Fiche de Biostatistique - Stage 8

Ordination sous contrainte spatiale

D. Chessel, S. Ollier & S. Dray

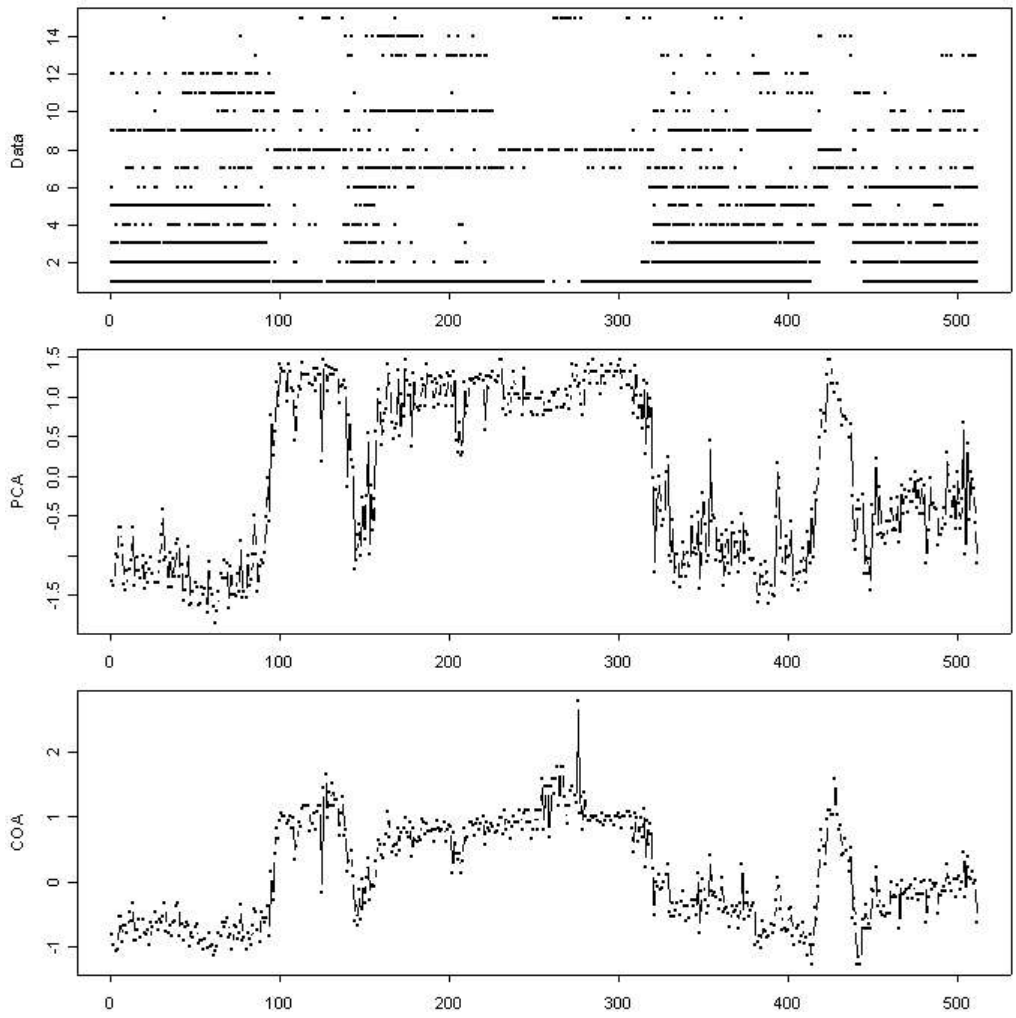
Résumé

La fiche décrit une méthode d'ordination sous contrainte spatiale. Elle a été écrite pour un cours intitulé "L'ACP avec contraintes spatiales" dans le cadre d'un stage "Formations aux statistiques spatiales" organisé par M. Arnaud au CIRAD (Montpellier 10/03 - 14/03/2003). Elle a été mise à jour pour la version 2004 de ce stage. On décrit les objets de base 'relations et pondérations de voisinage' en suivant la structure de la programmathèque **spdep** que R. Bivand a écrit pour le logiciel R. Est détaillé le fonctionnement de la procédure **multispati** généralisation dans la logique de la programmathèque **ade4** de la méthode de Wartenberg (Wartenberg, D. E. (1985) Multivariate spatial correlations: a method for exploratory geographical analysis. *Geographical Analysis* **17**: 263-283). Ce choix est commenté et illustré. Pour utiliser les illustrations charger et connecter les dernières versions de **tripack** (triangulation de Delaunay) **maptools** (outils graphiques d'unités surfaciques) **pixmap** (manipulations d'images bitmap) **spdep** (graphes et matrices de contiguïtés, statistique spatiale) **splancs** (processus de points) et **ade4** (analyse multivariée).

Plan

1. Quelques situations expérimentales

On se place ici en analyse des données multidimensionnelles quand l'information est disponible dans un grand nombre de descripteurs qui ont ou n'ont pas de significations personnelles particulières. 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) comme l'un des premiers articles sur l'AFC en écologie (Hatheway 1971) cartographie des coordonnées factorielles et notent l'efficacité de cette pratique. Hill (1974) puis Estève (1978) représentent des coordonnées factorielles le long d'un transect tout comme Dessier et Laurec (1978) les représentent en fonction du temps. Dans tous les cas, sans introduire la structure du plan d'observation (stations sur une carte, placettes sur un transect, prélèvements dans une chronique), on obtient avec les analyses classiques une expression parfaitement satisfaisante des résultats exprimés dans cette structure. On peut pour s'en convaincre reprendre l'exemple traité par J. Estève dans l'article précité.



Présence-absence de 15 espèces et coordonnées factorielles le long d'un transect de 512 placettes en steppe semi-aride.

```
data(steppe) # dans la librairie ade4
names(steppe)
[1] "tab"       "esp.names"
steppe$esp.names[1:15]
         "Noaea mucronata"         "Plantogo albicans"
  "Hernaria fontananensii"          "Stipa parviflora"
     "Helianthemum hirtum"               "Poa bulbosa"
    "Anabasis Oropediorum"       "Salsola vermiculata"
"Atractylis serratuloïdes"     "Artemisia Herba Alba"
  "Pithuranthos scoparium"           "Teucrium polium"
        "Fagonia kaherina"            "Lygeum spartum"
         "Peganum harmala"
```

Steppe$tab contient 512 relevés avec 37 espèces végétales d'une steppe semi-aride. Les 512 placettes sont alignées sur un transect de 5 km (Chessel and Donadieu 1977).

```
par(mfrow = c(3,1))
par(mar=c(2.1,4.1,1.1,1.1))
w1 <- col(as.matrix(steppe$tab[,1:15]))
w1 <- as.numeric(w1[steppe$tab[,1:15] > 0])
w2 <- row(as.matrix(steppe$tab[,1:15]))
w2 <- as.numeric(w2[steppe$tab[,1:15] > 0])
plot(w2, w1, pch = 20, ylab="Data", xlab="", cex=0.75)
plot(dudi.pca(steppe$tab, scan = FALSE, scale = FALSE)$li[,1],
         pch = 20, ylab = "PCA", xlab = "", type = "b", cex=0.75)
plot(dudi.coa(steppe$tab, scan = FALSE)$li[,1], pch = 20,
         ylab = "COA", xlab = "", type = "b", cex=0.75)
```

La figure restitue l'évolution, le long du transect, de la présence des 15 espèces principales puis celle de la première coordonnée de chaque analyse (ACP et AFC). 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.

On peut concevoir le problème de l'ordination spatiale comme celui de la synthèse d'un ensemble de cartes. On aborde ici la question remarquablement posée par Goulard et al. (1987) :

> *Tant au niveau de la description que de l'estimation, les méthodes géostatistiques se révèlent conceptuellement très adaptées mais leur application pratique n'est efficace que dans le cas d'un petit nombre de variables régionalisées stationnaires.*
>
> *Pour un nombre important de variables, il apparaît que les méthodes multidimensionnelles d'analyse des données sont encore les seules utilisables d'un point de vue concret. Elles peuvent être utilisées pour dégager les variables qui seront soumises ensuite à l'étude géostatistique.*

On retrouve partout cette omniprésence naturelle de l'espace dans les données multivariées. L'article de Grande et al. (2000) porte dans son titre *using factor analysis* mais dans la page de l'équipe (http://www2.uhu.es/rcagua/PublicacionesCongresos.htm) l'article en révision porte dans le titre *using spatial factor analysis*.

Cette intervention active de l'espace, qui semble souvent 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.

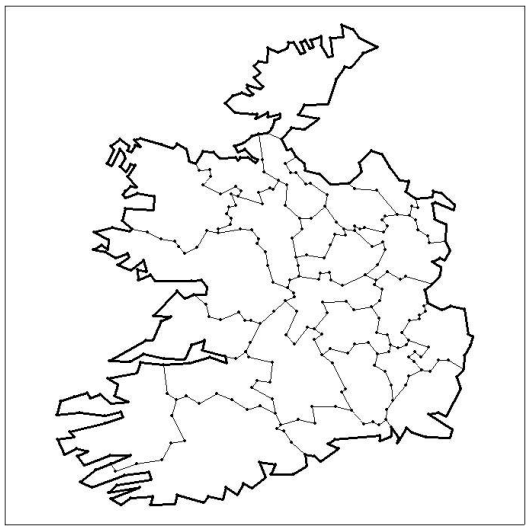
On intègre l'espace de multiples manières évidemment. Une des plus simples est de prendre deux coordonnées (x_i, y_i) pour chaque unité statistique ce qui associe à chaque couple de points une distance $d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$. En économie, les données sont généralement associées à des unités administratives, donc des enregistrements surfaciques qui supportent mal une telle réduction. C'est ainsi que la question s'est posée initialement. 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). Le matériel nécessaire est donné dans la liste `irishdata` de la librairie **ade4** dans laquelle on a dupliqué une partie des informations de la liste `eire` de la librairie **spdep**. Ceci évite des manipulations pour éliminer le contenu du compté de Dublin qui est trop étranger au reste du territoire pour ne pas perturber les analyses multivariées.

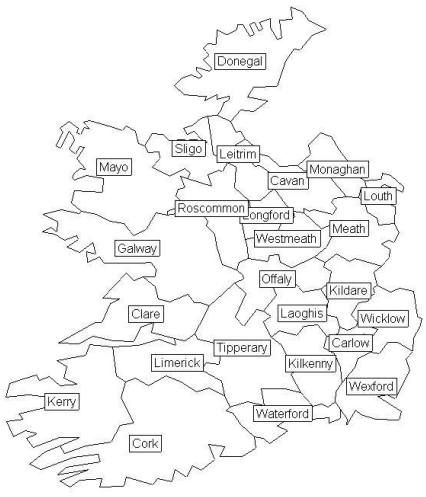
```
data(irishdata)
names(irishdata)
 [1] "area"         "county.names" "xy"           "tab"          "contour"
 [6] "link"         "area.utm"     "xy.utm"       "link.utm"     "tab.utm"
[11] "contour.utm"
```

irishdata$area contient des polygones de contour de 25 unités surfaciques :

```
   poly   x   y
1   S01 168  97
2   S01 178 101
3   S01 180  89
4   S01 168  68
5   S01 162  91
. . .
```

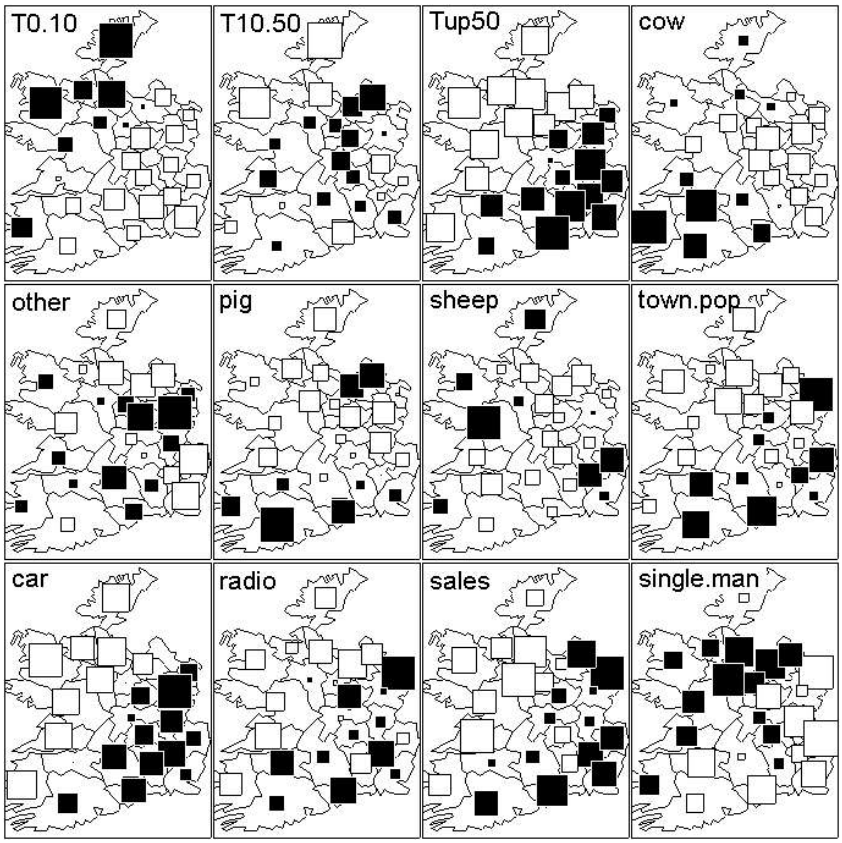
```
area.plot(irishdata$area.utm)
apply(irishdata$contour.utm, 1,
      function(x) segments(x[1],x[2],x[3],x[4], lwd = 3))
s.label(irishdata$area.utm[,2:3],cpoi=1,clab=0,add.p=T)
area.plot(irishdata$area.utm,
      lab = row.names(irishdata$xy.utm), clab = 1)
```






```
carto.irish <- function(x) {
      def.par <- par(no.readonly = TRUE)
      on.exit(par(def.par))
```

```
      par(mfrow=c(3,4))
      ncol = min(12,ncol(x))
      for (i in 1:12) {
            s.value(irishdata$xy.utm,x[,i],sub=names(x)[i],
possub="topleft",csub=3,

csi=2.5,cleg=0,area=irishdata$area.utm,include=F,addax=F,grid=F)
      }
}
w = as.data.frame(scalewt(irishdata$tab))
carto.irish(w)
```



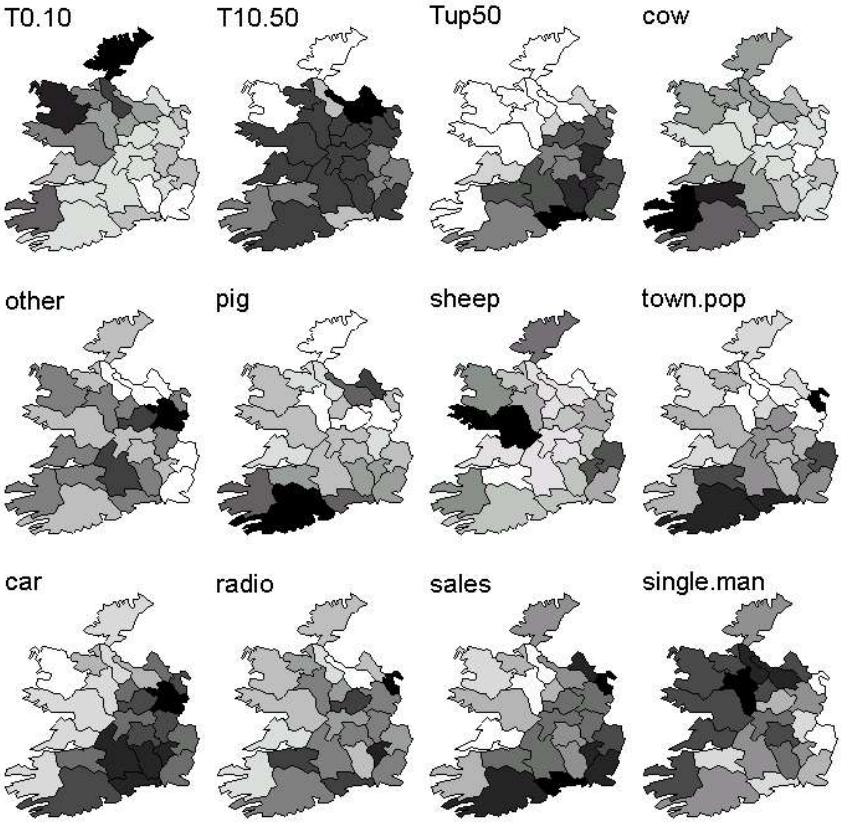
Cartographie par valeurs. 25 districts d'Irlande. Le district de Dublin est extrait du jeu de données. Tableau de 12 variables mesurées sur les 24 districts. Données célèbres reprises dans Cliff and Ord (1973 p. 53). Code des variables : 1-2-3 répartition (en 1 pour 1000) des propriétés agricoles en 3 groupes d'imposition (T0.10 <10 £, T10.50 10-50 £, Tup50 >50 £). 4-5-6-7 Nombres moyens d'animaux pour 1000 acres de prairies et cultures respectivement 4- cow vaches laitières, 5- other autres bestiaux, 6- pig cochons, 7- sheep moutons. 8- town.pop Pourcentage de population urbanisée (villes et villages) en 1 pour 1000 9- car Nombre de voitures pour 1000 habitants 10- radio Nombre de licences de radio pour 1000 habitants 11- sales Ventes de détail moyenne par habitant en £ 12- single.man Pourcentage de célibataires parmi les hommes de 30-34 ans en 1 pour 1000. Données normalisées.

Les données d'origine dans l'article fondateur de Geary sont curieusement multivariées. Si on sait faire l'analyse du tableau – ici, une analyse en composantes principales normée – la question est de reproduire cette analyse en l'optimisant du point de vue de l'intégration de l'espace sous-jacent au découpage administratif.

```
carto.irish1 <- function (x) {
    def.par <- par(no.readonly = TRUE)
    on.exit(par(def.par))
    par(mfrow=c(3,4))
    ncol = min(12,ncol(x))
    for (i in 1:12) {
```

```
                    area.plot(irishdata$area.utm,val   =   x[,i],sub=names(x)
[i],
                    possub="topleft",csub=3,cleg=0)
    }
}

carto.irish1 (w)
```

La même idée de synthèse multivariée se pose dans la figure :



Cartographie par unités surfaciques.

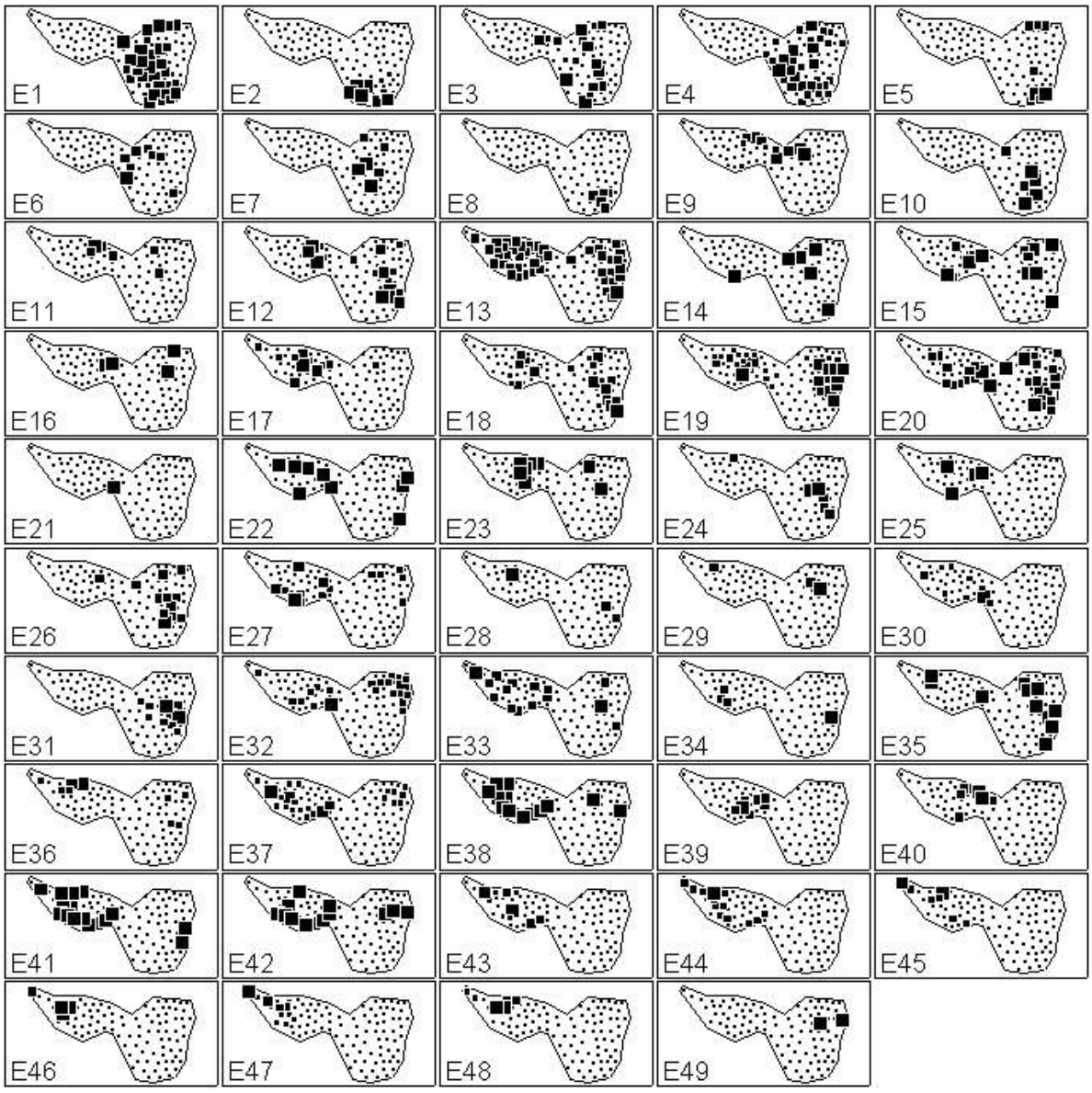
La même question est posée dans les données phyto-écologiques : le descripteur phytosociologique massivement multivarié porte sur des mesures élémentaires très simples (présence-absence 0-1, notes d'abondance-dominance entière 0-7, classe de recouvrement ou codage semi-quantitatif).

```
data(mafragh) # librairie ade4

flo = mafragh$flo
flo = mafragh$flo[,apply(mafragh$flo,2,sum)>3]

flonames = mafragh$espnames[apply(mafragh$flo,2,sum)>3]

dim(flo)
[1] 97 49
```

On a 97 sites et 49 espèces dont l'abondance vaut plus de 3.

```
carto.mafragh <- function (x=mafragh$flo) {
    def.par <- par(no.readonly = TRUE)
    on.exit(par(def.par))
```

```
                                                                    w1=c
(253,208,144,75,23,13,81,140,191,208,237,261,301,351,377,378,399,387,300,
252)
                                                  w2          =          c
(156,180,199,199,222,209,125,96,115,111,53,9,1,10,51,93,147,190,195,156)
    w3 = as.factor(rep("1",20))
    wessai = cbind.data.frame(w3,w1,w2)
    par(mfrow=c(10,5))
    ncol = min(49,ncol(x))
    for (i in 1:ncol) {
        area.plot(wessai,sub=names(x)[i],possub="bottomleft",csub=2.5)
        s.label(mafragh$xy,add.p=T,cpoi=0.25,clab=0)
        s.value(mafragh$xy,x[,i],add.p=T,csi=1,cleg=0)}
}
carto.mafragh()
```



49 espèces végétales non rares dans une enquête phytoécologique à vocation d'aménagement sur une plaine côtière derrière un cordon dunaire (de Belair 1981, de Belair and Bencheikh-Lehocine 1987) comportant 97 relevés (16x8 km) floristiques et autant d'analyses de sol.

Ce type d'expérience a été souvent reproduit en écologie. Utiliser la liste **oribatid**. Elle reproduit les données mises à disposition par P. Legendre à :

http://www.fas.umontreal.ca/biol/casgrain/fr/labo/oribates.html

Leur description est complète dans Borcard et al. (1992) et Borcard et Legendre (1994).

```
data(oribatid)
names(oribatid)
[1] "fau"   "envir" "xy"
```

Copyright, Ray Norton

http://www.fcps.k12.va.us/StratfordLandingES/Moran%20Website/mpages/soil_mite.htm

On a un tableau faunistique avec 35 espèces (colonnes) x 70 éléments d'échantillonnage (lignes). Les éléments d'échantillonnage sont des carottes de sol de 5 cm de diamètre et 10 cm de profondeur.

