on a précisément les distances deux à deux ? Et s'il existe comment le construire et retrouver un tableau :

0	3	3.6056	3.6056
3	0	2	3.1623
3.6056	2	0	1.4142
3.6056	3.1623	1.4142	0



2.2392	0.7407
0.3233	-1.5678
-1.2158	-0.2905
-1.3466	1.1176

Le problème et la solution sont de Gower [26].

En général, la question est : étant donnée une matrice de distances $\mathbf{D}=\left[d_{ij}\right]$ entre *n* entités existe t'il *n* points M_i dans un espace euclidien de dimension *p* tels que :

$$d_{ij}^2=\left\|M_i-M_j\right\|^2 \ ?$$

Si c'est le cas, l'espace possède une base orthonormée dans laquelle le point M_i a des coordonnées qu'on peut mettre sur la ligne *i* d'un tableau **X**. La matrice $\mathbf{XX}^t$ contient alors les produits scalaires $\left\langle M_i \middle| M_j \right\rangle$. Or :

$$d_{ij}^2=\left\|M_i-M_j\right\|^2=\left\|M_i\right\|^2+\left\|M_j\right\|^2-2\left\langle M_i \middle| M_j \right\rangle$$

Sans perte de généralités on peut supposer que le centre de gravité du nuage est à l'origine (sinon on le translate) et donc que $\sum_{i=1}^{n} M_i=0$. On considère alors la matrice des valeurs $\mathbf{H}=\left[-\frac{1}{2}d_{ij}^2\right]$. On calcule la moyenne par lignes m_i, la moyenne par colonne m_j et la moyenne générale *m* :

$$m_i=-\frac{1}{2}\left\|M_i\right\|^2-\frac{1}{2n}\sum_{j=1}^{n}\left\|M_j\right\|^2 \quad m_j=-\frac{1}{2}\left\|M_j\right\|^2-\frac{1}{2n}\sum_{i=1}^{n}\left\|M_i\right\|^2 \quad m=-\frac{1}{n}\sum_{i=1}^{n}\left\|M_i\right\|^2$$

La matrice **H** centrée par ligne et par colonne vaut alors :

$$\mathbf{H_{\bullet\bullet}}=\left[-\frac{1}{2}d_{ij}^2-m_i-m_j+m\right]=\mathbf{XX}^t$$

Toutes ses valeurs propres sont positives ou nulles. Réciproquement si la matrice des carrés des distances doublement centrées a toutes ses valeurs positives ou nulles :

$$\mathbf{H_{\bullet\bullet}}=\left[-\frac{1}{2}d_{ij}^2\right]_{\bullet\bullet}=\mathbf{U}\Lambda\mathbf{U}^t=\mathbf{U}\Lambda^{\frac{1}{2}}\Lambda^{\frac{1}{2}}\mathbf{U}^t=\mathbf{XX}^t$$

Les carrés des distances entre les lignes de **X** sont les carrés des distances d'origine. Donc pour une matrice de distance **D** quelconque de deux choses l'une :

- ou bien $\left[-\frac{1}{2}d_{ij}^2\right]_{\bullet\bullet}$ qui est symétrique a toutes ses valeurs propres positives ou nulles. On dit qu'elle est *euclidienne*. La matrice $\mathbf{X} = \mathbf{U}\Lambda^{\frac{1}{2}}$ donne les coordonnées d'un nuage de points dont les distances sont celles de départ. **X** est dite *représentation euclidienne* de **D**. Ces coordonnées sont rangées par ordre décroissant de variance et si on utilise les premières pour voir une projection du nuage sur ses axes principaux on dit qu'on fait une *Analyse en coordonnées principales (PCOA)*.

- ou bien $\left[-\frac{1}{2}d_{ij}^2\right]_{\bullet\bullet}$ qui est symétrique a des valeurs propres négatives. On dit qu'elle n'est pas *euclidienne*. La représentation euclidienne n'existe pas et la PCOA n'a pas lieu d'être.

En pratique on fait souvent "comme si" en ignorant le problème des valeurs propres négatives. Gower et Legendre [27] ont étudié le caractère euclidien ou non de nombreuses distances et on sait souvent si on choisit une méthode de calcul qui donnera ou non une matrice de distances euclidiennes. Par exemple toutes les distances basées sur des indices de similarité sont euclidiennes alors que les distances les plus utilisées en génétique ne le sont pas.