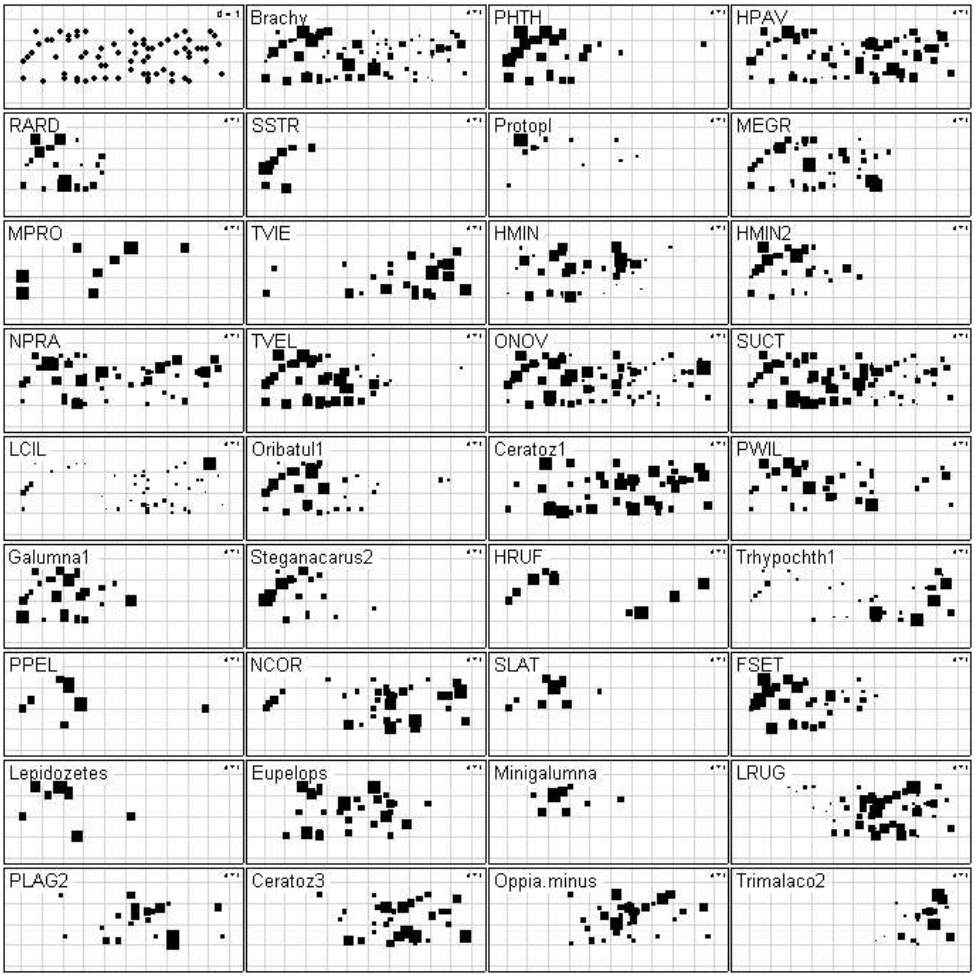
On a les coordonnées dans l'espace (en xy) des carottes.

On a un tableau de 70 carottes (lignes) et 5 variables environnementales (colonnes) :

```
names(oribatid$envir)
[1] "substrate" "shrubs"    "topo"      "density"   "water"
summary(oribatid$envir)
 substrate   shrubs        topo       density         water
 inter :27   few :26   blanket:44   Min.   :21.2   Min.   :134
 litter: 2   many:25   hummock:26   1st Qu.:30.0   1st Qu.:314
 peat  : 2   none:19                Median :36.4   Median :398
 sph1  :25                          Mean   :39.3   Mean   :411
 sph2  :11                          3rd Qu.:46.8   3rd Qu.:493
 sph3  : 1                          Max.   :80.6   Max.   :827
 sph4  : 2
```

substrate : Substrat - facteur à 7 modalités sph1 (Sphaignes, groupe d'espèces 1), sph2 (Sphaignes, groupe d'espèces 2), sph3 (Sphaignes, groupe d'espèces 3), sph4 (Sphaignes, groupe d'espèces 4), litter (litière), peat (tourbe nue) et inter (interfaces)).
shrubs : Buissons - facteur à 3 modalités none (aucun), few (un peu) et many (beaucoup).
topo : microtopography - facteur à modalités blanket (replat), hummock (butte).
density : numérique - Densité du substrat en $g.L^{-1}$ de matière sèche non comprimée
water : numérique - Contenu en eau du substrat en $g.L^{-1}$.

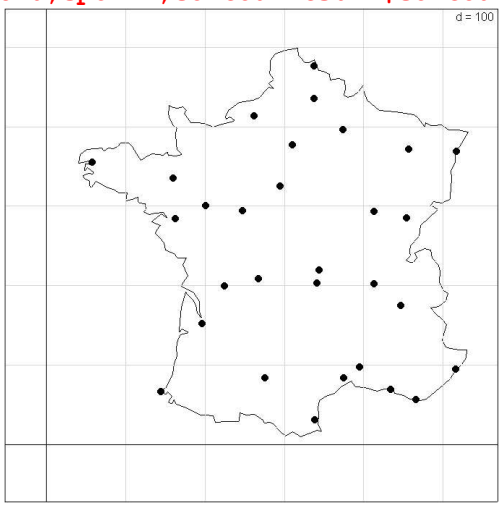
```
ori.xy=oribatid$xy[,c(2,1)] # pour avoir les figures dans la largeur de
la page
names(ori.xy)=c("x","y")

par(mfrow=c(9,4))
s.label(ori.xy,clab=0,incl=F,addax=F,cpoi=2)
for(j in 1:35)
      s.value(ori.xy,oribatid$fau[,j],cleg=0,,incl=F,addax=F,
            sub=names(oribatid$fau)[j],csub=2)
```



70 carottes et 35 espèces d'Oribates.

Les tableaux peuvent donc contenir des variables qualitatives, quantitatives ou distributionelles. Le tableau peut aussi être homogène : le tableau est homogène quand dans chaque cellule, à chaque ligne et chaque colonne, la mesure porte sur la même variable. Par exemple la température moyenne (°C x 10) dans 30 villes pour 12 mois (dans la liste t3012) forme un tableau homogène :

```
data(t3012)
s.label(t3012$xy, clab=0,cpoi=2,contour=t3012$contour)
```

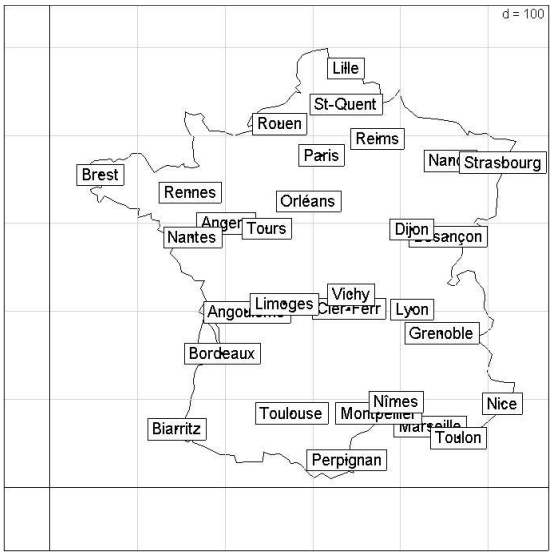


Pour ceux qui trouvent le cours indigeste : placer les 30 villes sur la carte (solution au dos).

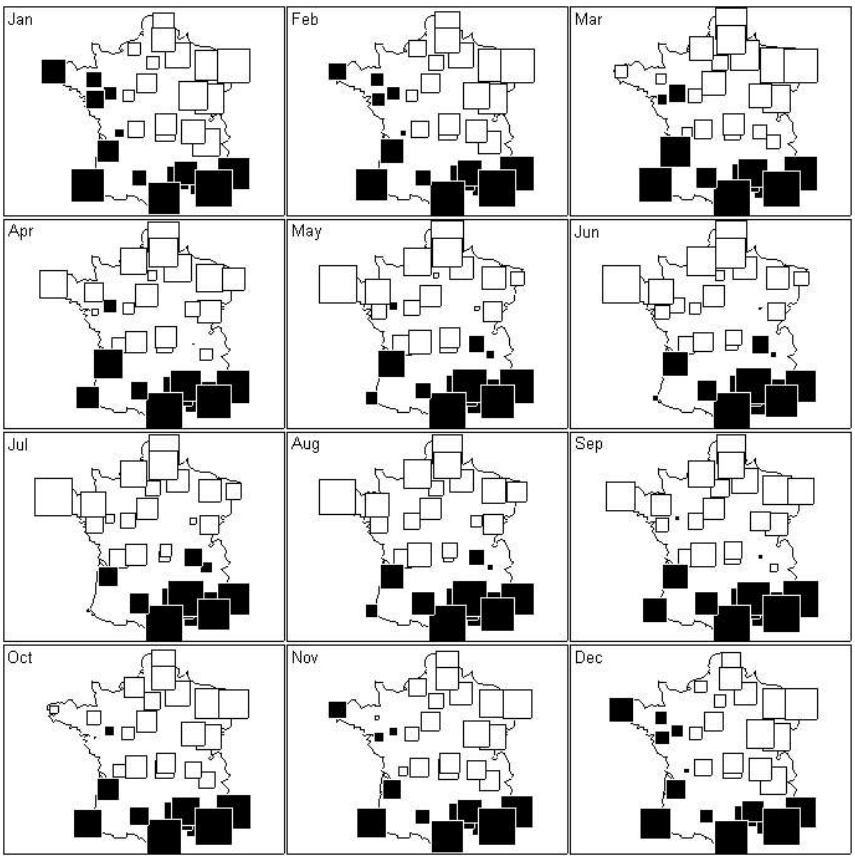
```
t3012$temp
```

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Angoulême	42	49	79	104	136	170	187	184	161	117	76	49
Angers	46	54	89	113	145	172	195	194	169	125	81	53
Besançon	11	22	64	97	136	169	187	183	155	104	57	20
Biarritz	76	80	108	120	147	178	197	199	185	148	109	82
Bordeaux	56	66	103	128	158	193	209	210	186	138	91	62
Brest	61	58	78	92	116	144	156	160	147	120	90	70
Cler-Ferr	26	37	75	103	138	173	194	191	162	112	66	36
Dijon	13	26	69	104	143	177	196	190	159	105	57	21
Grenoble	15	32	77	106	145	178	201	195	167	114	65	23
Lille	24	29	60	89	124	153	171	171	147	104	61	35
Limoges	31	39	74	99	133	168	184	178	153	107	67	38
Lyon	21	33	77	109	149	185	207	201	169	114	67	31
Marseille	55	66	100	130	168	208	233	228	199	150	102	69
Montpellier	56	67	99	128	162	201	227	223	193	146	100	65
Nancy	8	16	55	92	133	165	183	177	147	94	52	18
Nantes	50	53	84	108	139	172	188	186	164	122	82	55
Nice	75	85	108	133	167	201	227	225	203	160	115	82
Nîmes	57	68	101	130	166	208	236	229	197	146	98	65
Orléans	27	36	69	98	134	166	184	182	156	109	66	36
Paris	34	41	76	107	143	175	191	187	160	114	71	43
Perpignan	75	84	113	139	171	211	238	233	205	159	115	86
Reims	19	28	62	94	133	164	183	179	151	103	61	30
Rennes	48	53	79	101	131	162	179	178	157	116	78	54
Rouen	34	39	68	95	129	157	176	172	150	110	68	43
St-Quent	20	29	63	92	127	156	174	174	150	105	61	31
Strasbourg	4	15	56	98	140	172	190	183	151	95	49	13
Toulon	86	91	112	134	166	202	226	224	205	165	126	97
Toulouse	47	56	92	116	149	187	209	209	183	133	86	55
Tours	35	44	77	106	139	174	191	187	162	117	72	43
Vichy	24	34	71	99	136	171	193	188	160	110	66	34

```
s.label(t3012$xy, clab=0.75,cpoi=0,contour=t3012$contour)
```



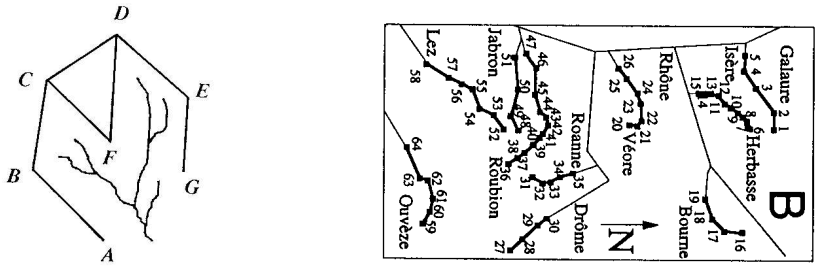

```
par(mfrow=c(4,3))
par(mar=c(0,0,0,0))
f1 <- function(z,sub){
    s.value(t3012$xy,z,addax=F,includ=F,cleg=0,csi=2,

contour=t3012$contour,cgrid=0,grid=F,sub=sub,csub=2)
}
for (j in 1:12)
      f1(scalewt(t3012$temp[,j]),names(t3012$temp)[j])
```



30 villes, 12 mois, la température moyenne normalisée par mois. Il fait plus chaud au Sud qu'au Nord ! Certes, mais encore ?

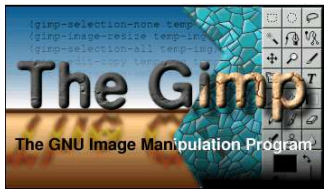
La liste contient en outre les coordonnées en xy et une image du contour de la France au format (x_1, y_1, x_2, y_2). Ces données ont servi d'illustrations dans la thèse de Ph. Besse (1979) et celle de S. Champely (1994).

Le dernier exemple est choisi dans une discipline ou le multivarié et le spatial font un couple particulièrement recherché, celui de la génétique. Multivarié est par essence l'enregistrement de la variabilité génétique (génotype d'un individu ou fréquences alléliques d'un groupe sur un ou plusieurs loci). Spatialisée est par essence la récolte des individus (de l'arbre en place à l'île dans l'océan, à toutes les échelles). L'espace est cependant celui du fonctionnement biologique. Voilà comment Smouse et Peakall (1999) voient le voisinage entre arbres d'une espèce dont la reproduction demande l'intervention d'un petit rongeur qui ne traverse jamais un cours d'eau (p. 566, à gauche) :



ce qui est l'exact complément du voisinage entre stations dans un réseau hydrographique (Thioulouse et al. 1995 p. 2, à droite). On a un exemple saisissant de cette plasticité de la notion spatiale, qui dépasse largement la notion de coordonnées ou même celle de distances, dans le travail de Fievet et al. (2001). L'analyse porte sur une crevette *Atya innocous* qui vit

et se reproduit en eau douce mais dont les larves dévalent et ont une période de croissance en mer. Les stations sont situées sur les rivière de Basse-Terre. La carte est scannée en mode binaire, et le format pnm est obtenu à partir d'un .tif avec Gimp :



http://www.gimp.org/

Utiliser la librairie **pixmap** de Friedrich Leisch and Roger Bivand :

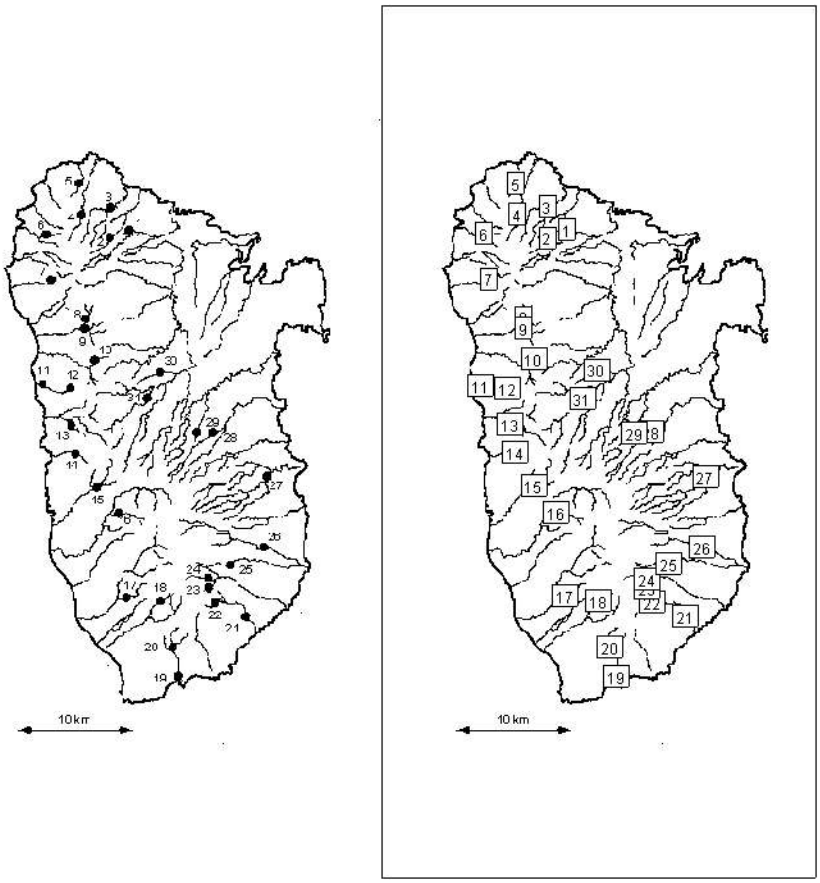
```
data(atya)
atya.digi <- read.pnm(system.file("pictures/atyadigi.pnm",
      package = "ade4"))
atya.carto <- read.pnm(system.file("pictures/atyacarto.pnm",
      package = "ade4"))
```

> Remarque : Si la librairie ade4 n'est pas dans le dossier standard de R, placer les fichiers .pnm dans le dossier de travail et les charger directement.

```
plot(atya.digi)
```

Pour digitaliser la position des stations (le résultat est déjà dans **atya$xy**) :

```
btxy=data.frame(locator(31))
par(mfrow = c(1,2))
plot(atya.digi)
plot(atya.carto)
s.label(btxy,add.plot=T,clab=0.75)
```



On dira ici que deux stations sont proches si elles sont dans le même bassin versant ou si elles sont dans deux bassins versants voisins au sens de la distance à parcourir le long de la côte entre les embouchures. L'hypothèse est que le mode de fonctionnement :

– soit génère le fonctionnement en une seule population avec brassage complet au cours de la migration (il n'y a pas de structure spatiale),

– soit induit une structure spatiale dans la composante génétique avec une ressemblance plus forte entre stations plus proches.

Pour en juger les données sont :

```
names(atya$gen)
 [1] "Fbp100"    "Fbp155"    "Gpi52"     "Gpi72"     "Gpi100"    "Gpi126"
 [7] "Gpi143"    "Mdh1.15"   "Mdh1.100"  "Mdh2.50"   "Mdh2.80"   "Mdh2.100"
[13] "Mdh2.115"  "Mdhp60"    "Mdhp84"    "Mdhp100"   "Mdhp140"   "Pgm88"
[19] "Pgm91"     "Pgm98"     "Pgm100"    "Pgm180"
```

On a 6 loci avec respectivement 2, 5, 2, 4, 4 et 5 allèles et des fréquences alléliques :

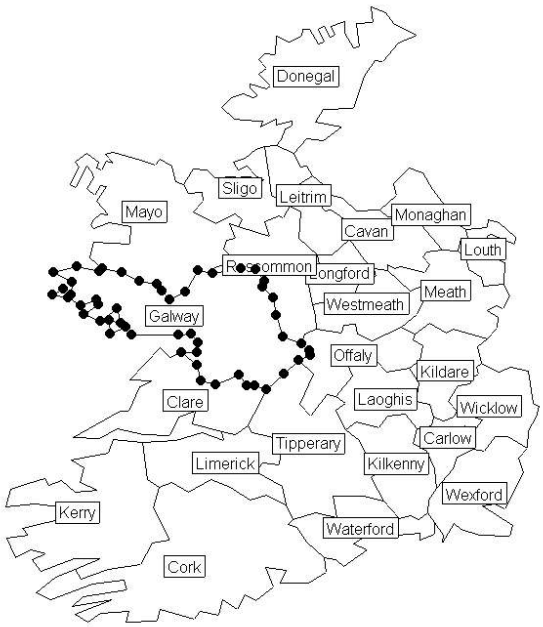
```
atya$gen[,3:7]
   Gpi52 Gpi72 Gpi100 Gpi126 Gpi143
1   0.00  0.02   0.98   0.00   0.00
2   0.00  0.02   0.96   0.02   0.00
. . .
19  0.00  0.02   0.93   0.05   0.00
20  0.00  0.05   0.95   0.00   0.00
. . .
30  0.00  0.00   1.00   0.00   0.00
31  0.00  0.00   0.95   0.03   0.02
```

2. Relations de voisinages

Dans tout ce qui suit, nous sommes donc concernés par un tableau de données et une structure spatiale au sens le plus large. En hydrobiologie terrestre, les unités statistiques sont des tronçons de rivière et la distance euclidienne n'a pratiquement aucun sens pour mesurer des proximités spatiales. Dans tous les cas, par contre on peut introduire l'espace en quantifiant, comme on le désire, le voisinage. La statistique spatiale s'appuie sur la quantification du voisinage et la structure spatiale s'exprime comme une relation quantitative, mesure sur chaque couple de deux points de l'importance du premier élément pour le second. Sont voisines, dans le cas le plus simple, deux unités surfaciques ayant une frontière commune et sont d'autant plus voisines que cette frontière est plus longue.

Par exemple, la matrice **`irishdata$link.utm`** `(25 lignes et 25 colonnes)` contient la longueur de frontières communes à deux unités. Les unités sont arbitraires (dans **`irishdata$link`**, `on a` des pixels issus de la digitalisation à l'écran). On note traditionnellement **W** une telle matrice de terme général w_{ij}. On dira qu'il s'agit d'une **matrice de voisinage** brute.

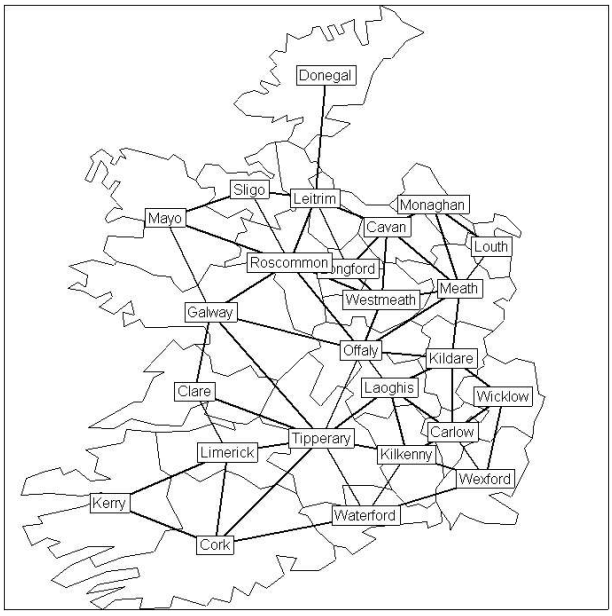
```
area.plot(irishdata$area.utm, lab=row.names(irishdata$xy.utm), clab = 1)
points(irishdata$area.utm[irishdata$area.utm[,1]=="Galway",2:3],
       cex=2,pch=20)
```




```
irishdata$link.utm[6,]
```

Carlow	Cavan	**Clare**	Cork	Donegal	Galway	Kerry	Kildare
0.00	0.00	**71.00**	0.00	0.00	0.00	0.00	0.00
Kilkenny	Laoghis	Leitrim	Limerick	Longford	Louth	**Mayo**	Meath
0.00	0.00	0.00	0.00	0.00	0.00	**85.13**	0.00
Monaghan	**Offaly**	Roscommon	Sligo	**Tipperary**	Waterford	Westmeath	Wexford
0.00	**23.95**	90.39	0.00	**37.85**	0.00	0.00	0.00
Wicklow							
0.00							

La première manière de s'en servir est de la réduire à deux valeurs "0 ou 1" pour "positif ou nul" comme pour "voisin ou non voisin". Quand on la réduit en binaire (1 deux unités sont voisines, 0 elles ne le sont pas) on obtient un **graphe de voisinage**. On notera **M** une telle matrice.



Un graphe de voisinage est donc une matrice carrée, symétrique, à diagonale nulle et à valeurs 0 (non voisins) ou 1 (voisins).

```
ir.xy=irishdata$xy.utm
ir.contour = irishdata$contour.utm
ir.w = irishdata$link.utm
ir.area = irishdata$area.utm
ir.01 = 1*(ir.w>0)
ir.neig = neig(mat01=ir.01)
ir.neig
```

```
m.irish
Carlow       .
Cavan        ..
Clare        ...
Cork         ....
Donegal      .....
Galway       ..1...
Kerry        ...1...
Kildare      1.......
Kilkenny     1........
Laoghis      1......11.
Leitrim      .1..1.....
Limerick     ..11..1.....
Longford     .1........1..
Louth        ..............
Mayo         .....1.........
Meath        .1.....1.....1..
Monaghan     .1...........1.1.
Offaly       .....1.1.1.....1..
Roscommon    .....1....1.1.1..1.
Sligo        ..........1...1...1.
Tipperary    ..11.1..11.1.....1...
Waterford    ...1....1...........1.
Westmeath    .1..........1..1.11....
Wexford      1.......1............1..
Wicklow      1......1..............1.
```

```
s.label(ir.xy,contour=ir.contour,inclu=F,neig=ir.neig,
            area=ir.area,ylim=range(ir.area[,3]),addax=F,cgrid=0,grid=F)
```

En fait, l'objet graphe de voisinage est conservé dans **ade4** comme une liste d'arêtes :

```
unclass(ir.neig)
```

```
      [,1] [,2]
 [1,]    3    6
 [2,]    4    7
 [3,]    1    8
. . .
[53,]    8   25
[54,]   24   25
attr(,"degrees")
   Carlow     Cavan     Clare      Cork   Donegal    Galway     Kerry
        5         5         3         4         1         5         2
. . .
Waterford Westmeath   Wexford   Wicklow
        4         5         4         3
attr(,"call")
neig(mat01 = m.irish)
```

Mais la vraie librairie des graphes de voisinages dans R est la librairie **spdep** de Roger Bivand . Dans **spdep** les graphes de voisinages sont conservés comme liste de voisins :

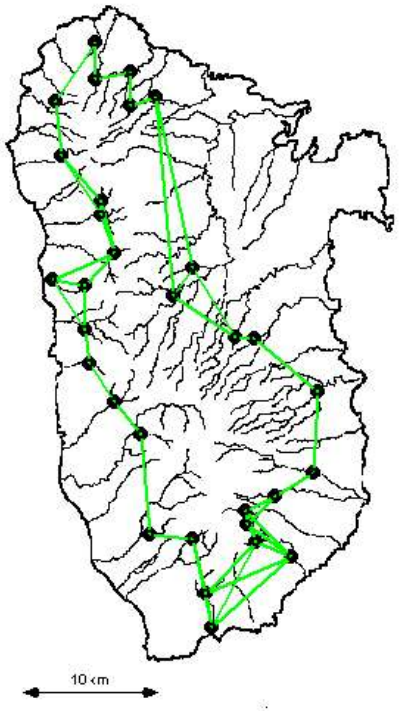
```
ir.nb = neig2nb(ir.neig)
ir.nb
```

```
$"1"
[1]  8  9 10 24 25

$"2"
[1] 11 13 16 17 23
```

```
. . .

$"24"
[1]  1  9 22 25

$"25"
[1]  1  8 24

attr(,"region.id")
 [1] "Carlow"    "Cavan"     "Clare"     "Cork"      "Donegal"   "Galway"
 [7] "Kerry"     "Kildare"   "Kilkenny"  "Laoghis"   "Leitrim"   "Limerick"
[13] "Longford"  "Louth"     "Mayo"      "Meath"     "Monaghan"  "Offaly"
[19] "Roscommon" "Sligo"     "Tipperary" "Waterford" "Westmeath" "Wexford"
[25] "Wicklow"
attr(,"gal")
[1] FALSE
attr(,"call")
neig2nb(neig = ir.neig)
attr(,"class")
[1] "nb"
```

La fonction **neig** dans **ade4** permet de récupérer les graphes de voisinages éventuellement implantés dans la version antérieure du logiciel **ADE-4** et **neig2nb** les transporte dans **spdep**. On peut donc faire les graphes de voisinage manuellement dans les cas originaux. Par exemple, utiliser la carte des stations de Basse-Terre, faire la liste des couples de voisins dans un fichiers **atyagraph.txt** :

```
1      2
1      30
1      31
2      3
...
26     27
27     28
28     29
29     30
29     31
30     31

w = as.matrix(read.table("atyagraph.txt", h=F))
t(w)
   1  2  3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28
V1 1  1  1 2 3 4 5 6 7  7  8  8  9 10 10 11 11 12 13 14 15 16 17 18 18 19 19 19
V2 2 30 31 3 4 5 6 7 8  9  9 10 10 11 12 12 13 13 14 15 16 17 18 19 20 20 21 22
   29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45
V1 20 20 21 21 21 22 22 23 23 24 25 26 27 28 29 29 30
V2 21 22 22 23 24 23 24 24 25 25 26 27 28 29 30 31 31

at.neig = neig(edges=w)
at.nb = neig2nb(at.neig)

Le résultat est disponible dans atya :

at.nb = neig2nb(atya$neig)
plot(atya.carto)
points(atya$xy,clab=0.75)
plot(at.nb, atya$xy, col="green", add=T, lwd=2)
```

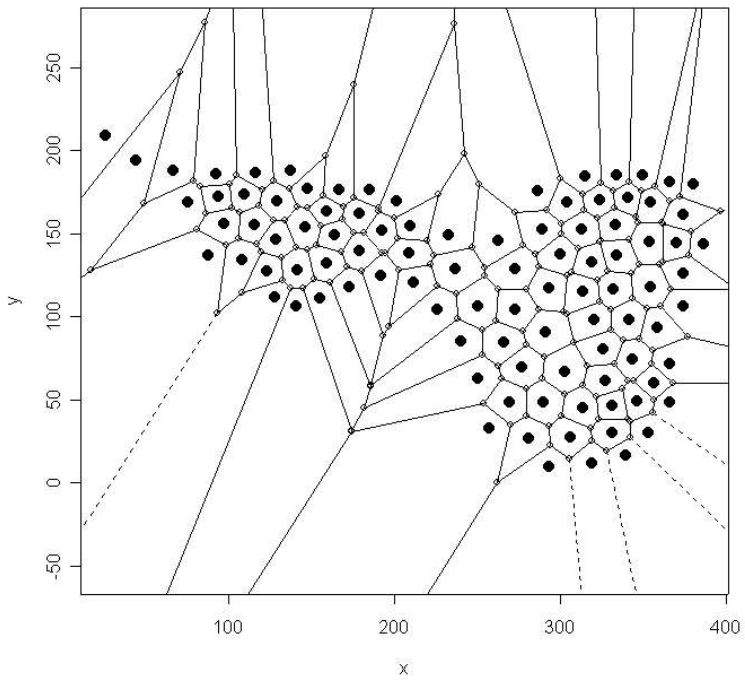


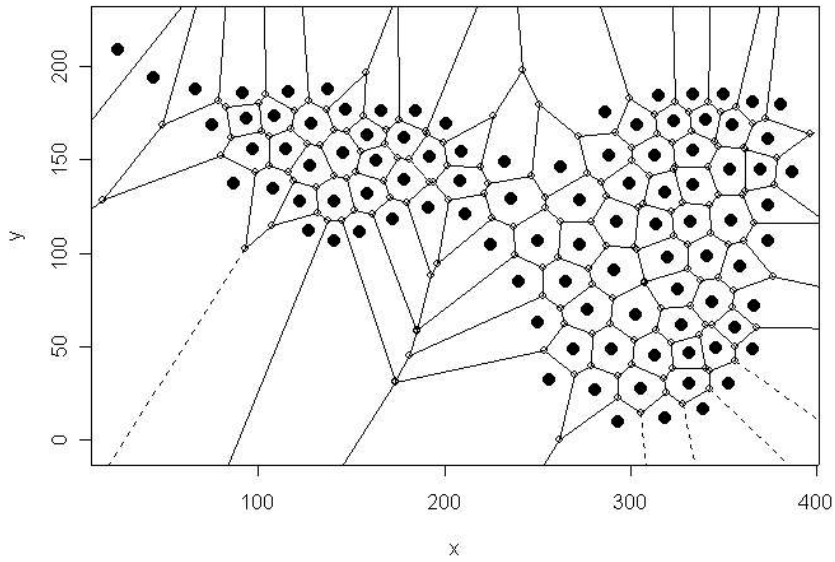
Graphe de voisinage de type circulaire

A noter : l'opération précédente génère un graphe symétrique et il suffit d'enregistrer le couple (x,y) ou le couple (y,x).

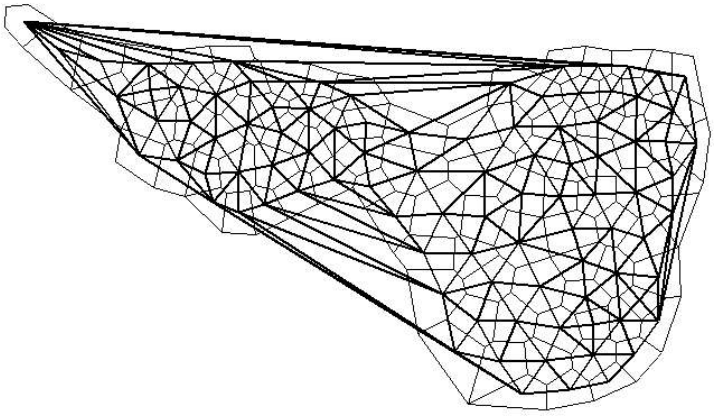
Le graphe de Voronoi (description dans Upton and Fingleton 1985) engendre des graphes de voisinages à partir des coordonnées. Utiliser la librairie **`tripack`** (code Fortran de R. J. Renka. Fonctions R de A. Gebhardt et contributions de S. Eglen et S. Zuyev) :

```
plot(mafragh$xy,asp=1,pch=20,cex=2)
plot(voronoi.mosaic(mafragh$xy),add=T)
```

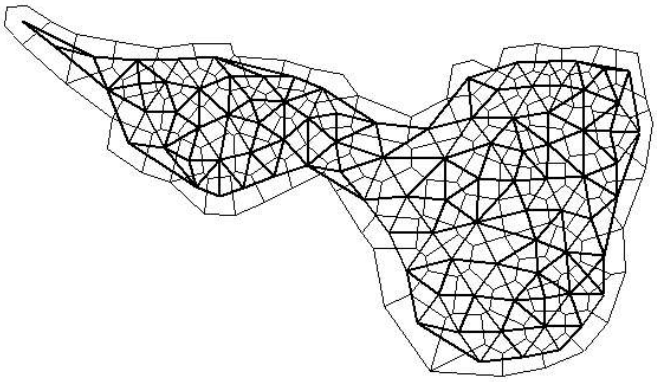





```
maf1 = tri2nb(mafragh$xy)
area.plot(mafragh$area,graph=nb2neig(maf1))
```



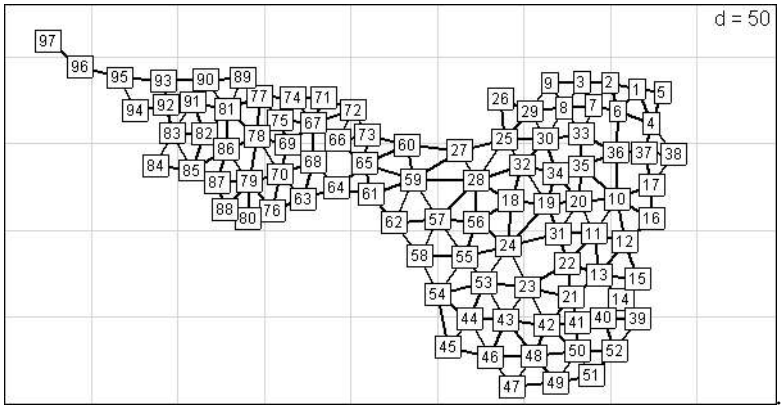
```
area.plot(mafragh$area,graph=mafragh$neig) # après manipulation
```



Graphe de voisinage dérivé d'une tesselation avec ajustement manuel (librairie tripack de R. J. Renka et Albrecht Gebhardt). Le même procédé est utilisé en milieu marin (Ghertsos et al. 2001)

Pour les unité ponctuelles, **spdep** permet de créer des graphes de voisinages variés.

```
w=gabrielneigh(as.matrix(mafragh$xy))
s.label(mafragh$xy,neig=nb2neig(graph2nb(w)),clab=0.75)
```



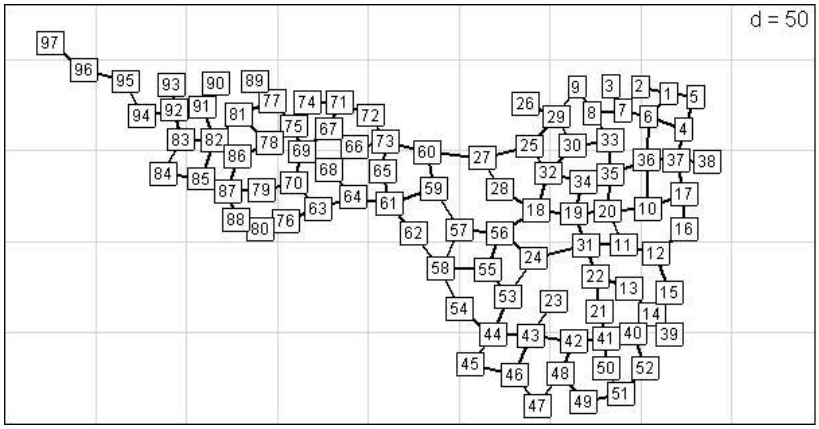
Le graphe de Gabriel est un sous graphe du graphe de Voronoi. Il est défini par x et y sont voisins si ils le sont au sens de la triangulation de Delaunay et si :

$$d(x,y) \leq \min_z \left(\sqrt{d^2(x,z) + d^2(y,z)} \right)$$

Deux points sont connectés si aucun autre point ne se trouve a l'intérieur du cercle de diamètre défini par ces 2 points (Gabriel and Sokal 1969).

```
library(mva)
ww=as.matrix(dist(mafragh$xy))
ww[9,26]
[1] 29.79
ww[29,26]
[1] 18.65
ww[29,9]
[1] 19.69
sqrt(ww[29,26]^2+ ww[29,9]^2)
[1] 27.12
9 et 26 ne sont pas voisins parce que le minimum réalisé par 29 est plus petit.
```

Définition, références dans la documentation de **gabrielneigh** dans **tripack**. **gabrielneigh** donne des objets de la classe **graph** (liste de couples de voisins et coordonnées), **graph2nb** transforme les objets de la classe **graph** en objet de la classe **nb**. Le graphe des voisins relatifs est aussi disponible :



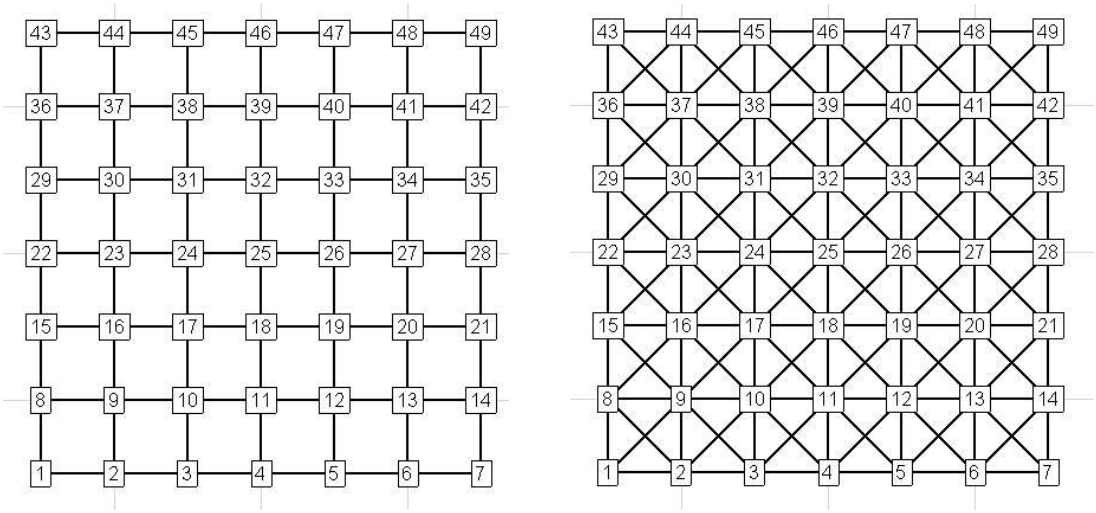
Le schéma de voisinage relatif

Dans ce graphe x et y sont voisins si ils le sont au sens de la triangulation de Delaunay et si $d(x,y) \leq \min_z \left(\max\left(d(x,z),d(y,z)\right)\right)$. C'est un sous-graphe du précédent. Il est ici trop faible pour l'objectif.

```
w=relativeneigh(as.matrix(mafragh$xy))
s.label(mafragh$xy,neig=nb2neig(graph2nb(w)),clab=0.75)
```

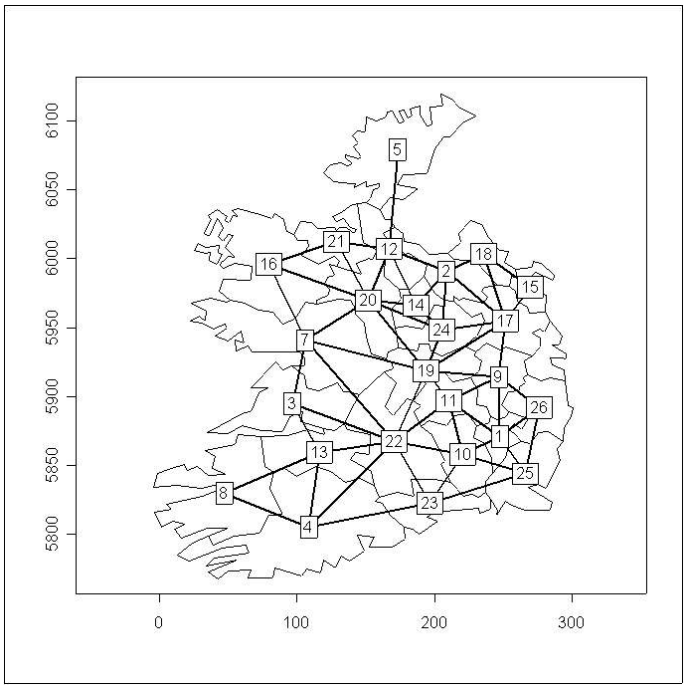
Mais pour travailler sur des grilles :

```
s.label(expand.grid(1:7,1:7),neig=nb2neig(cell2nb(7,7,type="rook")))
s.label(expand.grid(1:7,1:7),neig=nb2neig(cell2nb(7,7,type="queen")))
```



Relation de voisinage de la tour et de la reine sur une grille

```
data(eire) # dans spdep
plotpolys(eire.polys.utm, eire.bbs.utm)
s.label(eire.coords.utm[-6,],add.plot=T,neig=ir.neig)
```



Tous les calculs sur graphes de voisinage utiliseront les structures de données de **spdep**. Passer les graphes de voisinage (**neig** de **ade4**) aux graphes de voisinage (**nb** de **spdep**)

par **neig2nb**. Passer les fichiers **area** (**ade4**) aux listes de polygones (**spdep**) par **area2poly**.

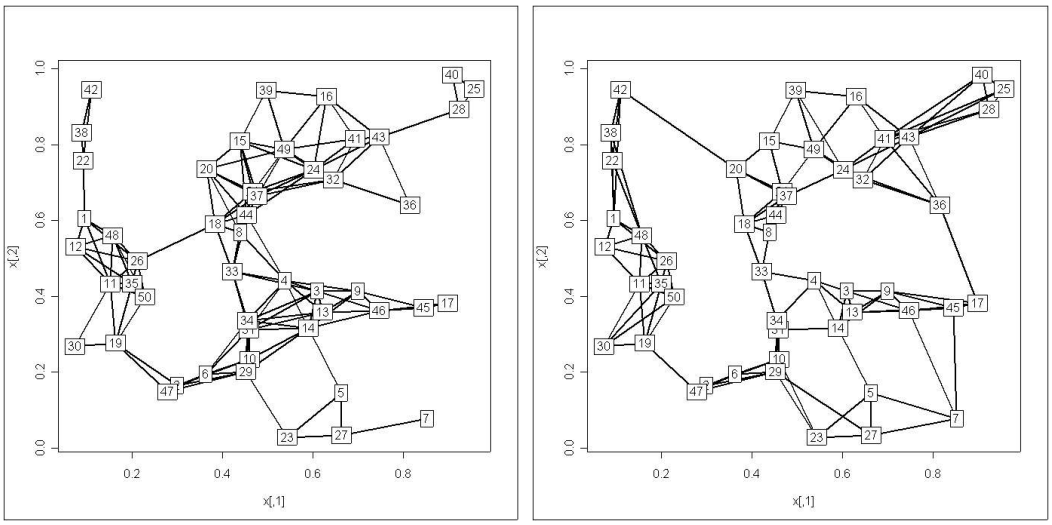
Les représentations graphiques sont équivalentes dans les deux librairies, mais les concepts de poids de voisinage sont en œuvre dans **spdep**.

On pourra aussi utiliser le **voisinage par distance**. Deux points sont voisins si et seulement si leur distance est supérieure à d1 et inférieure à d2 :

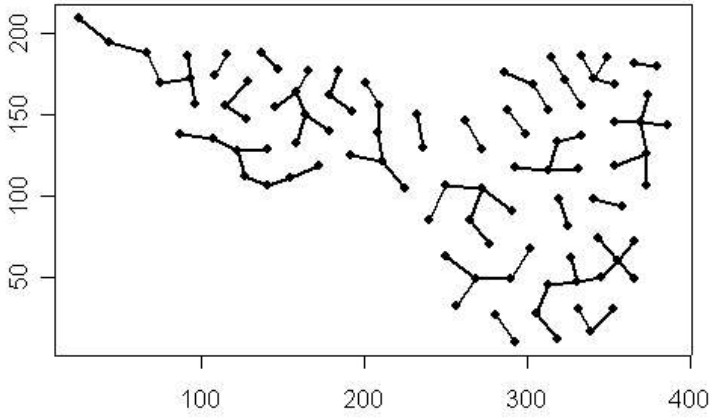
```
x = cbind(runif(50),runif(50))
plot(x)
w = dnearneigh(x,0,0.2)
s.label(as.data.frame(x),add.p=T,neig=nb2neig(w))
```

ou encore le **voisinage par plus proches voisins** :

```
w =knn2nb(knearneigh(x,4))
plot(x)
s.label(as.data.frame(x),add.p=T,neig=nb2neig(w))
```



à gauche : graphe du voisinage "deux points sont voisins si leur distance est inférieure à 0.2". à droite : graphe du voisinage i est voisin de j si il est un des 4 plus proches voisins de j. Cette matrice de voisinage n'est pas symétrique par défaut.



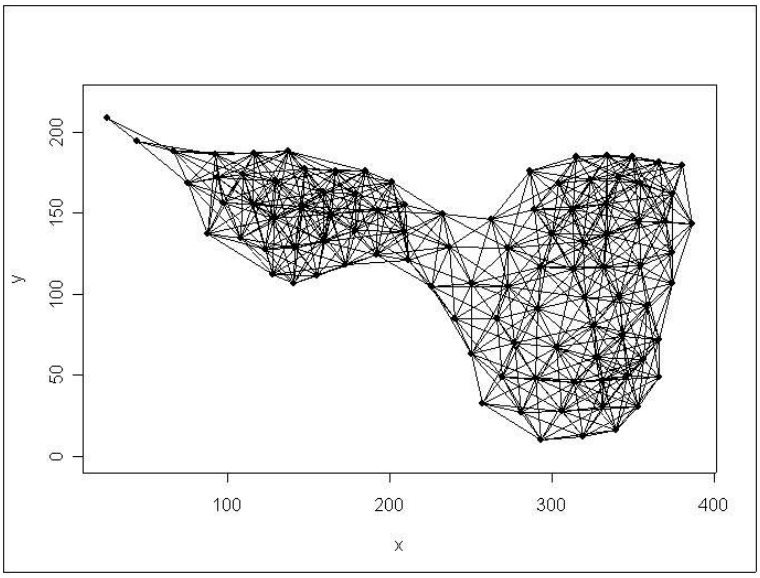
Un point a un seul plus proche voisin mais peut être le plus proche voisin de plusieurs autres.

```
plot(mafragh$xy,asp=1)
```

```
s.label(mafragh$xy,add.p=T,neig=nb2neig(knn2nb(
            knearneigh(as.matrix(mafragh$xy),1))),clab=0)
```

Le voisinage par les plus proches voisins conduit à un nombre constant de voisins, ce qui est un avantage (pondération uniforme par points) qu'on paye immédiatement par la non symétrie. La relation du plus proche voisin (ci-dessus), peut-être la partie commune à toutes les autres, a des propriétés fascinantes (Pace and Zou 2000) mais n'est pas utile ici.

```
plot(mafragh$xy,asp=1)
s.label(mafragh$xy,add.p=T,
      neig=nb2neig(dnearneigh(as.matrix(mafragh$xy),0,50)),
      clab=0,cneig=1)
```



Graphe de voisinage par distances : deux points sont voisins s'ils sont distants de 50 unités au plus (xy en pixels).

Le matériel de base des ordinations sous contraintes spatiales est le graphe de voisinage. En toute généralité, les graphes de la classe **nb** sont orientés : *i* peut être voisin de *j* sans que *j* soit voisin de *i*. C'est vrai pour deux stations sur un cours d'eau : l'amont influence l'aval et non l'inverse. La matrice du graphe **M** est définie par $m_{ij} = 1 \Leftrightarrow j$ est un voisin de i. La plupart des auteurs de l'école de Lebart n'envisagent que des graphes symétriques et ne manipulent pas de pondération de voisinage autre que directement induit par la relation binaire. Pour beaucoup d'autres, et on va voir pourquoi, le second élément de la prise en compte de l'espace est la pondération de voisinage. Un même graphe de la classe **nb** peut donner plusieurs pondérations de la classe **listw**.

On est alors dans la tradition des géographes et des économètres. Les écologues sont au courant, les généticiens – dans leur tour d'ivoire – reconstruisent le monde.