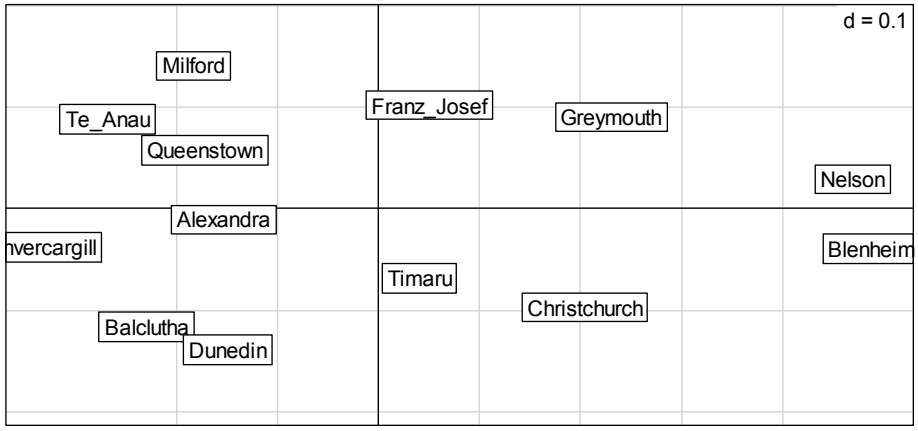
Deux résultats fondamentaux confortent l'intérêt des distances euclidiennes.

- Quand une matrice de distances n'est pas euclidienne, on sait calculer [28] la plus petite constante *c* qui assure que la distance $d_{ij} + c \quad \forall i \neq j$ est euclidienne.

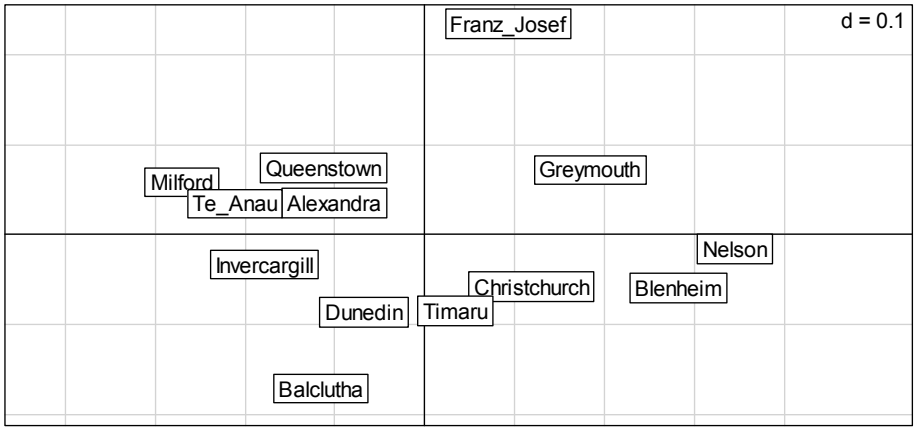
- Quand une matrice de distances n'est pas euclidienne, on sait calculer [29] la plus petite constante *c* qui assure que la distance $\sqrt{d_{ij}^2 + c} \quad \forall i \neq j$ est euclidienne.

a

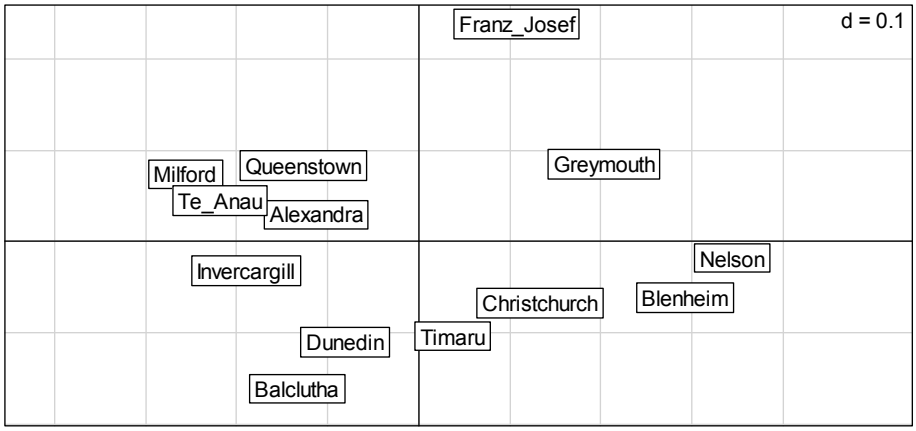
b



c



d



e

a - Carte de disposition de 13 *villes de Nouvelle-Zélande. b - Représentation des coordonnées cartésiennes après digitalisation. c - Représentation euclidienne de la distance canonique calculée sur les coordonnées cartésiennes. La distance routière n'est pas euclidienne. d - Représentation euclidienne (plan 1-2) de la matrice obtenue par la transformation de Lingoes. e - Représentation euclidienne (plan 1-2) de la matrice obtenue par la constante additive de Cailliez. La première transformation est conseillée dans* [30].