3. Pondérations de voisinage

Une pondération de voisinage est toujours associée à un graphe de voisinage. Ce qui est pondéré c'est le lien entre voisins. R. Bivand a représenté les principales options dans ses procédures. Pondérer un voisinage est essentiellement une question pratique qui fournit une matrice **W** à n lignes et n colonnes telles que $w_{ij} \geq 0$ et $w_{ij} = 0 \Leftrightarrow m_{ij} = 0$. Dans un objet de la classe **listw** on a d'abord une liste à n composantes qui sont des vecteurs donnant les numéros des voisins (on peut ou non tolérer des points sans voisins) puis une liste à n composantes qui sont des vecteurs donnant les poids des voisins.

Une remarque est très importante : la librairie de R. Bivand ne contient jamais de matrices et aucune des fonctions présentes ne manipule des matrices de voisinages (qui contiennent énormément de valeurs nulles). Ces fonctions n'ont donc pratiquement pas de limites en nombre de points, car elles n'utilisent que des listes de voisins et des listes de poids de voisinage. Les notations matricielles sont donc ici purement conceptuelles. Il y a au moins deux manières principales de pondérer pratiquement les voisinages. Le plus simple est de laisser agir la fonction **nb2listw**. Ces pratiques sont présentes dans l'ouvrage fondateur de Cliff et Ord (1973).

```
is.matrix(ir.neig) # ir.neig est un graphe de la classe neig
```

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

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

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

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

$call
neig(mat01 = ir01)

$class
[1] "neig"
```

```
is.list(ir.nb) # ir.nb est un objet de la classe nb
```

```
[1] TRUE

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

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

attributes(ir.nb)
$names
 [1] "1"  "2"  "3"  "4"  "5"  "6"  "7"  "8"  "9"  "10" "11" "12" "13" "14" "15"
```

```
[16] "16" "17" "18" "19" "20" "21" "22" "23" "24" "25"

$region.id
 [1] "Carlow"    "Cavan"     "Clare"     "Cork"      "Donegal"   "Galway"
. . .
[19] "Roscommon" "Sligo"     "Tipperary" "Waterford" "Westmeath" "Wexford"
[25] "Wicklow"

$gal
[1] FALSE

$call
neig2nb(neig = ir.neig)

$class
[1] "nb"
```

Les deux objets contiennent la même information dans des formats différents :

```
ir.neig[which(ir.neig[,1]==1),]
```

```
     [,1] [,2]
[1,]    1    8
[2,]    1    9
[3,]    1   10
[4,]    1   24
[5,]    1   25

ir.nb[[1]]
[1]  8  9 10 24 25
```

Les deux objets contiennent la même information mais seul le second fournit des pondérations de voisinages de la classe **listw** :

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

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

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

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

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

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

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

L'option W donne une *row standardized spatial weights matrix* comme dans Cliff and Ord (1973 p. 13) ou Anselin and Hudak (1992 p. 514) .

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

```
[1] 0.2212
```

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

Pour éviter les complications nous dirons que cette matrice est encore de type *n***F**.

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

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

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

```
sumlig=apply(ir.w,1,sum)
ir.w[1,]/sumlig[1]
   Carlow     Cavan     Clare      Cork   Donegal    Galway     Kerry
   0.0000    0.0000    0.0000    0.0000    0.0000    0.0000    0.0000
  Kildare  Kilkenny   Laoghis   Leitrim  Limerick  Longford     Louth
   0.1031    0.2528    0.1008    0.0000    0.0000    0.0000    0.0000
     Mayo     Meath  Monaghan    Offaly Roscommon     Sligo Tipperary
   0.0000    0.0000    0.0000    0.0000    0.0000    0.0000    0.0000
Waterford Westmeath   Wexford   Wicklow
   0.0000    0.0000    0.2445    0.2988

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

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

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

B ***basic binary coding*** : l'option B attribue à chaque couple la valeur brute qui lui revient :

```
pond.ext.b$weights[[1]]
 Kildare Kilkenny  Laoghis  Wexford  Wicklow
   19.87    48.75    19.44    47.13    57.62
 ir.w[1,]
   Carlow     Cavan     Clare      Cork   Donegal    Galway     Kerry
     0.00      0.00      0.00      0.00      0.00      0.00      0.00
  Kildare  Kilkenny   Laoghis   Leitrim  Limerick  Longford     Louth
    19.87     48.75     19.44      0.00      0.00      0.00      0.00
     Mayo     Meath  Monaghan    Offaly Roscommon     Sligo Tipperary
     0.00      0.00      0.00      0.00      0.00      0.00      0.00
Waterford Westmeath   Wexford   Wicklow
     0.00      0.00     47.13     57.62
```

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

```
25*ir.w[1,]/sum(ir.w)
   Carlow     Cavan     Clare      Cork   Donegal    Galway     Kerry
  0.00000   0.00000   0.00000   0.00000   0.00000   0.00000   0.00000
  Kildare  Kilkenny   Laoghis   Leitrim  Limerick  Longford     Louth
```

```
  0.09632   0.23629   0.09422   0.00000   0.00000   0.00000   0.00000
     Mayo     Meath  Monaghan    Offaly Roscommon     Sligo Tipperary
  0.00000   0.00000   0.00000   0.00000   0.00000   0.00000   0.00000
Waterford Westmeath   Wexford   Wicklow
  0.00000   0.00000   0.22846   0.27928
```

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

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

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

C'est typiquement *n***F.**

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

```
ir.w[1,]/sum(ir.w)
   Carlow     Cavan     Clare      Cork   Donegal    Galway     Kerry
 0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000
  Kildare  Kilkenny   Laoghis   Leitrim  Limerick  Longford     Louth
 0.003853  0.009452  0.003769  0.000000  0.000000  0.000000  0.000000
     Mayo     Meath  Monaghan    Offaly Roscommon     Sligo Tipperary
 0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000
Waterford Westmeath   Wexford   Wicklow
 0.000000  0.000000  0.009138  0.011171
pond.ext.u$weights[[1]]
 Kildare Kilkenny  Laoghis  Wexford  Wicklow
0.003853 0.009452 0.003769 0.009138 0.011171
```

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

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

W normalisée par ligne : matrice de type L

U globalement normalisée : matrice de type F

$$\mathbf{L} = \left[f_{j/i} \right] \Rightarrow \mathbf{L}\mathbf{1}_n = 1 \quad \mathbf{F} = \left[f_{ij} \right] \Rightarrow \mathbf{1}_n^t \mathbf{F} \mathbf{1}_n = 1$$

La question des poids de voisinage est le talon d'Achille des méthodes multivariées intégrant l'espace. On utilise en général le terme générique **W** pour parler de **F**, *n***F, M, L** ou *n***L**, ce qui ne simplifie pas les choses. C'est tellement instable qu'il faut vérifier systématiquement ce qui s'est fait. Dans Cliff et Ord (*op. cit.* p. 13) les auteurs proposent un système quelconque où le poids est une fonction des longueurs des frontières communes et des distances entre centres, ou encore l'inverse de la distance entre centres, mais aussi une matrice **W** de somme unité par lignes non symétrique pour faire en sorte que **Wz** calcule les moyennes des voisins d'une unité statistique. Cet aspect sans pratique canonique, ou pour le

moins à deux options dominantes, ouvre la porte à toutes manipulations arbitraires qu'on retrouve dans l'usage de l'indice de Moran dans l'étude de la relation entre un trait biologique et une phylogénie (Gittleman and Kot 1990).

Deux pratiques cependant se retrouve constamment chez les principaux auteurs, la normalisation par ligne et la double normalisation. Les termes sont bizarres mais d'usage général.

Pour les praticiens de l'autocorrélation spatiale et de l'école fondée par Moran (cf. ci-dessous) **la normalisation par lignes** est fréquente. Le terme désigne simplement la division de chaque valeur par la somme de la ligne où on la trouve. On retrouve **L** la matrice normalisée par lignes, L pour profil par lignes, de terme général :

$$l_{ij} = \frac{w_{ij}}{\sum_j w_{ij}} = \frac{w_{ij}}{w_{i\bullet}}$$

Pour le voisinage par distance maximum la pondération de voisinage définie par Pace et Barry (1997) (eq. 6) est :

$$d_{ij} \leq d_{\max} \Rightarrow w_{ij} = 1$$

où $d_{\max}$ est la distance maximum d'influence fixée. Comme indiqué par les auteurs cette pondération est ensuite normalisée par lignes et donc du type **L**.

La relation de voisinage des *m* plus proches voisins définie par Pace et al (2003) (eq. 12 p. 14) s'écrit :

$$0 < d_{ij} < d_i^m \Rightarrow w_{ij} = \frac{1}{m}$$

où d_i^m est la distance de *i* à son *m*-ième plus proche voisin. Comme indiqué par les auteurs cette pondération de voisinage est non symétrique mais *row-stochastic*, c'est-à-dire de somme unité par lignes, et donc du type **L**.

Pour Bavaud (1998) la définition d'une pondération de voisinage est très précise. Soit $S = \{1,...n\}$ un ensemble de points. Une matrice de pondération de voisinage est une matrice **W** à *n* lignes et *n* colonnes telle que a) $w_{jk} \geq 0$ b) $\sum_{k=1}^{n} w_{jk} = 1 \quad \forall j \in S$. Les poids w_{jj} ne sont pas forcément nuls. On utilise les termes équivalents de ***contiguity***, ***connectivity***, ***adjacency***, ***association*** pour ce type de matrices. La matrice **W** n'est pas forcément symétrique car elle donne une indication sur l'influence potentielle de *i* sur *j*. Il s'agit alors de la matrice de transition d'un processus de Markov. Elle est de type **L**.

Certes, une proposition remarquable est faite encore par Pace et LeSage (2002) pour combiner les relations au *k* premiers voisins pour obtenir une matrice de poids bistochastiques dites *the doubly spatial model* (distributions de fréquences par lignes et par colonnes), ce qui est très alléchant, mais pas encore en routine. Les matrices de type **L**, stochastiques par ligne, ou markoviennes, décrivent la répartition de l'influence du point *i* sur l'ensemble des autres par une distribution de fréquence. C'est l'option par défaut dans **`spdep`** et ce n'est sûrement pas un hasard.

Pour les praticiens de la variance locale et de l'école fondée par Geary (cf. ci-dessous) et relancée en multivarié par Lebart, c'est au contraire les matrices de type **F** qui s'impose. La double normalisation est simplement la division par la somme de toutes ses valeurs qui

donne une **matrice de poids de voisinages**. On retrouve **F** la matrice globalement normalisée, F pour tableau de fréquences. Son terme général :

$$f_{ij} = \frac{w_{ij}}{\sum_{i,j} w_{ij}}$$

contient le poids du voisinage entre les unités *i* et *j*. Le poids de voisinage n'a guère de sens pour le couple (i,i) et pour éviter les sommes pour $i \neq j$ on simplifie en posant $w_{ii} = 0$. Une matrice de poids de voisinage est donc une matrice carrée, symétrique, à diagonale nulle et somme unité. A partir de maintenant nous utiliserons **L** ou **F** pour désigner des pondérations de voisinages et **W** quand les deux cas sont concernés. L'usage des matrices **F** en autocorrélation est aussi largement répandu. P. Aubry (2000) qui fait une analyse bibliographique hors du commun utilise directement sans notions de voisinages les pondérations (p. 58) :

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

La plus simple des matrices de poids de voisinage se rapporte à un *espace inconcevable*, celui dans lequel tout point serait également voisin de tout autre, celui où tout couple de points aurait un poids égal à celui de tout autre, c'est-à-dire le poids $1/n(n-1)$ car il y a $n(n-1)$ couples. La remarque n'est pas innocente.

La question des poids de voisinage n'est donc pas fixée et ne le sera sans doute pas, une solution universelle pour tous les problèmes et tous les matériaux n'ayant pas de sens.

On se pose la question de l'analyse en composantes principales de ce type de données en introduisant la notion de voisinages comme contrainte. L'intérêt est d'aborder des tableaux massivement multivariés comme le sont par exemple des relevés de faune ou de flore. Ces méthodes s'appuient sur les éléments univariés de base que sont les indices de structures spatiales.

4. Moran et Geary : la fondation de deux écoles

Dans les problèmes largement multivariés, on a d'abord besoin de trier les variables en fonction de leur *pattern* ou mode de variation dans l'espace, pour le moins sur l'existence d'une information spatialisée. Mais plus profondément une méthode multivariée a pour fonction essentielle de réduire le nombre de variables en faisant des combinaisons linéaires qui optimise ce qu'on sait faire en univarié. La régression multiple donne la combinaison qui offre la meilleure régression simple, l'analyse discriminante donne la combinaison qui offre la meilleure analyse de variance, l'analyse canonique donne la combinaison des variables du premier tableau et celle du second tableau qui optimise la corrélation. En multivarié spatial, il en sera de même. La seule nouveauté c'est qu'il y a deux critères en compétition depuis 50 ans. Il y aura donc deux écoles (au moins) de statistiques multivariées spatialisées.

4.1. Mesure d'auto-corrélation

Les indices de Geary (1954) et de Moran (1948, 1950) sont à la base de deux écoles de statistiques spatiales. On utilise directement la présentation de Cliff et Ord (*op. cit.* p. 8). Un résumé efficace par L. Anselin est disponible à :

http://geog55.geog.uiuc.edu/sa/pdf/w9_global.pdf

n est le nombre de mesures (unités statistiques) et **W** est la matrice des poids de voisinages. x_i est la valeur de l'unité statistique *i* et $z_i = x_i - \overline{x}$. La notation classique est :

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

Le *I* de Moran est en général :

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

ce qui désigne quelquefois (définition de **F**) :

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

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

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

Le c de Geary vaut :

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

qui semble plutôt utilisé comme :

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

Ces notions introduisent la dualité variance en $1/n$ et variance en $1/(n-1)$. C'est une différence mineure mais deux définitions principales et deux usages de la matrice **W** dont on peut se demander si ils ont la même signification.

Notons que, dans tous les cas, on retrouve la division par $\sum_{(2)} w_{ij}$ et $\sum_{i=1}^{n} z_i^2 / n$ ou l'équivalent en *n*-1 et que le centrage à pondération uniforme est préalable dans l'indice de Moran et sans effet dans l'indice de Geary puisque :

$$z_i - z_j = x_i - \overline{x} - x_j + \overline{x} = x_i - x_j$$

L'absence de structure spatiale est décrite par l'hypothèse nulle "z_i est la réalisation d'une variable aléatoire gaussienne de loi $N(\mu, \sigma^2)$" (modèle gaussien) ou par l'hypothèse nulle

"les observations sont distribuées dans l'espace par tirage au hasard dans l'espace des $n!$ permutations des n premiers entiers" (modèle non paramétrique). Dans ce cas, on peut soit utiliser une approximation de la loi de la statistique basée sur les moments ou générer des titages aléatoires (test de randomisation). Cette dernière technique l'emporte sur les autres et limitent les discussions byzantines.

La signification de ces indices ne posent pas de problèmes majeurs. Ils ont été abondamment commentés. Si on fait l'impasse sur le $1/n$ ou $1/(n-1)$, les deux indices utilise une variable **z** sous sa forme centrée (moyenne nulle) et réduite (variance 1). L'indice de Moran est en général utilisé dans une des trois possibilités :

1) graphe de voisinage binaire, non orienté, symétrique de matrice d'incidence **M** avec m arêtes, soit $2m$ couples de voisins ou encore $\mathbf{1}_n^t \mathbf{M} \mathbf{1}_n = 2m$:

$$I = \frac{1}{2m} \mathbf{z}^t \mathbf{M} \mathbf{z} = \mathbf{z}^t \mathbf{F} \mathbf{z}$$

2) pondération de voisinage binaire, symétrique de matrice **F** après normalisation globale :

$$I = \mathbf{z}^t \mathbf{F} \mathbf{z}$$

3) pondération de voisinage markovienne, normalisée par lignes, de matrice **L** :

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

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

Dans cette optique, la fonction importante est **`lag.listw`** : elle calcule pour un vecteur **x** de longueur n et de composantes x_i le vecteur de composantes $y_i = \sum_{j \text{ voisin de } i} w_{ij} x_j$ ou encore $\mathbf{y} = \mathbf{L}\mathbf{x}$ dit *lag vector*.

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

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

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

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

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

```
sum(lag.listw(pond.ext.u,rep(1,25)))
[1] 1
```

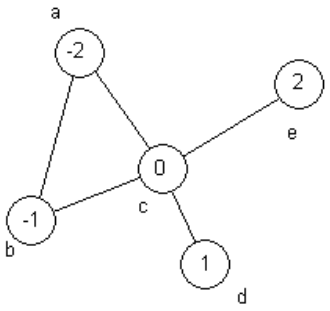
Pour l'option C, on trouve les moyennes par voisin déformées par le rapport du poids de voisinage sur le poids uniforme. Nous n'utiliserons par la suite que l'option W pour laquelle l'opération à un sens ordinaire en analyse des données.

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

print(lag.listw(pond.ext.s,rep(1,25)))
 [1] 1.0765 1.1317 0.8415 0.9527 0.5228 1.0481 0.7178 1.1089 1.1240 1.0994
[11] 1.1071 1.0101 1.0175 0.7265 0.8836 1.0749 0.8828 1.1593 1.2411 0.8909
[21] 1.3662 0.9962 1.1092 1.0359 0.8754

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

4.2. Mesure de variance locale

La présence du carré de la différence qui ne distingue pas les couples (i,j) et (j,i) fait que l'indice de Geary n'a de sens que dans les deux premières situations. Dans ce cas, les deux expressions jumelles $\sum_{(2)} w_{ij} z_i z_j$ (Moran) et $\sum_{(2)} w_{ij} \left(x_i - x_j\right)^2$ (Geary) semblent devoir être liées.

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



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

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

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 :

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

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 :

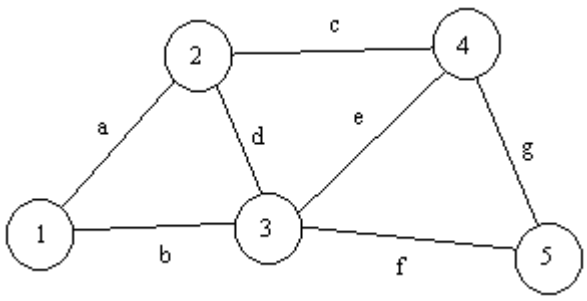
$$\sum_{i=1}^{n}\sum_{j=1}^{n} p_i p_j \left(x_i - x_j\right)^2 = 2\sum_{i=1}^{n} p_i \left(x_i - \bar{x}\right)^2$$

La variance à pondération quelconque est *la moitié* de la moyenne des carrés des différences élémentaires. Si on revient à l'*espace inconcevable*, celui dans lequel chaque couple porte le poids $1/n(n-1)$:

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

On retrouve 2 fois la variance en $1/(n-1)$. Pour une vraie pondération de voisinage l'indice de Geary mesure la variabilité locale et l'indice de Moran mesure la covariance locale (ou autocorrélation). Ces deux approches sont presque complémentaires sans l'être tout à fait. Ce qui sépare les deux indices est qu'en général on les emploie sur des systèmes de poids différents. On ne peut pas les comparer.

Pour redéfinir la famille des indices de Geary de manière plus efficace, on utilise la remarque fondamentale de départ dans Banet et Lebart (1984). 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 :

$$\mathbf{L} = \begin{array}{c} a \\ b \\ c \\ d \\ e \\ f \\ g \end{array} \begin{bmatrix} 1 & -1 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 \\ 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 1 & -1 \end{bmatrix} \qquad \mathbf{M} = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{bmatrix}$$

$$\mathbf{N} = \begin{bmatrix} 2 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 \\ 0 & 0 & 4 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 2 \end{bmatrix}$$

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

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

$$\sum_{(2)} f_{ij}\left(x_i - x_j\right)^2 = \frac{1}{2m}\sum_{i \text{ voisin de} j}\left(x_i - x_j\right)^2 = 2\mathbf{x}^t\left(\mathbf{D}-\mathbf{F}\right)\mathbf{x} = 2\mathbf{z}^t\left(\mathbf{D}-\mathbf{F}\right)\mathbf{z}$$

D'où :

$$I = \frac{\mathbf{z}^t\mathbf{F}\mathbf{z}}{\sum_{i=1}^n z_i^2 \Big/ n} \qquad c = \frac{\mathbf{z}^t\left(\mathbf{D}-\mathbf{F}\right)\mathbf{z}}{\sum_{i=1}^n z_i^2 \Big/ \left(n-1\right)}$$

Ces propriétés restent vraie pour une matrice de poids de voisinage **W** quelconque. On peut simplifier en normalisant à priori. Une modification mineure donne alors :

$$I^* = \mathbf{z}^t\mathbf{F}\mathbf{z} \qquad c^* = \mathbf{z}^t\left(\mathbf{D}-\mathbf{F}\right)\mathbf{z}$$

Ces relations extrêmement simples cachent en fait de nombreux problèmes qui ont beaucoup nui à leur usage effectif. Rappelons ici que **F** est une matrice de poids de voisinage sur couple de points et que **D** est la matrice diagonale des poids de voisinage des points qui en découle, soit :

$$f_{i.} = \sum_{j=1}^n f_{ij} \quad \mathbf{D} = diag\left(f_{1.},...,f_{n.}\right)$$

4.3. Tests contre l'absence de structures spatiales

Pour faire le *test de Geary*, dans le modèle non paramétrique de l'équiprobabilité des *n*! permutations des données :

```
unclass(geary.test(irishdata$tab$cow,pond.ext.b))
$statistic
Geary C statistic standard deviate
                             -3.37
$p.value
 Kildare
0.000376
$estimate
Geary C statistic       Expectation          Variance
           0.3110            1.0000            0.0418
$alternative
[1] "less"
$method
[1] "Geary's C test under randomisation"
$data.name
[1] "irishdata$tab$cow \nweights: pond.ext.b \n"
```

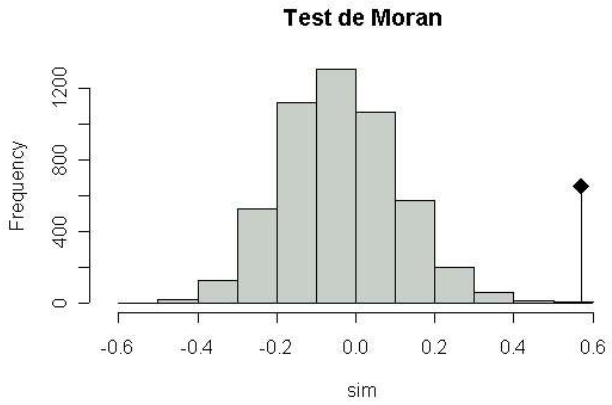
```
geary.test(irishdata$tab$cow,pond.ext.b)

        Geary's C test under randomisation

data:  irishdata$tab$cow
weights: pond.ext.b

Geary C statistic standard deviate = -3.37, p-value = 0.000376
alternative hypothesis: less
sample estimates:
Geary C statistic       Expectation          Variance
           0.3110            1.0000            0.0418
```

Pour faire le *test de Moran*, dans le modèle non paramétrique de l'équiprobabilité des *n*! permutations des données :

```
unclass(moran.test(irishdata$tab$cow,pond.ext.w))
$statistic
      [,1]
[1,] 5.253
attr(,"names")
[1] "Moran I statistic standard deviate"
$p.value
          [,1]
[1,] 7.479e-08
$estimate
Moran I statistic       Expectation          Variance
          0.67358          -0.04167           0.01854
$alternative
[1] "greater"
$method
[1] "Moran's I test under randomisation"
$data.name
[1] "irishdata$tab$cow \nweights: pond.ext.w \n"
```

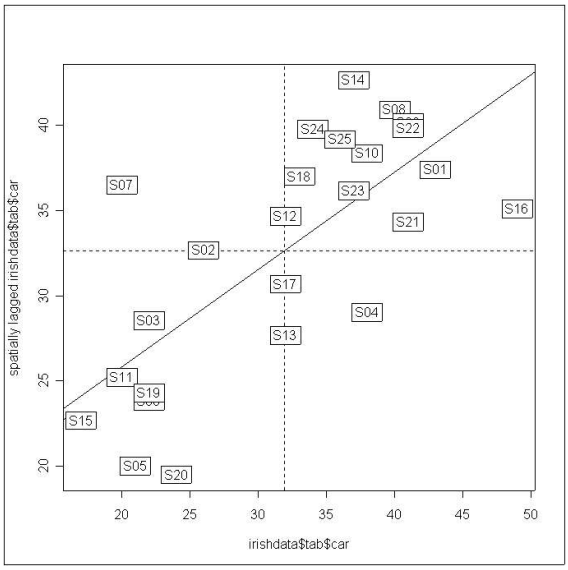
Pour faire la version de Monte-Carlo :

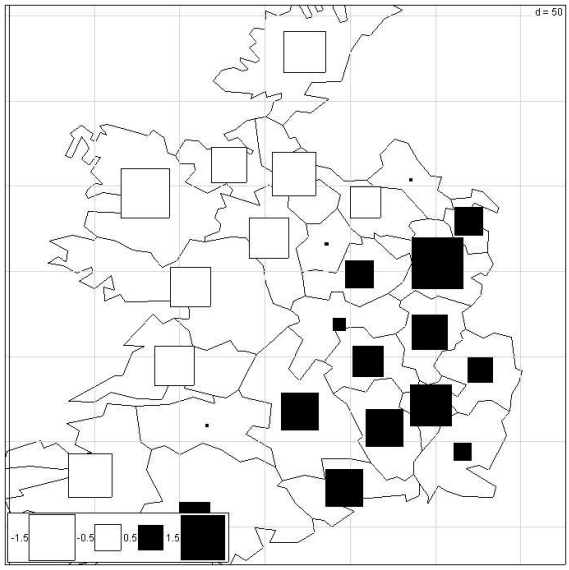
```
t1 = moran.mc(irishdata$tab$car,pond.ext.w,5000)
plot(as.randtest(t1$res,t1$statistic),main="Test de Moran")
```




```
moran.plot(irishdata$tab$car,pond.ext.w)
w = cbind.data.frame(irishdata$tab$car,
      lag.listw(pond.ext.w,irishdata$tab$car))
s.label(w,add.pl=T,label=row.names(irishdata$tab))
```

```
Potentially influential observations of
         lm(formula = wx ~ x) :

   dfb.1_ dfb.x dffit   cov.r   cook.d hat
7    0.86  -0.75  0.92_*  0.73_*  0.34   0.12
15 -0.13    0.12 -0.13    1.29_*  0.01   0.16
```





Scatter-plot de Moran (Anselin 1995, 1996). *En abscisse les valeurs d'une variable. En ordonnée la moyenne des valeurs des voisins (la variable 'retard'). La droite est l'estimation du modèle y = ax+b. Les deux droites pointillées passent par les moyennes. Se lit avec la formule : 4 quadrants :*

grand-grand ou petit-petit groupement spatial
petit-grand ou grand-petit aberration spatiale.

La pente reflète l'autocorrélation.

Pour faire le test pour toutes les variables d'un tableau :

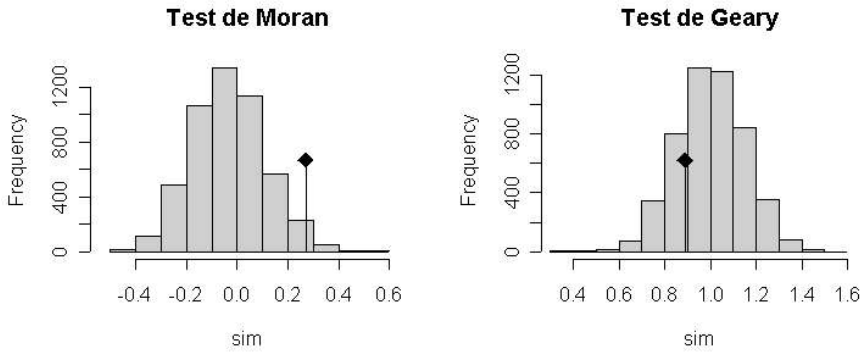
```
geary.s = as.numeric(unlist(lapply(apply(irishdata$tab,2,geary.test,
          listw=pond.ext.s),function(x) x$p.value)))
geary.b = as.numeric(unlist(lapply(apply(irishdata$tab,2,geary.test,
```

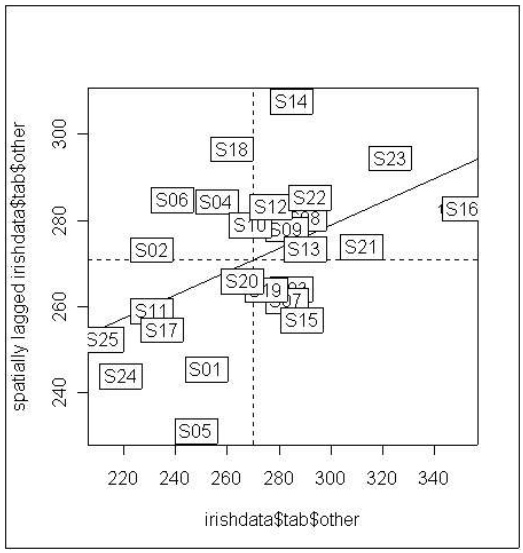
```
            listw=pond.ext.b),function(x) x$p.value)))
moran.w = as.numeric(unlist(lapply(apply(irishdata$tab,2,moran.test,
            listw=pond.ext.w),function(x) x$p.value)))
w = cbind.data.frame(geary.s,geary.b,moran.w)
row.names(w) = names(irishdata$tab)
w
```

```
             geary.s   geary.b   moran.w
T0.10      4.185e-05 6.530e-04 1.465e-05
T10.50     1.026e-02 2.455e-02 5.083e-02
Tup50      5.048e-07 3.757e-06 4.395e-07
cow        1.027e-05 3.760e-04 7.479e-08
other      1.389e-01 1.759e-01 1.330e-02
pig        3.069e-03 1.603e-02 1.523e-04
sheep      1.637e-01 2.300e-01 5.818e-02
town.pop   8.644e-03 6.580e-03 2.169e-02
car        7.173e-05 4.474e-04 1.114e-05
radio      7.141e-02 4.684e-02 2.060e-01
sales      2.902e-04 7.149e-04 5.887e-04
single.man 2.278e-03 1.981e-03 2.542e-03
```

On a beaucoup discuté de la puissance de ces tests. Globalement le *I* de Moran l'emporte sur le *c* de Geary. En cas de doute les tests de permutations sont les meilleurs :

```
t1 = moran.mc(irishdata$tab$other,pond.ext.w,5000)
t2 = geary.mc(irishdata$tab$other,pond.ext.s,5000)
plot(as.randtest(t1$res,t1$statistic),main="Test de Moran")
plot(as.randtest(t2$res,t2$statistic),main="Test de Geary")
```






```
moran.plot(irishdata$tab$other,pond.ext.w)
w = cbind.data.frame(irishdata$tab$other,
            lag.listw(pond.ext.w,irishdata$tab$other))
```

```
s.label(w,add.pl=T,label=row.names(irishdata$tab))
```

Seule la variable radio semble n'être pas spatialement structurée. On la laisse comme témoin.

5. Hésitations méthodologiques

On arrive à la question de fond. *n* unités statistiques portent une pondération de voisinage du type W de R. Bivand. On note **L** la matrice à *n* lignes et *n* colonnes dont le terme général est L_{ij} poids de voisinage du point j pour le point *i*. On a :

$$\sum_{j=1}^{n} L_{ij} = L_{i\bullet} = 1 \Leftrightarrow \mathbf{L}\mathbf{1}_n = \mathbf{1}_n$$

La matrice **L** n'a qu'une existence théorique, les calculs se faisant à partir de liste de voisins (classe **listw** de **spdep**). Un schéma de dualité, objet de la classe **dudi** ou triplet statistique à *n* lignes et *p* colonnes s'écrit $(\mathbf{X},\mathbf{Q},\mathbf{D})$. Le but est d'introduire le point de vue de voisinage **L** dans l'analyse de $(\mathbf{X},\mathbf{Q},\mathbf{D})$ en partant de l'indice de Moran. Ce choix est loin d'être naturel et Lebart a fait le contraire en partant de l'indice de Geary.

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

Une remarque s'impose sur le choix par Lebart de *c* au détriment de *I* dans l'approche multivariée. On comprend facilement que $c^* = \mathbf{z}^t(\mathbf{D}-\mathbf{W})\mathbf{z}$ mesure la variance locale. Elle s'écrit :

$$c^* = \mathbf{z}^t(\mathbf{D}-\mathbf{W})\mathbf{z} = \frac{1}{2}\sum_{i,j} f_{ij}(x_i - x_j)^2 = \frac{1}{2}\sum_{i,j} f_{ij}(z_i - z_j)^2$$

La généralisation de Lebart (1969) introduit la matrice de covariance spatiale $\mathbf{X}^t(\mathbf{D}-\mathbf{P})\mathbf{X}$ à partir des graphes de voisinage.

L'idée a été reprise par Monestiez (1978) et généralisée aux pondérations quelconques dans le cadre de l'ACP par Le Foll (1982). Le Foll met en évidence l'opérateur **D**-symétrique $\mathbf{I}_n - \mathbf{D}^{-1}\mathbf{P}$. Mom (1988) admet une pondération extérieure **D** qui sommée sur les voisins donne une surpondération de voisinage $\mathbf{D}_*$, l'opérateur de lissage $\mathbf{H} = \mathbf{D}_*^{-1}\mathbf{M}\mathbf{D}$ et l'analyse de $((\mathbf{I}_n - \mathbf{H})\mathbf{X},\mathbf{Q},\mathbf{D})$. Méot et al (1993), dans la même situation introduise l'opérateur **D**-symétrique $\mathbf{D}_* - \mathbf{M}\mathbf{D}$ mais tous conservent des formes quadratiques positives donc le point de vue initial de la variance locale qui donne pour deux tableaux l'analyse de covariance locale (Chessel and Sabatier 1993). Toutes ces approches portent sur la variabilité de voisinage dites encore analyses factorielle des différences locales (Benali and Escofier 1990).

Pourquoi donc l'idée de Lebart ne s'est-elle pas imposée concrètement. Le doute s'installe quand Benali et Escofier, dans le même article, mettent en avant l'existence de l'objectif inverse sous la forme de **l'analyse factorielle lissée** ou analyse de $(\mathbf{L}\mathbf{X},\mathbf{Q},\mathbf{D})$, l'analyse du tableau des moyennes de voisinage. On y diagonalise encore un opérateur positif. L. Lebart a certainement eu le mérite d'ouvrir le débat et de connecter le multivarié et le spatial. Il l'a

fait sur la base du multivarié en introduisant le spatial par le biais d'une métrique euclidienne.

Ce faisant on restait fortement attaché au standard, ce qui est particulièrement sensible dans son AFC d'un graphe de voisinage (Lebart 1984) véritable erreur stratégique qui mélange les autocorrélations positives et négatives par le biais de leurs carrés. Cette communauté est restée fermée comme en témoigne encore l'intervention de Aufaure et al. (2000) et la généralisation aux cubes de données de Cornillon et al. (1999), à l'exception cependant des contacts avec l'école italienne : Di Bella et Jona-Lasinio (1996) l'utilise dans le champ de l'ordination multi-échelles ouvert par Ver Hoef et Glenn-Lewin (1989).

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

$$\langle \mathbf{y} | \mathbf{z} \rangle_c = \mathbf{y}^t (\mathbf{D} - \mathbf{W}) \mathbf{z} = \sum_{i,j} w_{ij} (y_i - y_j)(z_i - z_j)$$

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

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

Seul Wartenberg (1985b) a osé casser la contrainte qu'une analyse doit donner des valeurs propres positives. Il diagonalise $\mathbf{M} = \mathbf{X}^t\mathbf{W}\mathbf{X} = \mathbf{X}^t\mathbf{F}\mathbf{X}$ non sans précaution :

> *An important difference between this approach and PCA must be pointed out. Unlike* **R***, the product-moment correlation matrix that is decomposed in PCA,* **M** *is not positive definite. That is,* **M** *can have negative eigenvalues, which* **R** *cannot. These negative eigenvalues are as important as positive eigenvalues but are of a qualitatively different type. They represent spatial interaction (covariance) that is more important than spatial pattern (variance). ... To avoid this situation, data yielding negative eigenvalues are not used in this paper. All examples have large eigenvalues that are positive only.*

Il sait que son analyse pourrait donner de grandes valeurs propres négatives ayant du sens mais le cache provisoirement. Il y a cependant une contradiction en ce sens que l'indice de Moran prend tout son intérêt sur un lien **L** et que l'analyse utilise l'indice de Moran sur un lien **F**. Ces hésitations font qu'il y a peu d'utilisateurs de ces propositions auxquelles on préfère les classifications sous contraintes spatiales (spatial clustering) ou les méthodes géostatistiques multivariées (Wackernagel 2003) comme dans Monestiez et al. (1994). Les méthodes d'ordination sous contraintes spatiales sont les grandes absentes de la synthèse remarquable qui vient d'être publiée dans *Ecography* par le groupe de travail *"Integrating the Statistical Modeling of Spatial Data in Ecology"* (Dale et al. 2002, Dungan et al. 2002, Keitt et al. 2002, Legendre et al. 2002, Liebhold and Gurevitch 2002, Perry et al. 2002).

Mais en géologie, en particulier en minéralogie, la situation est différente. Si on appelle **MSC** pour ***M**ultivariate **S**patial **C**orrelation* l'analyse de Wartenberg, **MSC** est voisine de variantes nées à la même époque. Elle n'est pas isolée conceptuellement mais le développement de méthodes nouvelles se fait souvent sur des idées voisines dans des environnements séparés. Est souvent mentionnée **SFA** pour ***S**patial **F**actor **A**nalysis* une analyse proposée et défendue par Grunsky et Agterberg (1988, 1989, 1991, Grunsky et al. 1996) alors que le terme "spatial factor analysis" renvoie souvent à **MAF** pour ***M**in/**M**ax **A**utocorrelation **F**actor **A**nalysis* créé par Switzer et Green (1984) dans un rapport souvent cité qui a été ensuite redécrit et utilisé à plusieurs reprises par Nielsen et son équipe (Conradsen et al. 1985, Ersbll 1989, Nielsen and Larsen 1994, Nielsen 1995a, b, Nielsen and Conradsen 1997, Nielsen et al. 1997, Nielsen et al. 1998, Nielsen 1999, Flesche et al. 2000).

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

Le lien spatial utilisé dans **SFA** en fait plutôt une curiosité. Le meilleur article de Grunsky (2002) est celui qui dit aux géologues que le logiciel R (http://lib.stat.cmu.edu/R/CRAN/) est une excellente plate-forme pour faire de la statistique appliquée, ce qui est bien vrai. Le lien de Wartenberg est le plus général et convient parfaitement en écologie et économie. Le lien de Nielsen est celui qui est adaptée à l'analyse des images de télédétection. Entre les deux, l'ajustement est purement technique.

Mais il y a entre les deux une différence de taille. La **MSC** diagonalise directement $\mathbf{R}_V$ alors que **MAF** est basée sur la diagonalisation de $\mathbf{R}^{-1}\mathbf{R}_V$ où $\mathbf{R}$ est la matrice de corrélation ordinaire. La **MSC** est une **ACP** (pour augmenter la covariation spatiale, on doit d'abord augmenter la variance) sous contrainte (la variance ne doit pas augmenter trop au détriment de l'indice de Moran). La **MAF** est apparentée aux analyses discriminantes, elle fournit des scores canoniques de moyenne 0 et variance 1 qui maximise strictement l'indice de Moran. Elle est invariante par combinaisons linéaires de rang plein des données départ. Ceci n'est possible, sans d'énormes problèmes de stabilité numérique, que pour des nombres de lignes considérables, ce qui est le cas en analyse d'images. De la télédétection à l'imagerie du cerveau en passant par les réseau de stations, il n'y a guère de sens à penser une méthode unique, pas plus qu'il y a unicité de la définition des pondérations de voisinage.

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

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

Cette quantité est d'autant plus grande que de grandes valeurs positives (respectivement négatives) se trouvent associées sur des couples ayant un grand poids de voisinage. Wartenberg (1985c) a utilisé l'autocorrélation spatiale dans l'interprétation d'une analyse ordinaire en sélectionnant les coordonnées des analyses ordinaires qui donnait une forte valeur, pour approfondir l'usage des coordonnées concrètes dans l'espace comme données numériques (Wartenberg 1985a) puis en diagonalisant la matrice des covariance spatiale définie par les produits (Wartenberg 1985b) :

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

Cette quantité malheureusement n'est un coefficient de corrélation exactement que si le centrage est fait avec une moyenne calculée avec les poids de voisinage des points et si la normalisation est faite en divisant par l'écart-type utilisant la même pondération. En outre, cette forme quadratique n'est pas positive et l'analyse peut avoir des valeurs propres négatives. Cette insertion n'est pas optimum du point de vue mathématique, tout en étant très légitime du point de vue expérimental. C'est moins beau, mais c'est beaucoup plus utile.

On peut concilier les deux points de vue (Thioulouse et al. 1995) en n'utilisant que des données **D**-normalisées, c'est à dire en prenant :

$$\bar{x} = \sum_{i=1}^{n} f_{i.} x_i \quad \mathrm{var}(\mathbf{x}) = \sum_{i=1}^{n} f_{i.} (x_i - \bar{x})^2 \quad z_i = \frac{x_i - \bar{x}}{\sqrt{\mathrm{var}(\mathbf{x})}}$$

Alors pour toute variable ayant subi ce traitement :

$$\mathbf{z}^t \mathbf{D} \mathbf{z} = \mathbf{1} = \mathbf{z}^t (\mathbf{D} - \mathbf{F}) \mathbf{z} + \mathbf{z}^t \mathbf{F} \mathbf{z}$$

Variance totale = 1=variance locale +autocovariance.

Cette décomposition est curieuse en ce que deux termes seulement sur les trois sont toujours positifs. Elle est abondamment commentée par Durand et al. (1999) et Ghertsos et al. (2001). Pour un processus "lisse" donc fortement cartographiable la variance locale est faible (mais positive) et la covariance locale est positive et forte. Pour un processus à forte variation entre voisins, 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 pourrait croire la question résolue mais ce point de vue cache un gros inconvénient. Dans l'approche inférentielle, en effet, la pondération non uniforme qui intervient dans le calcul de la moyenne et la variance fait que cette moyenne et cette variance ne sont pas des invariants dans l'espace des *n*! permutations des données.

Si une variable n'a pas de structure spatiale (hypothèse nulle), sa moyenne et sa variance sont estimées par des unités statistiques toutes égales. Si au contraire elle en a une et c'est ce qu'on veut mettre en évidence au niveau multivarié les points sont d'autant plus importants qu'ils ont un poids de voisinage plus grands. Un point relativement isolé dans une telle analyse est une perturbation qui n'a pas grand chose à dire, un point central joue un grand rôle dans l'analyse de la structure. La pondération uniforme des données impose la pondération uniforme, le bilan multivarié impose plutôt la pondération de voisinage. Nous nous en tiendrons ici au point de vue de Wartenberg dans la lignée de Moran.

6. La fonction `multispati`

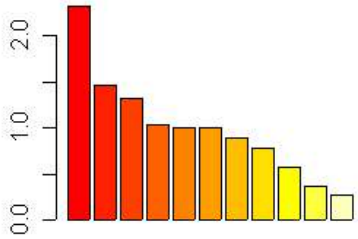
6.1. Paramètres

La fonction utilise un quadruplet $\left(\underset{n\times p}{\mathbf{X}}, \underset{p\times p}{\mathbf{Q}}, \underset{n\times n}{\mathbf{D}}, \underset{n\times n}{\mathbf{L}}\right)$ dont les dimensions sont indiquées en associant une pondération de voisinage (classe **`listw`**) à un schéma de dualité (classe **`dudi`**).

$(\mathbf{X},\mathbf{Q},\mathbf{D})$ est un schéma de dualité ou analyse de premier niveau (en cas de besoin, voir la fiche http://pbil.univ-lyon1.fr/R/stage/stage3.pdf). **X** est un tableau, **Q** une pondération de ses colonnes et **D** une pondération de ses lignes. Le plus utile des objets de ce type dérive d'un data frame quelconque contenant des variables quantitatives (***`numeric`***) et des variables qualitatives (***`factor`***) voire même des qualitatives à modalités ordonnée (***`ordered`***).

Les quantitatives sont centrées et réduites, les qualitatives décomposées en indicatrices de classes puis centrées correctement et les pondérations font en sorte que chaque variable ait le même poids que les autres. La fonction **`dudi.mix`** assure cette opération.

```
data(oribatid)
names(oribatid)
[1] "fau"   "envir" "xy"

ori.mix=dudi.mix(oribatid$envir)
Select the number of axes: 3
```




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

$nf: 3 axis-components saved
# Une valeur propre intéressante et beaucoup d'inertie désorganisée
$rank: 11
eigen values: 2.312 1.456 1.316 1.031 1 ...

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

  data.frame nrow ncol content
1 $tab       70   14   modified array
2 $li        70   3    row coordinates
3 $l1        70   3    row normed scores
4 $co        14   3    column coordinates
5 $c1        14   3    column normed scores
other elements: assign index cr
```

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

```
ori.mix$cw
 subst.inter subst.litter   subst.peat   subst.sph1   subst.sph2   subst.sph3
     0.38571      0.02857      0.02857      0.35714      0.15714      0.01429
  subst.sph4    shrub.few   shrub.many   shrub.none topo.blanket topo.hummock
     0.02857      0.37143      0.35714      0.27143      0.62857      0.37143
     density        water
     1.00000      1.00000
sum(ori.mix$cw)
[1] 5
```

La pondération des lignes est uniforme :

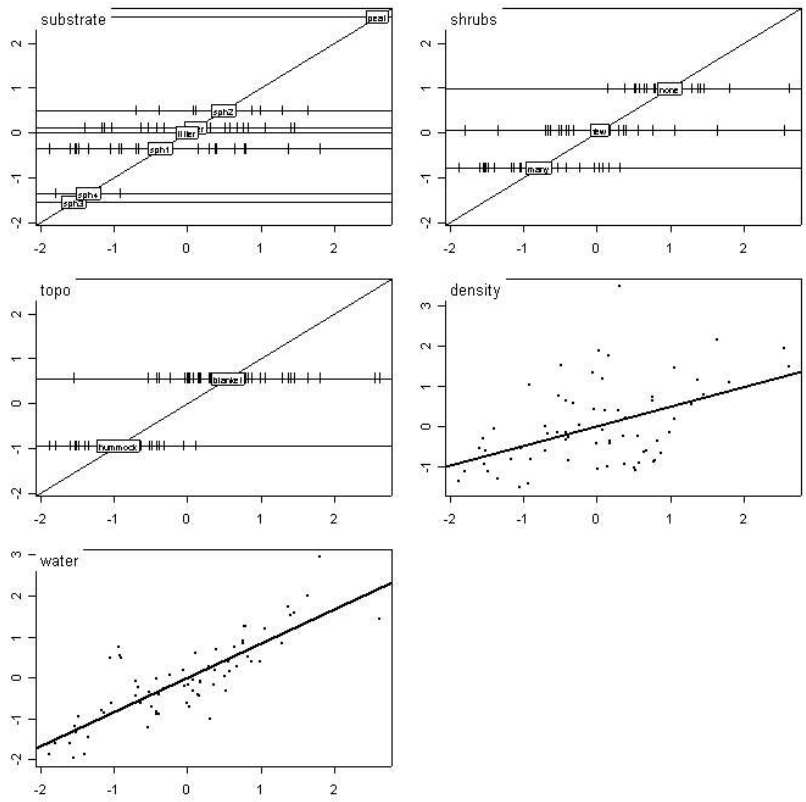
```
unique(ori.mix$lw)
[1] 0.01429
1/nrow(ori.mix$tab)
[1] 0.01429
```

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

$$\mathbf{U}_s'\mathbf{Q}\mathbf{U}_s = \mathbf{I}_s$$

Les coordonnées de l'analyse simple $\mathbf{L}_s = \mathbf{X}\mathbf{Q}\mathbf{U}_s$ maximise successivement l'inertie projetée sur un axe **u** soit $\|\mathbf{X}\mathbf{Q}\mathbf{u}\|_{\mathbf{D}}^2$. Les maxima successifs sont les valeurs propres de l'analyse simple qu'on notera $\sigma_1,...,\sigma_s$. Dans cet exemple, cela signifie que la première coordonnée est un score numérique **z** qui maximise la somme des quantités $corr^2(\mathbf{z},\mathbf{X}[,j])$ quand la variable $\mathbf{X}[,j]$ est quantitative et $\eta^2(\mathbf{z},\mathbf{X}[,j])$ quand elle est qualitative. Cette analyse est une ACP normée sur matrice de corrélation quand il n'y a que des quantitatives et une Analyse des Correspondances Multiples quand il n'y a que des qualitatives.