On peut donc, en analyse multivariée de données spatialisées introduire les données soit comme tableaux soit comme distances et l'espace soit comme tableau, soit comme distance, soit comme graphe de voisinage. Il s'en suit une très grande diversité des pratiques et l'absence de méthodes standardisées.

6. Un exemple

Un exemple très caractéristique de la diversité des méthodes couplant espace et multivarié est proposé dans [31]. Le justificatif 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 generaly 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 massivement multivarié (à deux niveaux) avec l'espace.

> *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. Les distances génétiques sont déterminées en général entre groupes ou populations au sens large à partir des fréquences alléliques dans les groupes alors qu'ici on a besoin d'une distance entre individus. Le codage est du type 002000 pour un homozygote et du type 010100 pour un hétérozygote. Pour un locus la distance proposée entre deux individus x et y est la moitié de la métrique euclidienne canonique :

$$d_{xy}^2 = \frac{1}{2}\sum_{k=1}^{K}\left(x_k - y_k\right)^2$$

On peut introduire une pondération pour tenir plus compte des allèles rares avec :

$$d_{xy}^2 = \frac{1}{2}\sum_{k=1}^{K} w_k \left(x_k - y_k\right)^2 \text{ avec } w_k = \frac{1}{2Kp_k} \text{ et la recommandation } p_k = \frac{N_k + \frac{1}{k}}{2N+1}.$$

On reconnaît, à une constante près, la métrique de l'analyse des correspondances modifiée par un argument d'estimateur moins biaisé. L'essentiel est que les données sont importées dans l'analyse par une matrice de distance euclidienne, canonique ou non. On a une distance par locus et une distance totale soit $L + 1$ matrices de distances. Les matrices de distances sont directement écrites avec les carrés :

$$\mathbf{D}_1 = \left[d_{ij1}^2\right], \mathbf{D}_2 = \left[d_{ij2}^2\right], ..., \mathbf{D}_L = \left[d_{ijL}^2\right], \mathbf{D} = \left[d_{ij1}^2 + d_{ij2}^2 + ... + d_{ijL}^2\right]$$

Le problème des K-tableaux se pose ici comme problèmes des K-distances en débat en statistique mathématique depuis **32**. L'analyse se poursuivant pour chacune des distances et la distance globale, nous en conservons une, par exemple **D**.

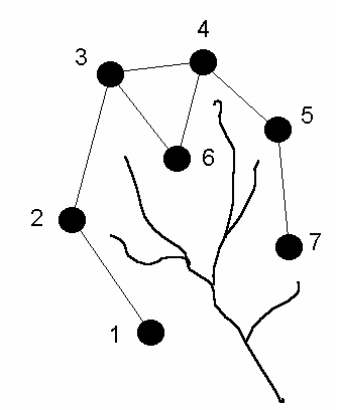
En se référant à **26**, les auteurs définissent alors la *genetic covariance matrix* **C** par :

$$c_{ij} = \left[-d_{ij}^2 + \left(\sum_{i=1}^{N} d_{ij}^2 + \sum_{j=1}^{N} d_{ij}^2 \right) \Big/ N - \left(\sum_{i \neq j}^{N} d_{ij}^2 \right) \Big/ N^2 \right] / 2$$

soit exactement la matrice $\mathbf{C} = \left[-\frac{1}{2} d_{ij}^2 \right]_{\bullet\bullet} = \mathbf{X}\mathbf{X}^t$.

Donc les données sont traitées par la *matrice des produits scalaires de la représentation euclidienne* associée à la distance génétique.

Est alors 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 :



$$\mathbf{X}^{(1)} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 2 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 3 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 3 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 2 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \end{bmatrix} \quad \mathbf{X}^{(2)} = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 2 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 2 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 2 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 2 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 \end{bmatrix} \ldots$$

On reconnaît les matrices du type **N + 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) etc.