```
ori.mix$cr
             RS1      RS2     RS3
substrate 0.3655 0.555784 0.76607
shrubs    0.4874 0.354482 0.25668
topo      0.5242 0.031934 0.20871
density   0.2370 0.509574 0.01215
water     0.6982 0.003817 0.07214

sum(ori.mix$cr[,1])
[1] 2.312

ori.mix$eig[1]
[1] 2.312

score(ori.mix)
```



Expression du lien entre les variables par le lien de chacune d'entre elles avec un score de synthèse.

Une analyse de base ou triplet $(\mathbf{X},\mathbf{Q},\mathbf{D})$ ou schéma de dualité (Escoufier 1987) a de nombreuses variantes comme :

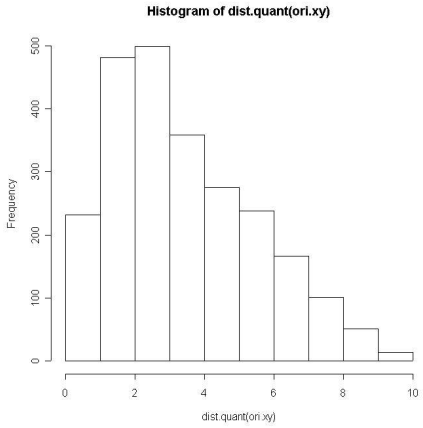
Fonction	Analyse	Référence
dudi.pca	Principal component Analysis	1
dudi.coa	Correspondence Analysis	2
dudi.acm	Multiple Correspondence Analysis	3
dudi.fca	Fuzzy Correspondence Analysis	4
dudi.mix	Mixture of numeric and factors	5
dudi.nsc	Non Symetric Correspondence Analysis	6
dudi.dec	Decentred Correspondence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

Les fonctions dudi. 1—équivalent de prcomp/princomp, 2—Greenacre (1984), 3—Tenenhaus and Young (1985), 4—Chevenet et al.(1994), 5—Hill and Smith (1976), Kiers (1994) 6—Kroonenberg and Lombardo (1999), 7—Dolédec et al. (1995).

A ce triplet on associe un opérateur de retard $\mathbf{L}$ qui permet de calculer $\mathbf{Y}=\mathbf{LX}$ où chaque valeur initiale au point *i* de la variable *j* est remplacée par la moyenne des valeurs des voisins de *i* pour la même variable *j*. Pour une variable on a $\mathbf{y}=\mathbf{Lx}$ et le graphe du couple $(\mathbf{x},\mathbf{y})$ est la *Moran-scatterplot* d'Ansellin. Ainsi étendu, l'opération génère un deuxième tableau totalement apparié au premier et donc un deuxième nuage de *n* points de $\mathbb{R}^p$ qu'on peut projeter sur les axes principaux.

```
hist(dist.quant(ori.xy))
1 = Canonical
d1 = ||x-y|| A=Identity
2 = Joreskog
d2=d2 = ||x-y|| A=1/diag(cov)
3 = Mahalanobis
d3 = ||x-y|| A=inv(cov)
```

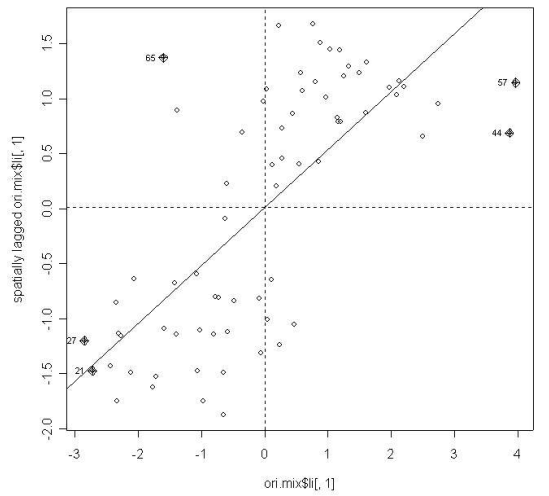
```
Select an integer (1-3): 1
```

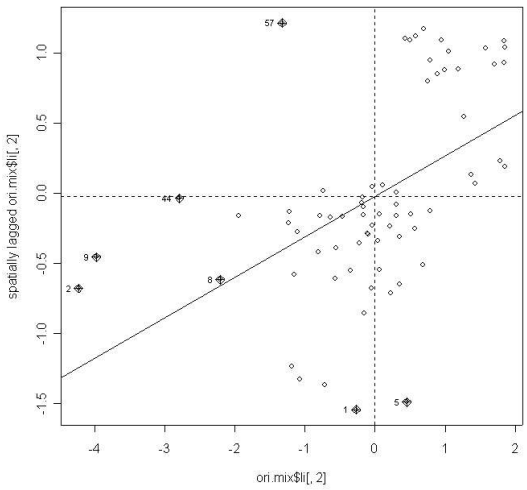



```
ori.dn=dnearneigh(as.matrix(ori.xy),0,1)
print(card(ori.dn))
# le nombre de voisin par points
 [1]  1  3  3  3  3  6  0  6  7  5  6  3  4  8  6  5  5  6  5  2  3  6  5  6  6
[26]  6  6  4  7  5  4  7 10 10  4 10 11  8 14  9  9  7  8  9 14  9  9  9  8  7
[51]  7  5  5  8  7  3 10  5  3  8  9 10  9 10  7 10  5  4  5  1

ori.dn[[7]]
# un point sans voisin
[1] 0

ori.listw=nb2listw(ori.dn)
Error in nb2listw(ori.dn) : Empty neighbour sets found

ori.dn=dnearneigh(as.matrix(ori.xy),0,1.5)
print(card(ori.dn))
 [1]  4  5  9  9  6 10  7 11 11 15 10  8  9 11 11  9 11 12 10 17 11 17  9 13 14
[26] 15 12 12 13 13 14 17 19 21 10 19 14 17 18 16 13 13 15 16 21 19 19 17 17 19
[51] 15 11 14 21 11 10 16 13  8 16 19 15 14 13 14 13 10  9  9  3
ori.listw=nb2listw(ori.dn)
```

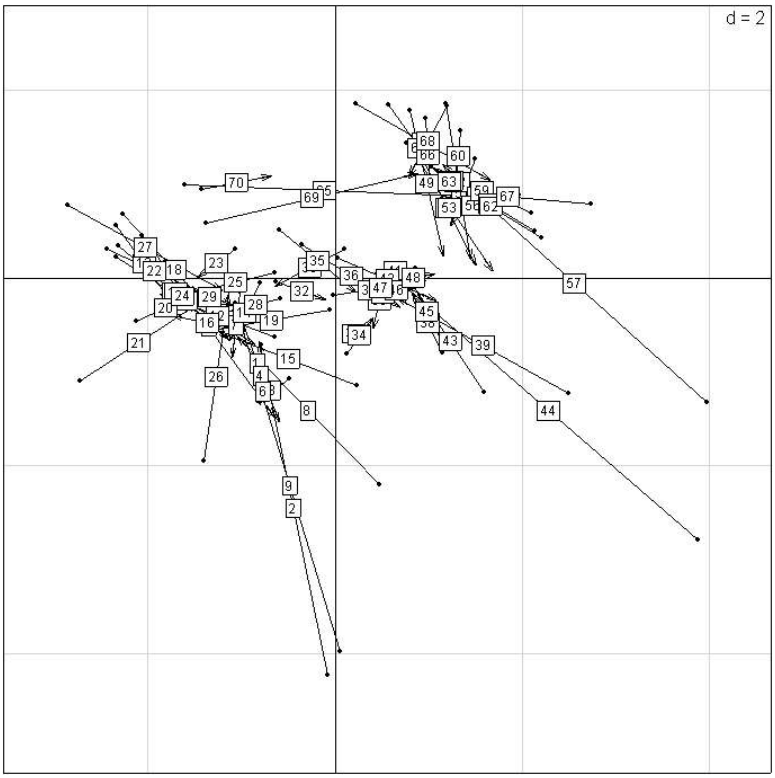
On peut donc calculer l'autocorrélation des coordonnées :

```
lapply(ori.mix$li,moran.mc,listw=ori.listw,nsim=999)
$Axis1 statistic = 0.5274, observed rank = 1000, p-value = 0.001
$Axis2 statistic = 0.2888, observed rank = 1000, p-value = 0.001
$Axis3 statistic = 0.0953, observed rank = 981, p-value = 0.019

moran.plot(ori.mix$li[,1],ori.listw)
moran.plot(ori.mix$li[,2],ori.listw)
```





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



La carte factorielle ordinaire est celle des points à l'origine des flèches. L'extrémité de la flèche est la position moyenne des voisins du point. La carte factorielle ordinaire maximise la variance projetée. La nouvelle analyse gardera une part de cette propriété mais intégrera le lien de voisinage.

6.2. Principes

On comprend que chaque axe de $\mathbb{R}^p$ définit un système de coordonnées qui est plus ou moins autocorrélé. Les axes principaux de l'analyse simple maximise la variance projetée et n'on aucune propriété d'autocorrélation particulière. On cherche alors ceux qui maximise l'autocorrélation. La solution n'est pas ordinaire car le critère est :

$$I(\mathbf{XQv}) = \frac{\mathbf{v}^t\mathbf{Q}^t\mathbf{X}^t\mathbf{DLXQv}}{\mathbf{v}^t\mathbf{Q}^t\mathbf{X}^t\mathbf{DXQv}}$$

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

$$Max_{\|\mathbf{v}\|_\mathbf{Q}=1}\langle\mathbf{Hv}|\mathbf{v}\rangle_\mathbf{Q} = \mathbf{v}^t\mathbf{Q}^t\mathbf{X}^t\mathbf{DLXQv} = \|\mathbf{XQv}\|_\mathbf{D}^2 I(\mathbf{v}) = \mathrm{var}(\mathbf{XQv})I(\mathbf{XQv})$$

Le cas particulier pour une ACP normée est l'analyse de Wartenberg (1985b) quand on utilise un lien normalisé par lignes ou la **MAF** de Switzer et Green (1984) étendue à une pondération de voisinage quelconque mais sans inversion de métrique. On appelera **MS**

pour multivarié spatial la recherche de la base de vecteurs propres de **H** et son usage. La fonction multispati fait cela.

6.3. Un test de permutation multivarié

Notons d'abord qu'on retrouve un élément des calculs du corrélogramme de Smouse et Peakall (1999). L'argument 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 retrouve ce calcul explicité par Kuo (2002) :

d_{ij}^2 is a distance based on a equilateral tetrahedron method (Smouse and Peakall, 1999).

- $d_{ij}^2 = 0$ for 2 identical genotypes.
- $d_{ij}^2 = 1$ when 2 genotypes shared 1 allele.
- $d_{ij}^2 = 2$ when a heterozogote-heterozygote pair didn't share any alleles.
- $d_{ij}^2 = 3$ when a homozygote-heterozygote pair didn't share any alleles.
- $d_{ij}^2 = 4$ for 2 different homozygotes.
- Max. possible distance is 4 for each locus, and the distance is added across loci.

C'est clair mais il suffit de savoir qu'il s'agit de la métrique canonique sur les codages 2000 (homozygotes) et 1100 (hétérozygotes). 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 :

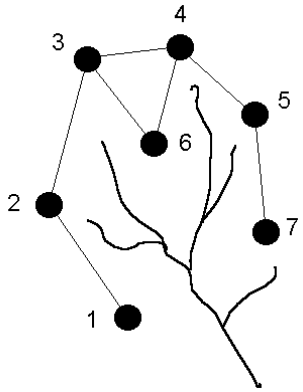
$$\mathbf{D}_1 = \left[d_{ij1}^2\right], \mathbf{D}_2 = \left[d_{ij2}^2\right], ..., \mathbf{D}_L = \left[d_{ijL}^2\right], \mathbf{D} = \left[d_{ij1}^2 + d_{ij2}^2 + ... + d_{ijL}^2\right]$$

L'analyse se poursuivant pour chacune des distances et la distance globale, nous en conservons une, par exemple **D**. En se référant à Gower (1966), 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 $\mathbf{C} = \left[-\frac{1}{2} d_{ij}^2\right]_{\bullet\bullet} = \mathbf{X}\mathbf{X}^t$ de l'analyse en coordonnées principales. Donc les données sont traitées par la *matrice des produits scalaires de la représentation euclidienne* associée à la distance génétique. Le tableau de départ est évidemment une représentation euclidienne de la matrice de distance qui en découle.

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 toute les matrices sont symétriques :

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

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

$$r = \frac{Trace\left(\mathbf{X}^t\mathbf{DLXQ}\right)}{Trace\left(\mathbf{X}^t\mathbf{DXQ}\right)}$$

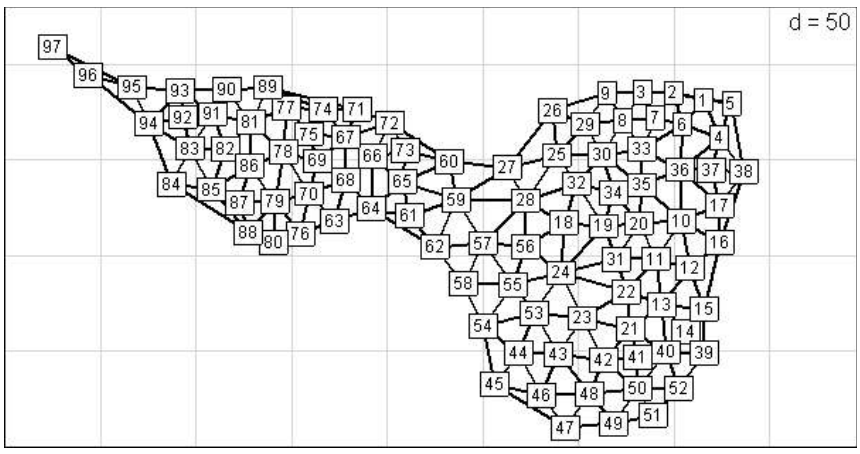
La définition de Smouse et Peakall est donc étendue à toute pondération de voisinages et à tout type d'analyse élémentaire. Le test de permutations associé considère que les lignes du tableau et leur poids dans l'analyse sont attribué au hasard dans l'espace. La fonction **`multispati.rtest`** fait le calcul dans R et **`multispati.randtest`** le fait en C avec une fonction externe.

7. Illustrations

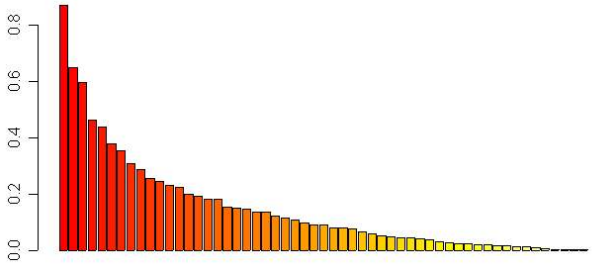
Nous pouvons alors reprendre les exemples introduits au début de cette fiche.

7.1. Analyse des correspondances à composantes cartographiables

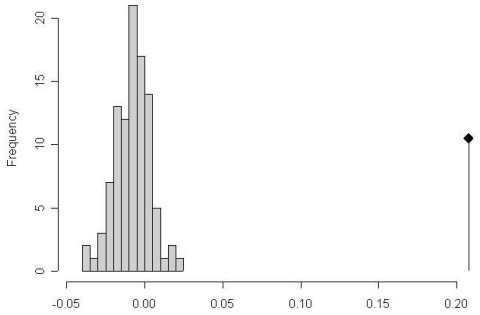
```
data(mafragh)
maf.xy=mafragh$xy
maf.flo=mafragh$flo
maf.listw=nb2listw(neig2nb(mafragh$neig))
s.label(maf.xy,neig=mafragh$neig,clab=0.75)
```



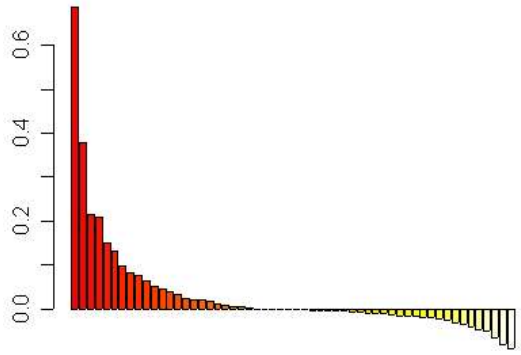
```
maf.coa = dudi.coa(maf.flo)
Select the number of axes: 3
```

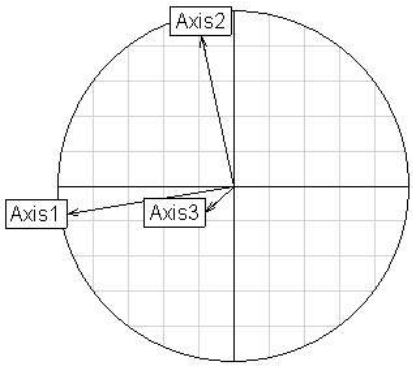



```
maf.coa.test=multispati.randtest(maf.coa,maf.listw)
plot(maf.coa.test)
```




```
maf.coa.ms=multispati(maf.coa,maf.listw)
Select the first number of axes (>=1): 2
Select the second number of axes (>=0): 0
```



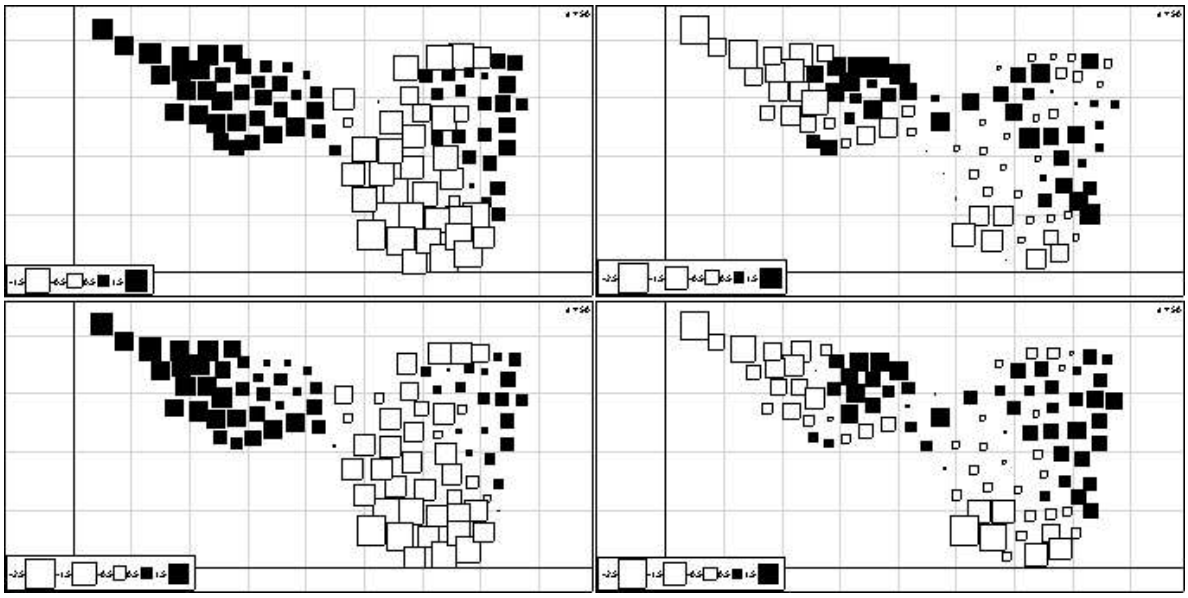


A

Comparer l'analyse simple et l'analyse spatialisée par :

```
summary(maf.coa.ms)

Multivariate Spatial Analysis
Call: multispati(dudi = maf.coa, listw = maf.listw)

Scores from the first duality diagramm:
       var    cum  ratio  moran
RS1 0.8691 0.8691 0.1043 0.7250 # 0.8691 > 0.8333
RS2 0.6491 1.5183 0.1823 0.4834 # 0.8691 + 0.6491 > 0.8333 + 0.5866
RS3 0.5975 2.1158 0.2540 0.2264

Eigenvalues decomposition:
       eig    var  moran
CS1 0.6855 0.8333 0.8226 # 0.6855 > 0.8691 * 0.7250 = 0.6301
CS2 0.3785 0.5866 0.6453

s.corcircle (maf.coa.ms$as) # A
```

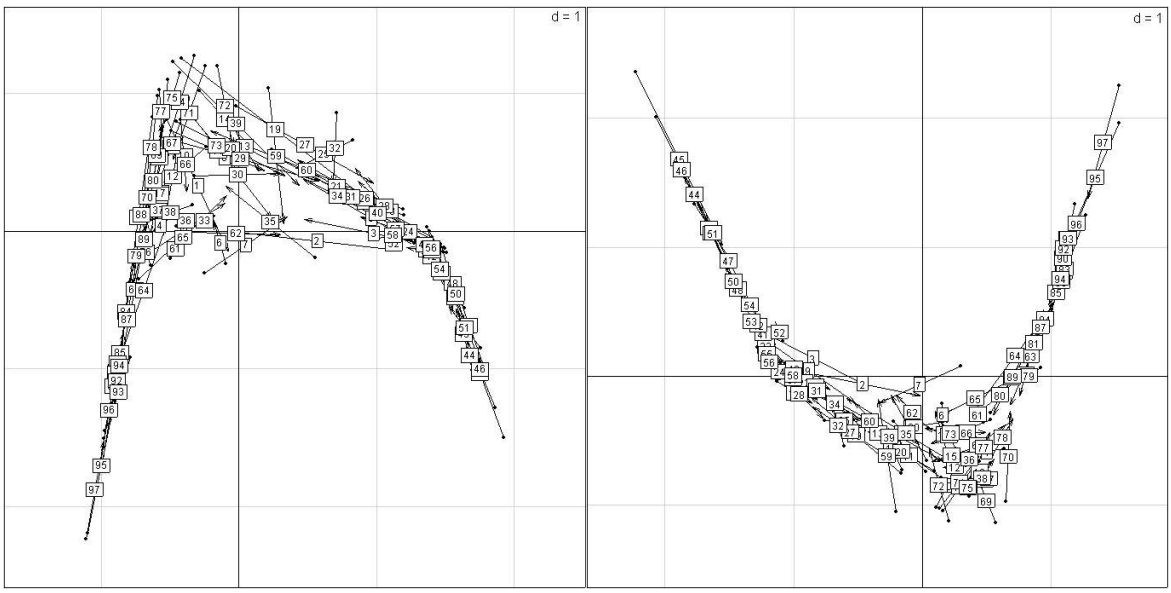
Représenter la projection des trois premiers axes de l'analyse simple sur le plan des deux premiers axes de l'analyse spatialisée. Globalement le plan 1-2 est largement conservé.



```
s.value(mafragh$xy,-maf.coa$li[,1])
s.value(mafragh$xy,maf.coa$li[,2])
s.value(mafragh$xy,maf.coa.ms$li[,1])
s.value(mafragh$xy,maf.coa.ms$li[,2])
```

Mais la lecture des résultats est simplifiée.

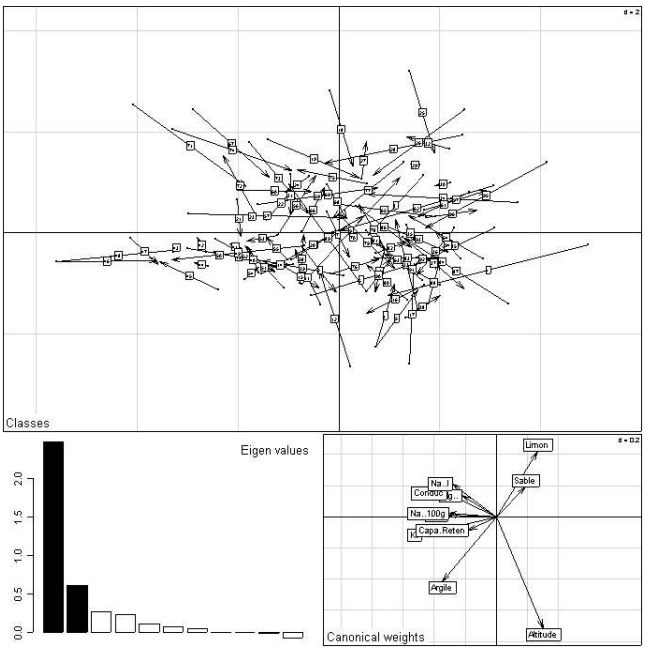
```
w1=maf.coa$li[,1:2]
w1m=apply(w1,2,function(var) lag.listw (x=maf.listw, var))
s.match(w1,w1m,clab=0.75)
w1=maf.coa.ms$li[,1:2]
w1m=apply(w1,2, function(var) lag.listw (x=maf.listw, var))
s.match(w1,w1m,clab=0.75)
```



L'analyse nous apprend que la partie spatialisée de l'interprétation se réduit à deux dimensions. Vérifier que `maf.coa$c1` contient des scores des espèces de moyenne nulle et de variance unité pour la distribution de `maf.coa$cw`. Vérifier que `maf.coa$li` place les

sites à la moyenne des espèces qu'ils contiennent. Vérifier que les mêmes propriétés sont vraies pour **maf.coa.ms$c1** et **maf.coa.ms$li**.

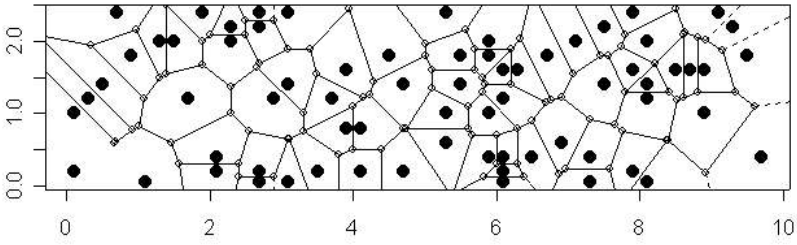
```
plot(maf.pca.ms<-multispati(dudi.pca(mafragh$mil,scan=F),maf.listw))
```



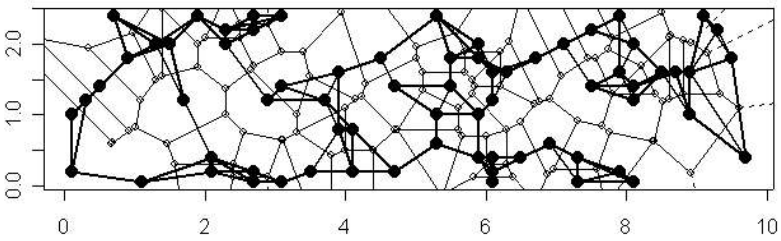
La végétation est beaucoup plus spatialisée que l'analyse de sol.

7.2. Gradients

```
data("oribatid.rda")
names(oribatid)
# [1] "fau"   "envir" "xy"
ori.xy=oribatid$xy[,c(2,1)] # pour avoir les figures dans la largeur de
la page
names(ori.xy)=c("x","y")
plot(ori.xy,pch=20,cex=2,asp=1)
plot(voronoi.mosaic(ori.xy),add=T)
```

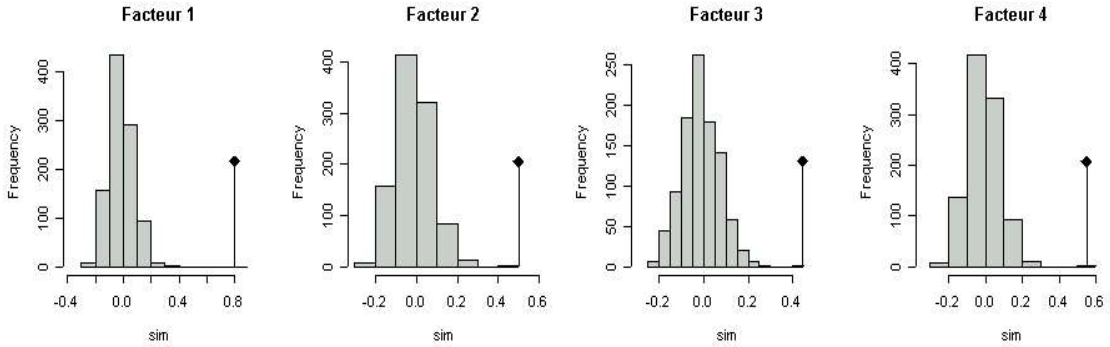



```
s.label(ori.xy,add.p=T,
     neig=nb2neig(knn2nb(knearneigh(as.matrix(ori.xy),3))),clab=0)
```

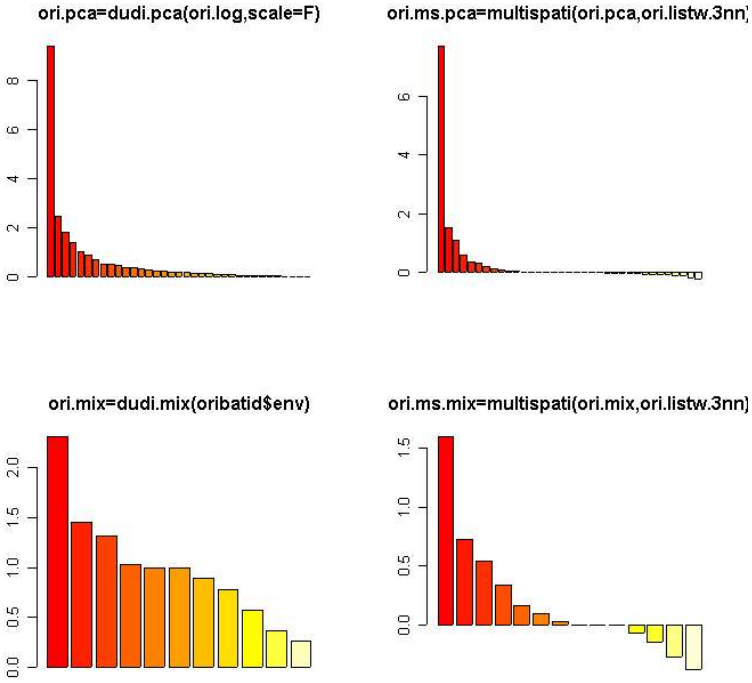


Implanter la relation de voisinage aux 3 plus proches voisins. On s'intéresse aux variations liées au voisinage immédiat :

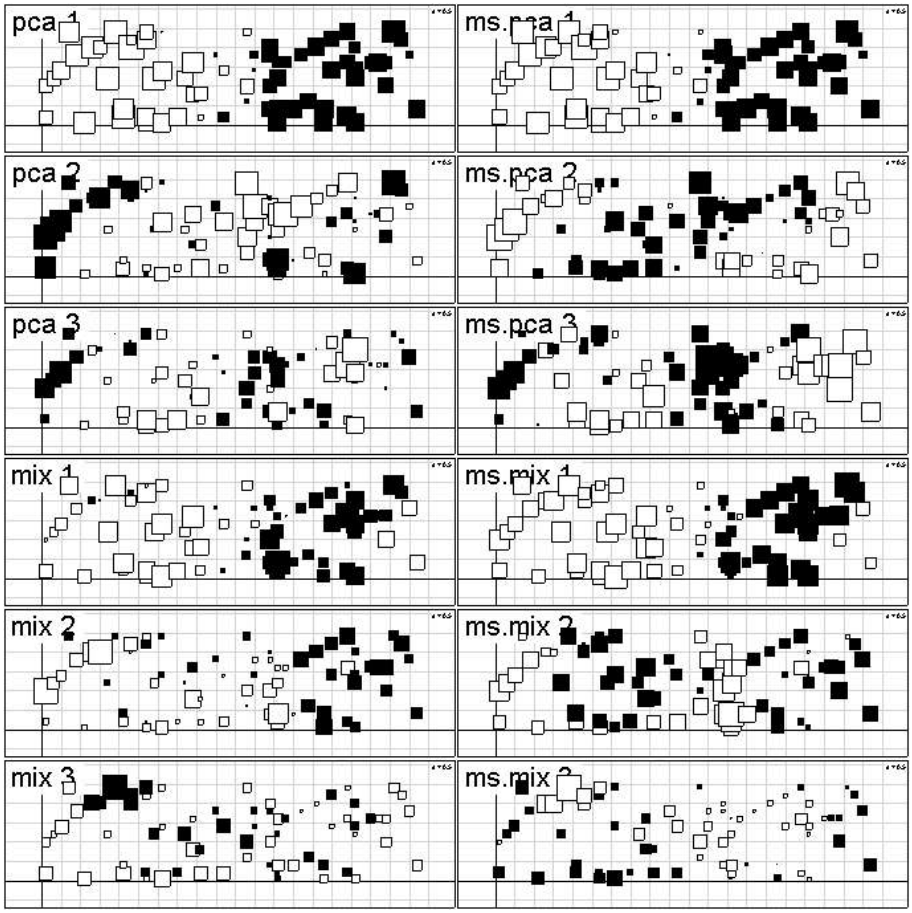
```
ori.listw.3nn=nb2listw(knn2nb(knearneigh(as.matrix(ori.xy),3)))
ori.log=log(oribatid$fau+1)
ori.pca=dudi.pca(ori.log,scale=F)
Select the number of axes: 4

par(mfrow=c(1,4))
for (i in 1:4) {
      w=moran.mc(ori.pca$l1[,i],ori.listw.3nn,999)
      plot(as.randtest(w$res,w$statistic),main=paste('Facteur',i))
}
```



Faire l'analyse du tableau faune. La structure faunistique est-elle spatialisée ? Faire l'analyse du tableau milieu. Les coordonnées ont-elles une structure spatiale ?




```
par(mfcol=c(6,2))
for (i in 1:3) {s.value(ori.xy,ori.pca$l1[,i],sub=paste("pca",i),csub=3,cleg=0)}
for (i in 1:3) {s.value(ori.xy,ori.mix$l1[,i],sub=paste("mix",i),csub=3,cleg=0)}
for (i in 1:3) {s.value(ori.xy,ori.ms.pca$li[,i],sub=paste("ms.pca",i),csub=3,cleg=0)}
for (i in 1:3) {s.value(ori.xy,ori.ms.mix$li[,i],sub=paste("ms.mix",i),csub=3,cleg=0)}
```



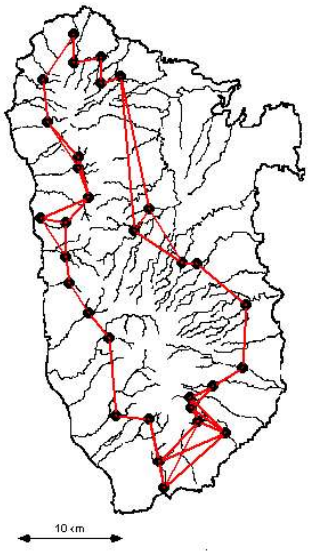
Caractériser les gains des analyses sous contrainte.

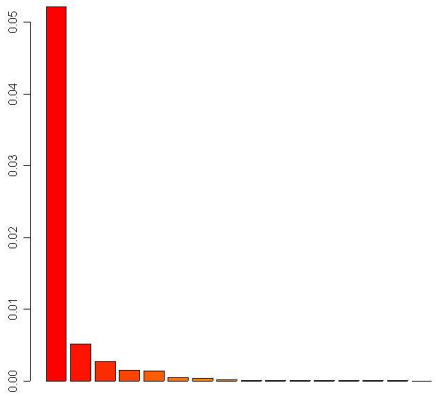
Le lien très fort des deux tableaux avec l'espace est une contrainte énorme qui cache peut-être les relations espèces-environnement ayant un intérêt écologique. D'où les travaux qui vise à débarrasser les données des composantes spatiales (Borcard et al. 1992, Borcard and Legendre 1994, Méot et al. 1998).

De manière générale, la contrainte simplifie l'interprétation quand on la cherche dans l'espace.

7.3. Variations locales

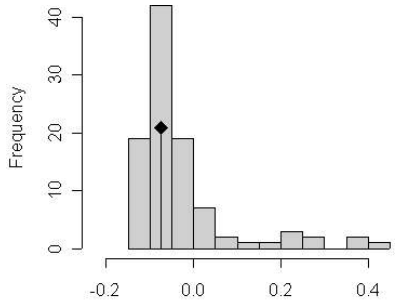
```
data(atya)
at.xy = atya$xy
at.gen = atya$gen
at.pnm =  read.pnm(system.file("pictures/atyacarto
            package = "ade4"))
atya$carto.pnm
at.nb = neig2nb(atya$neig)
at.listw=nb2listw(at.nb)
plot(at.pnm)
points(at.xy, pch=20,cex=2)
plot(at.nb, at.xy, col="red", add=T, lwd=2)
```



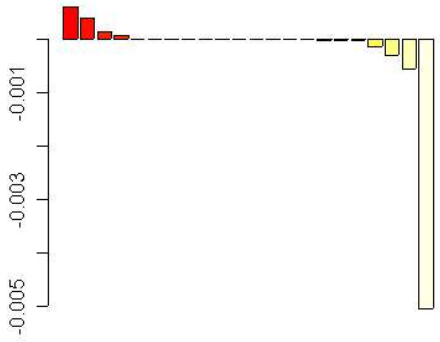



```
at.pca=dudi.pca(at.gen,scale=F)
Select the number of axes: 2
```

```
plot(multispati.randtest(at.pca,at.listw))
```




```
at.ms=multispati(at.pca,at.listw)
Select the first number of axes (>=1): 1
Select the second number of axes (>=0): 1
```



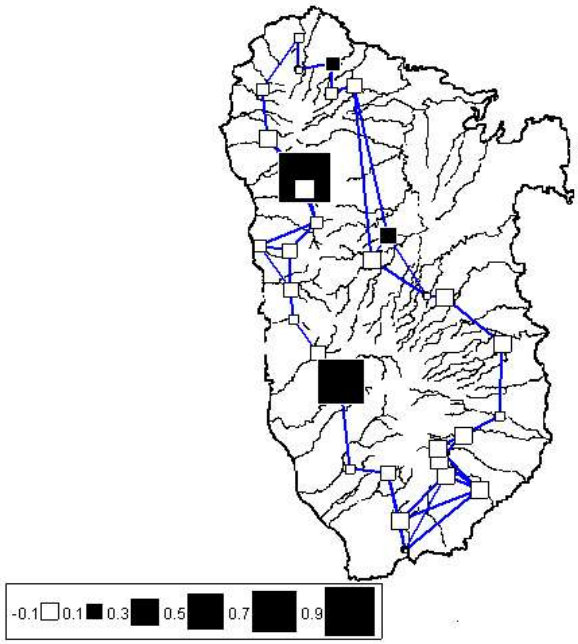
Des valeurs propres négatives qui ne se cachent pas !

```
summary(at.ms)
```

```
Multivariate Spatial Analysis
Call: multispati(dudi = at.pca, listw = at.listw)

Scores from the first duality diagramm:
         var     cum  ratio     moran
RS1 0.052108 0.05211 0.8133 -0.092958
RS2 0.005177 0.05728 0.8941  0.009872

Eigenvalues decomposition:
           eig      var   moran
CS1   0.0006051 0.002371  0.2552
CS22 -0.0050308 0.050243 -0.1001
```

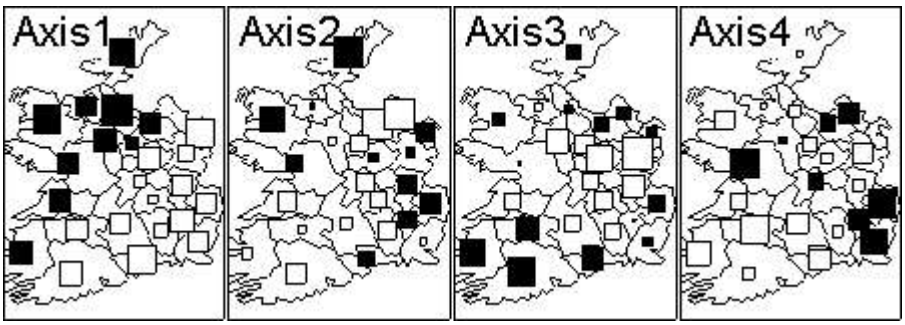
la totalité de information est locale

```
plot(at.pnm)
plot(at.nb, at.xy,col="blue",add=T,lwd=2)
s.value(at.xy,at.ms$li[,2],add.p=T,csi=1)
```



7.4. Cartes factorielles et cartes spatiales

```
data(irishdata)
names(irishdata)
 [1] "area"          "county.names" "xy"           "tab"          "contour"
 [6] "link"          "area.utm"     "xy.utm"       "link.utm"     "tab.utm"
[11] "contour.utm"

ir.xy=irishdata$xy.utm
ir.pca=dudi.pca(irishdata$tab)
Select the number of axes: 4
```

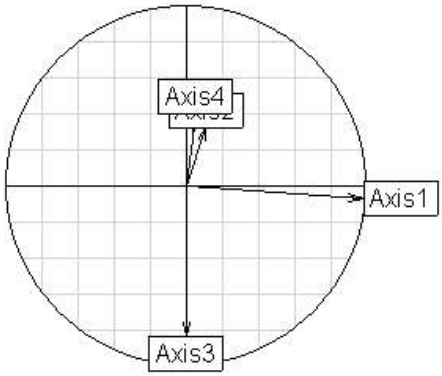



```
lapply(ir.pca$li,moran.mc,listw=pond.ext.w,nsim=5000)
$Axis1        statistic = 0.567, observed rank = 5001, p-value = 0.0002
$Axis2        statistic = 0.1754, observed rank = 4672, p-value = 0.066
$Axis3        statistic = 0.4677, observed rank = 4998, p-value = 0.0006
$Axis4        statistic = 0.2716, observed rank = 4916, p-value = 0.017
```

La première et la troisième composante sont très structurées dans l'espace concret, la quatrième beaucoup moins, la seconde à peine.

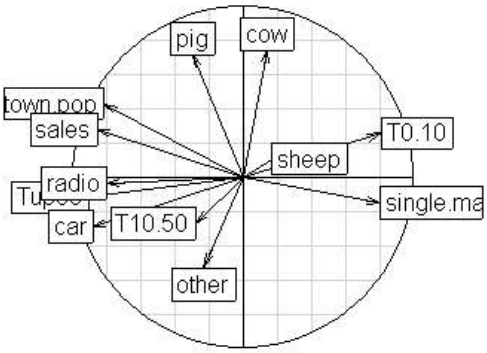
```
ir.nb=neig2nb(neig(mat01=1*(irishdata$link.utm>0)))
ir.listw=nb2listw(ir.nb,apply(irishdata$link.utm,1,function(x) x[x>0]))
ir.ms=multispati(ir.pca,ir.listw)
```

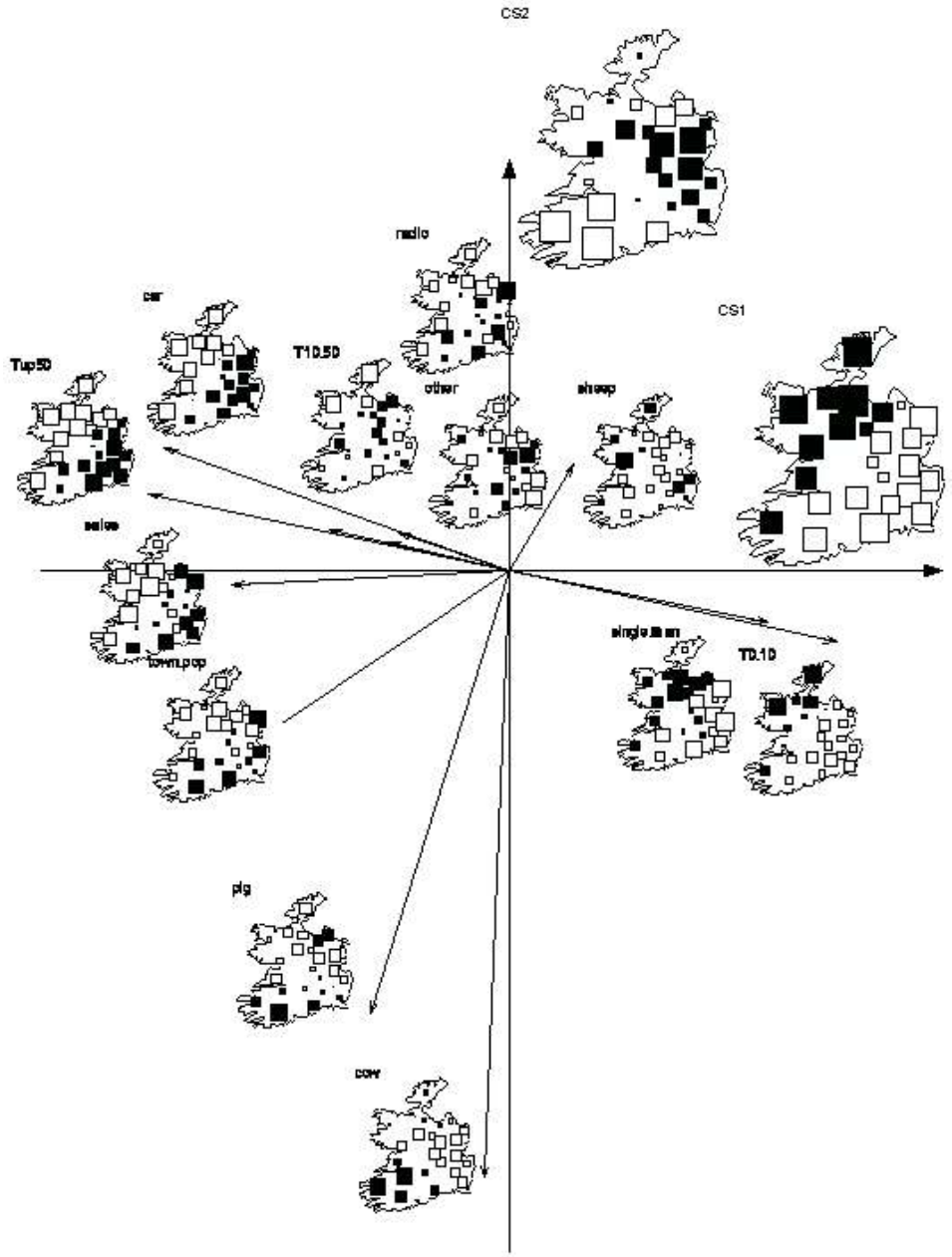
```
Select the first number of axes (>=1): 2
Select the second number of axes (>=0): 0
summary(ir.ms)

Multivariate Spatial Analysis
Call: multispati(dudi = ir.pca, listw = ir.listw)

Scores from the first duality diagramm:
      var    cum  ratio  moran
RS1 5.186  5.186 0.4322 0.5670
RS2 1.922  7.108 0.5923 0.1754
RS3 1.877  8.985 0.7488 0.4677
RS4 1.228 10.213 0.8511 0.2716

Eigenvalues decomposition:
      eig   var  moran
CS1 3.074 5.002 0.6146
CS2 1.156 1.748 0.6614

s.corcircle(ir.ms$as)
```



Le plan 1-3 de l'analyse simple est le plan 1-2 de l'analyse spatiale.

```
s.corcircle(ir.pca$co,1,3)
```





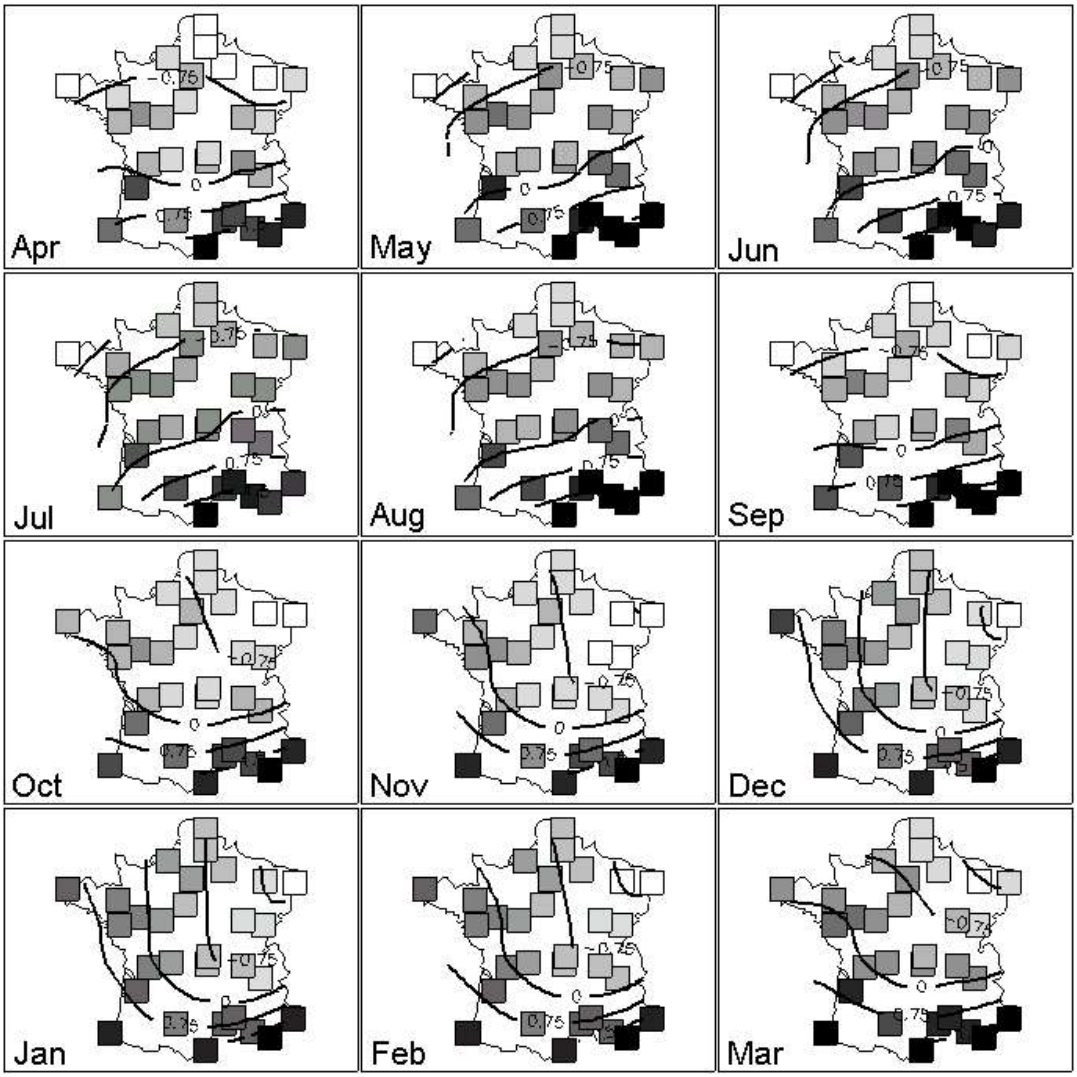
Les flèches sont positionnées par `ir.ms$c1` : coefficients des combinaisons linéaires des variables donnent les composantes cartographibles (`ir.ms$li`). Ceci forme une sorte de typologie de cartes exactement comme le cercle des corrélations forment une typologie de variables.

7.5. Une information exclusivement spatiale