Ceci permet d'obtenir une autocorrélation spatiale au pas *h* par (formule 15 p. 566) :

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

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

$$r^{(h)} = \frac{Trace\left(\mathbf{MXX}^t\right)}{Trace\left(\mathbf{NXX}^t\right)} = \frac{Trace\left(\mathbf{X}^t\mathbf{PX}\right)}{Trace\left(\mathbf{X}^t\mathbf{DX}\right)}$$

On est donc exactement sur l'indice de Moran multivarié avec un test de permutation du type test de Mantel exécuté sur la représentation euclidienne. Dans la version d'ADE-4 les variables sont normalisées pour la pondération de voisinage et le dénominateur est égal au nombre de variables. Ici on fait le rapport de la covariance de voisinage à la variance totale, ce qui est un choix aussi légitime. Une question est ouverte : on peut se demander si le passage matrice de données vers matrice de distance puis matrice de distances vers matrice de produits scalaires ne donne pas le même résultat que le passage direct matrice de données vers matrice de produits scalaires.

7. Références

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

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

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

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

5 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.

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

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

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

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

10 Mantel, N. (1967) The detection of disease clustering and a generalized regression approach. *Cancer Research* : 27, 209-220.

11 Manly, B.F. (1994) *Multivariate Statistical Methods. A primer*. Second edition. Chapman & Hall, London. 1-215.

12 Gros, G. (1978) *Structure et échantillonnage des peuplements spontanés de framboisiers (Rubus Idaeus L.) dans les Vosges*. Thèse de 3° cycle, INRA, Colmar. 1-60.

13 Besse, Ph. (1979) *Etude descriptive d'un processus ; approximation, interpolation*. Thèse de 3° cycle, Université Paul Sabatier. Toulouse.

14 Champely, S. (1994) *Analyse de données fonctionnelles - Approximation par les splines de régression*. Thèse de Doctorat, Université Lyon 1. 1-250.

15 Banet, T.A. & Lebart, L. . (1984) Local and Partial Principal Component Analysis (PCA) and Correspondence Analysis (CA). In : *COMPSTAT 84*. International Association for Statistical Computing. (Ed.) Physica-Verlag, Vienna. 113-123.

16 Manly, B.F.J. (1991) *Randomization and Monte Carlo methods in biology*. Chapman and Hall, London. 1-281.

17 Legendre, L. & Legendre, P. (1984b) *Ecologie numérique*. Tome 2 - La structure des données écologiques. Masson, Paris. 2ème édition revue et augmentée : 1-344. Voir p. 6 et suivantes.

18 Digby, P. G. N. & Kempton, R. A. . (1987*) Multivariate Analysis of Ecological Communities*. Chapman and Hall, Population and Community Biology Series, London. 1-205. (p.96).

19 Rogers, J.S. (1972) Measures of genetic similarity and genetic distances. *Studies in Genetics*, Univ. Texas Publ. 7213: 145-153.

20 Avise, J.C. (1994) *Molecular markers, natural history and evolution*. Chapman & Hall, London. 1-511 (p. 95).

21 Nei, M. (1972) Genetic distances between populations. *American Naturalist* : 106. 283-292.

22 Edwards, A.W.F. (1971) Distance between populations on the basis of gene frequencies. *Biometrics* : 27, 873-881.

23 Hartl, D.L. & Clark, A.G. (1989) *Principles of population genetics*. Sinauer Associates, Sunderland, Massachussetts. 1-682 (p. 303).

24 Lebart, L., Morineau, A. & Piron, M. (1995) *Statistique exploratoire multidimensionnelle*. Dunod, Paris. 1-439.

25 Friedman, J.H. & Rafsky, L.C. (1979) Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. *The Annals of Statistics* : 7, 697-717.

26 Gower, J.C. (1966) Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika* : 53, 325-338.

27 Gower, J.C. & Legendre, P. (1986) Metric and Euclidean properties of dissimilarity coefficients. *Journal of Classification* : 3, 5-48.

28 Cailliez, F. (1983) The analytical solution of the additive constant problem. *Psychometrika* : 48, 305-310.

29 Lingoes, J.C. (1971) Somme boundary conditions for a monotone analysis of symmetric matrices. *Psychometrika* : 36, 195-203.

30 Legendre, P. & Anderson, M.J. (1999) Distance-based redundancy analysis: testing multispecies responses in multifactorial ecological experiments. *Ecological Monographs* : 69, 1-24.

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

32 Gower, J.C. (1975) Generalized procustes analysis. *Psychometrika* : 40, 33-51.