```
data(t3012)
```

Donner une représentation de l'information, par exemple :

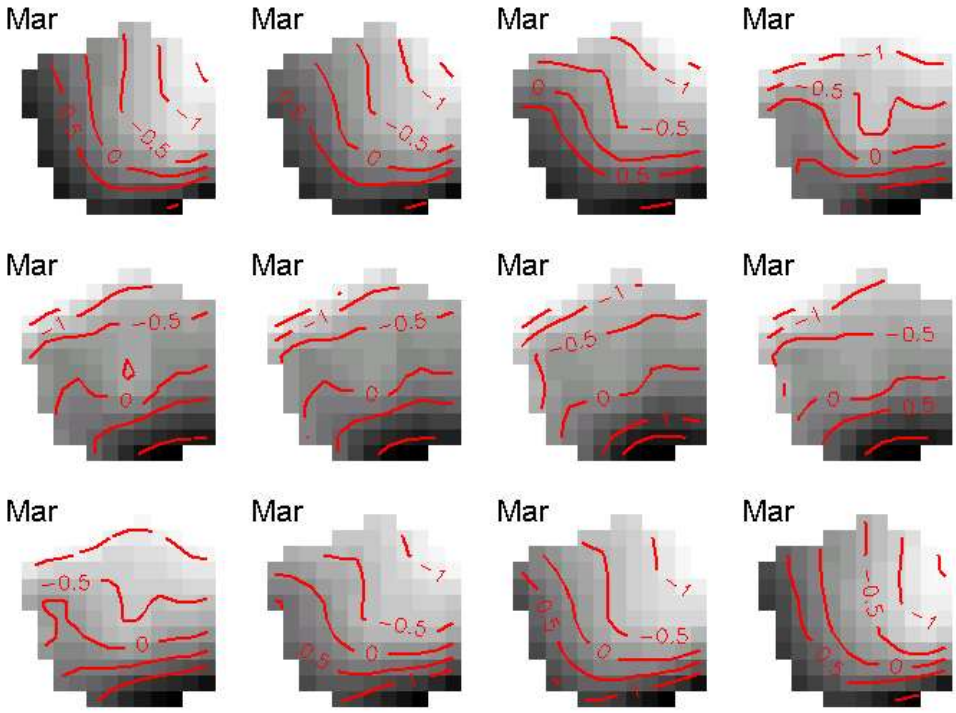
```
par(mfrow=c(4,3))
for (j in c(4:12,1:3)) {
      cdn(t3012$xy,scalewt(t3012$temp[,j]),
            t3012$contour,0.7,names(t3012$temp)[j])}
```




```
cdn <- function(xy,z,poly4, span=0.25, sub){
    par(mar=c(0,0,0,0))
    if (!require(splancs)) stop ("splancs required for inout")
    if (!require(modreg)) stop ("modreg required for loess")
    w = cbind.data.frame(xy,z)
    lo=loess(z~x+y,data=w,span=span)
    xg = seq(min(xy[,1]),max(xy[,1]),le=nrow(xy))
    yg = seq(min(xy[,2]),max(xy[,2]),le=nrow(xy))
    gr=expand.grid(xg, yg)
    names(gr)=names(xy)
    polyin = xy[chull(xy),]
    grin = inpip(gr,polyin)
    mod = predict(lo,newdata=gr)
    mod[-grin] = NA
    mod = matrix(mod,nrow(xy),nrow(xy))
    s.label(xy,include.ori=F, addax=F, cpoi=1, clab=0,
       contour=poly4,cgrid=0,grid=F,sub=sub,csub=3)
    s.value(xy,z,add.p=T,cleg=0, meth="greylevel")
    contour(xg,yg,mod,add=T,labcex=0.75,lwd=2,nlevels=5,
       levels=c(-1.5,-0.75,0,0.75,1.5))
}
```

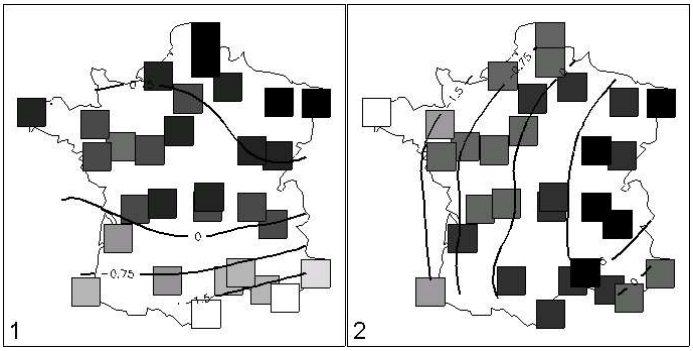
Ou encore :

```
par(mfrow = c(4,4))
for(k in 1:12)  s.image(t3012$xy,scalewt(t3012$temp[,k]),
kgrid = 3,sub=names(t3012$temp)[j],csub=3)
         par(mfrow = c(1,1))
```



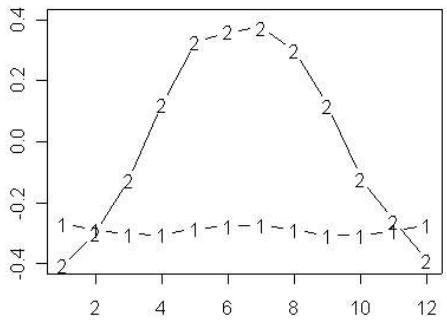
Commenter.

Faire l'ACP normée du tableau et cartographier les deux premières coordonnées :



Dépouiller :

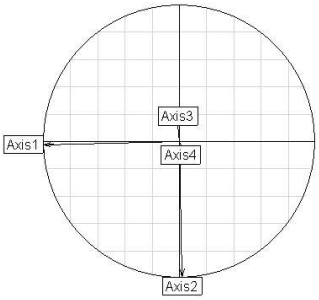
```
plot(1:12,t3012.pca$c1[,1],ylim=c(-0.4,0.4),type="b",pch="1")
points(1:12,t3012.pca$c1[,2],type="b",pch="2")
```



Choisir une pondération de voisinage, par exemple :

```
w1 = nb2listw(knn2nb(knearneigh(as.matrix(t3012$xy),3)))
```

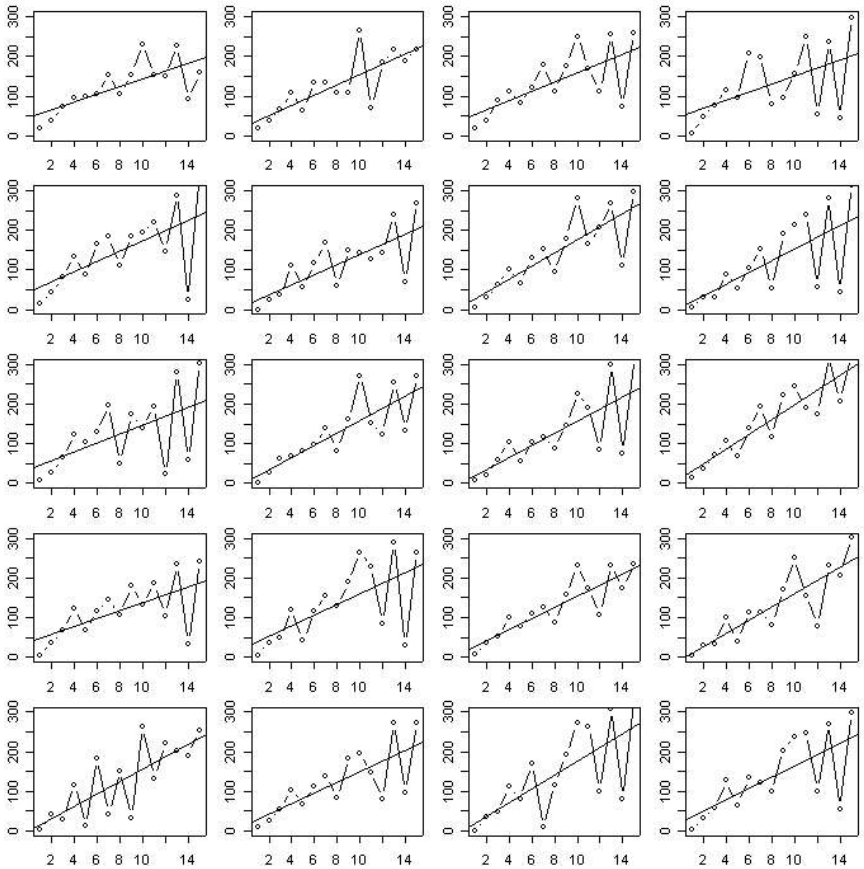
Faire l'ACP sous contrainte de voisinage et comparer les deux :



Commenter ce qu'on a gagné !

7.6. Croissance et alternance, global et local

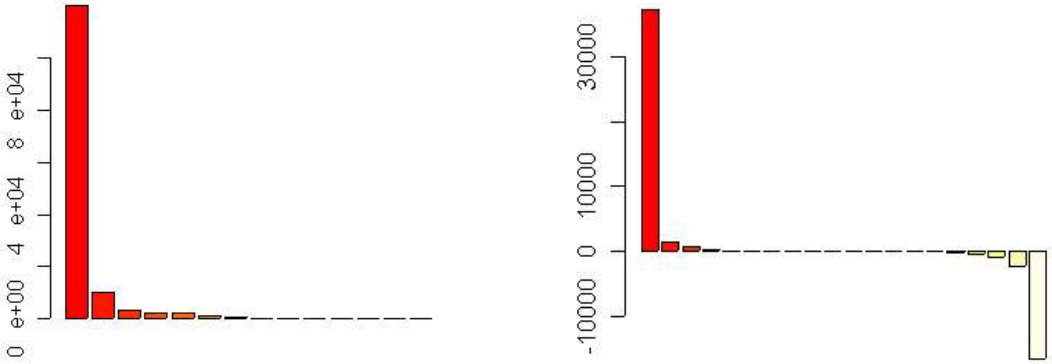
```
data(clementines) # voir la documentation

op <- par(no.readonly = TRUE)
par(mfrow = c(5,4)) ; par(mar = c(2,2,1,1))
w0 <- 1:15
for (i in 1:20) {
      plot(w0, clementines[,i], type = "b",ylim=c(0,300))
      abline(lm(clementines[,i] ~ w0))
}
par(op)
```



Production de 20 arbres fruitiers pendant 15 ans

```
clem.pca=dudi.pca(clementines,scale=F,scannf=F)
```

```
clem.neig=neig(n.line=15)
clem.nb=neig2nb(clem.neig)
clem.listw=nb2listw(clem.nb)
clem.ms=multispati(clem.pca,clem.listw)
Select the first number of axes (>=1): 2
Select the second number of axes (>=0): 2
```

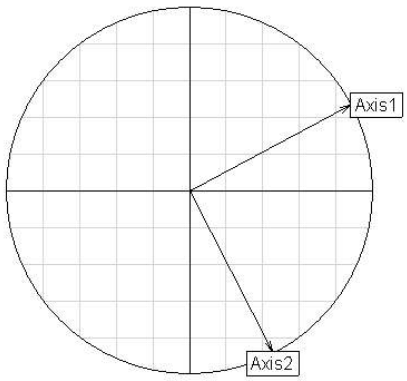


A gauche les valeurs propres de l'ACP, à droite celle de l'ACP sous contraintes spatiales.

```
summary(clem.ms)


Multivariate Spatial Analysis
Call: multispati(dudi = clem.pca, listw = clem.listw)

Scores from the first duality diagramm:
       var    cum  ratio   moran
RS1 119761 119761 0.8510  0.2096
RS2  10059 129821 0.9225 -0.5008

Eigenvalues decomposition:
        eig   var   moran
CS1   37104 94738  0.3916
CS2    1475  2157  0.6838 # on peut faire mieux (1) mais ce n'est pas collectif
CS19  -2285  3637 -0.6284 # on peut faire mieux (20) mais ce n'est pas collectif
CS20 -16496 33789 -0.4882
```

Le plan des deux premiers axes de l'ACP simple maximisent l'inertie projetée. Le total est 129821 soit 92.25%. Il ne peut être dépassé par le plan 1-20 de l'ACP spatiale (94738 + 33789 = 128527) qui donne cependant un résultat très proche.

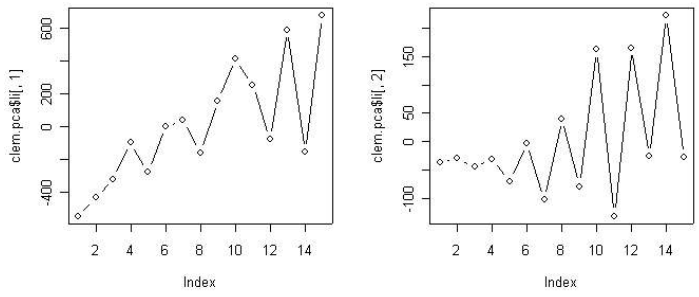
```
s.corcircle(clem.ms$as,1,4)
```

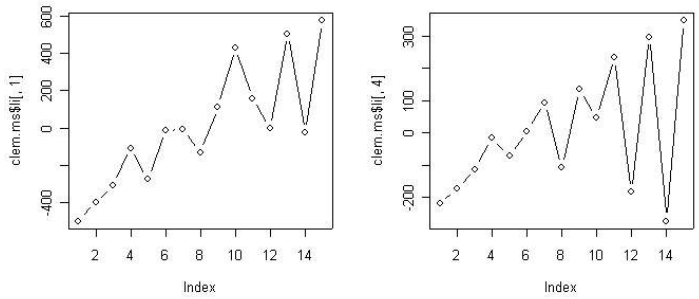


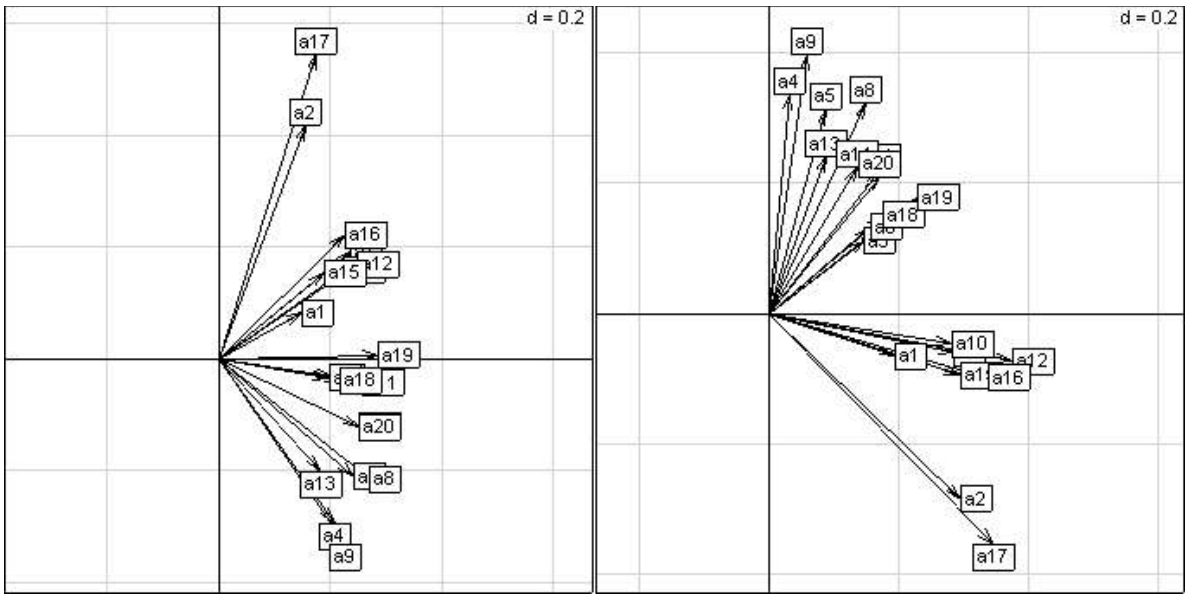
Les deux plans sont les mêmes.

```
par(mfrow=c(2,2))
plot(clem.pca$li[,1],type="b")
```

```
plot(clem.pca$li[,2],type="b")
plot(clem.ms$li[,1],type="b")
plot(clem.ms$li[,4],type="b")
```



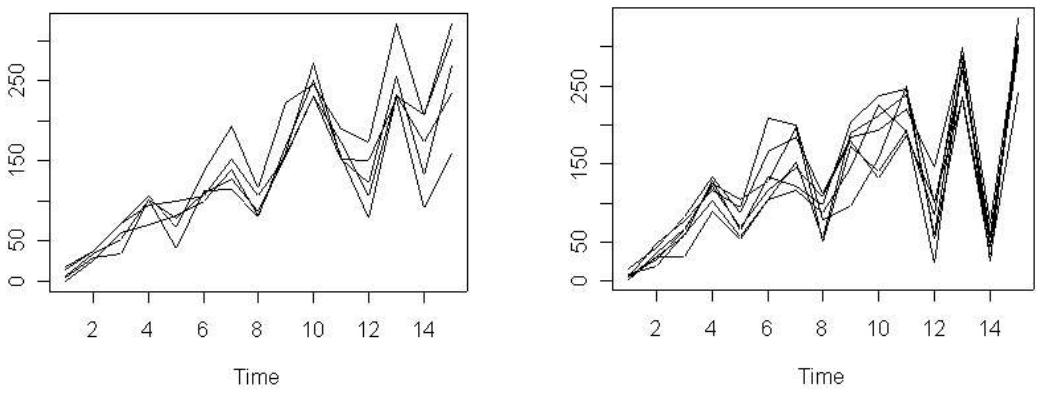




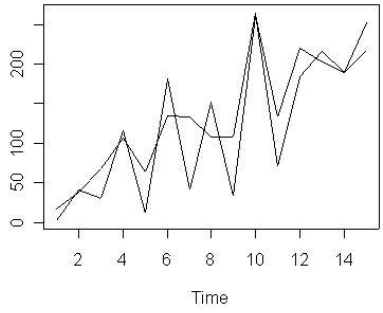

```
s.arrow(clem.pca$c1)
s.arrow(clem.ms$c1,1,4)
```

La croissance et l'alternance sont deux composantes de la variabilité. L'analyse spatiale les sépare clairement et identifie le groupe des croissances les plus régulières (1,10,12,15 et 16) et des alternances les plus marquées (4,9,5,13,8,11,20,13).

```
library(ts)
ts.plot(clementines[,c(1,10,12,15,16)])
ts.plot(clementines[,c(4,9,5,13,8,11,20,13)])
```




```
ts.plot(clementines[,c(2,17)])
```



L'analyse sous contrainte spatiale reste une analyse d'inertie fortement orientée dans l'interprétation vers la lecture de *l'autocorrélation*.

8. Références

Anselin, L. 1995. Local indicators of spatial association. Geographical Analysis 27:93-115.

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

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

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

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

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

Banet, T. A., and L. Lebart. 1984. Local and Partial Principal Component Analysis (PCA) and Correspondence Analysis (CA). Pages 113-123 in I. A. f. S. Computing., editor. COMPSTAT 84. Physica-Verlag, Vienna.

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

Benali, H., and B. Escofier. 1990. Analyse factorielle lissée et analyse factorielle des différences locales. Revue de Statistique Appliquée 38:55-76.

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

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

Borcard, D., P. Legendre, and P. Drapeau. 1992. Partialling out the spatial component of ecological variation. Ecology 73:1045-1055.

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

Chessel, D., and P. Donadieu. 1977. Introduction à l'étude de la structure horizontale en milieu steppique. III Dispersion locale, densité et niveaux d'implantation chez les ligneux bas. Œcologia Plantarum 12:221-224.

Chessel, D., and R. Sabatier. 1993. Couplage de triplets statistiques et graphes de voisinage. Pages 28-37 in B. C. Asselain, editor. Biométrie et Données spatio-temporelles. Société Française de Biométrie, ENSA, Rennes.

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

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

Conradsen, K., B. K. Nielsen, and T. A. Thyrsted. 1985. Comparison of min/max autocorrelation factor analysis and ordinary factor analysis. in Proceedings from Symposium in Applied Statistics. Technical University of Denmark, Lyngby, Denmark.

Cornillon, P.-A., P. Amenta, and R. Sabatier. 1999. Three-way data arrays with double neighbourhood relations as a tool to analyze a contiguity structure. Pages 263-270 in M. Vichi and O. Opitz, editors. Classification and data analysis. Theory and Application. Springer-Verlag, Berlin.

Dale, M. R. T., P. Dixon, M. J. Fortin, P. Legendre, D. Myers, and M. Rosenberg. 2002. Conceptual and mathematical relationships among methods for spatial analysis. ecography 25:558-577.

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

de Belair, G., and M. Bencheikh-Lehocine. 1987. Composition et déterminisme de la végétation d'une plaine côtière marécageuse : La Mafragh (Annaba, Algérie). Bulletin d'Ecologie 18:393-407.

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

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

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

Dungan, J. L., J. Perry, M. R. T. Dale, S. Citron-Pousty, M. J. Fortin, A. Jakomulska, A. Legendre, M. Miriti, and M. S. Rosenberg. 2002. A balanced view of scaling in spatial statistical analysis. Ecography 25:626–640.

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

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

Escoufier, Y. 1987. The duality diagramm : a means of better practical applications. Pages 139-156 in P. Legendre and L. Legendre, editors. Development in numerical ecology. NATO advanced Institute , Serie G .Springer Verlag, Berlin.

Estève, J. 1978. Les méthodes d'ordination : éléments pour une discussion. Pages 223-250 in J. M. Legay and R. Tomassone, editors. Biométrie et Ecologie. Société Française de Biométrie, Paris.

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

Flesche, H., A. A. Nielsen, and R. Larsen. 2000. Supervised mineral classification with semiautomatic training and validation set generation in scanning electron microscope energy dispersive spectroscopy images of thin sections. Mathematical Geology 32:337-366.

Gabriel, K. R., and R. R. Sokal. 1969. A new statistical approach to geographic variation analysis. Systematic Zoology 18:259-278.

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

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

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

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

Goulard, M., M. Voltz, and P. Monestiez. 1987. Comparaison d'approches multivariables pour l'étude de la variabilité spatiale des sols. Agronomie 7:657-665.

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

Grande, J. A., J. Borrego, and J. Morales. 2000. Study of heavy metal pollution in the Tinto-Odiel estuary in SW of Spain using factor analysis. Environmental Geology 39:1095-1101.

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

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

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

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

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

Grunsky, E. C., Q. Chen, and F. P. Agterberg. 1996. Applications of spatial factor analysis to multivariate data. Pages 229-261 in A. Foerster and D. F. Merriams, editors. Geologic Modeling and Mapping. Plenum, New York.

Hatheway, W. H. 1971. Contingency table analysis of rain forest vegetation. Pages 271-314 in G. P. Patil, E. C. Pielou, and W. E. Waters, editors. Statistical Ecology. III Many species populations ecosystems and systems analysis. Pennsylvania State University Press.

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

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

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

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

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

Kuo, C.-H. 2002. IGD, Among-individual level genetic distance (relatedness) calculation. Department of Zoology and Genetics, owa State University.

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

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

Lebart, L. 1984. Correspondence analysis of graph structure. Bulletin technique du CESIA, Paris 2:5-19.

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

Liebhold, A. M., and J. Gurevitch. 2002. Integrating the statistical analysis of spatial data in ecology. ecography 25:553-557.

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

Méot, A., D. Chessel, and R. Sabatier. 1993. Opérateurs de voisinage et analyse des données spatio-temporelles. Pages 45-72 in B. Asselain, editor. Biométrie et Environnement. Masson, Paris.

Méot, A., P. Legendre, and D. Borcard. 1998. Partialling out the spatial component of ecological variation: questions and propositions in the linear modelling framework. Environmental and Ecological Statistics 5:1-27.

Mom, A. 1988. Méthodologie statistique de la classification des réseaux de transports. thèse de doctorat, USTL, Montpellier.

Monestiez, P. 1978. Présentation de deux méthodes utilisant la notion de contiguïté pour l'analyse des données géographiques. Thèse de Docteur-Ingénieur, Paris VI.

Monestiez, P., M. Goulard, and G. Charmet. 1994. Geostatistics for spaial genetic structures: study of wild populations of perennial ryegrass. Theoretical and Applied Genetics 88:33-41.

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

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

Nielsen, A. A. 1995a. Change detection in multi-spectral, bi-temporal spatial data using orthogonal transformations. in .

Nielsen, A. A. 1995b. Multi-channel remote sensing data and orthogonal transformations for change detection. in .

Nielsen, A. A. 1999. C04351 Statistical Image Analysis, Spring 1999 Orthogonal Transformations. in .

Nielsen, A. A., and K. Conradsen. 1997. Multivariate alteration detection (MAD) in multispectral, bi-temporal image data: a new approach to change detection studies. in . Tech. rep. 199711, Department of Mathematical Modelling, Technical University of Denmark.

Nielsen, A. A., K. Conradsen, J. L. Pedersen, and A. Steenfelt. 1997. Spatial factor analysis of stream sediment geochemistry data from South Greenland. Pages 955-960 in V. Pawlowsky-Glahn, editor. Proceedings of ther Third Annual Conference of the International Association for Mathematical Geology, Barcelona, Spain.

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

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

Pace, R. K., and R. Barry. 1997. Sparse spatial autoregressions. Statistics and Probability Letters 33:291-297.

Pace, R. K., R. Barry, V. C. Slawson, and C. F. Sirmans. 2003. Simultaneous spatial and functional form transformations. Pages in R. Florax and L. Anselin, editors. Advances in Spatial Econometrics.

Pace, R. K., and J. P. LeSage. 2002. Semiparametric maximum likelihood estimates of spatial dependance. Geographical Analysis 34:76-90.

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

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

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

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

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

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

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

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

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

Wackernagel, H. 2003. Multivariate Geostatistics : An introduction with applications, Springer Verlag edition.

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

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

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