Analyse statistique des données spatiales I

Christine Thomas-Agnan

Toulouse School of Economics

30 octobre 2012

Données spatiales

- Données spatiales ou géoréférencées : données pour lesquelles une information géographique est attachée à chaque unité statistique. L'information géographique est en général la position de l'unité sur une carte ou dans un référentiel spatio-temporel, et peut par exemple prendre la forme de latitude et longitude ou de coordonnées UTM.

Données spatiales

- Données spatiales ou géoréférencées : données pour lesquelles une information géographique est attachée à chaque unité statistique. L'information géographique est en général la position de l'unité sur une carte ou dans un référentiel spatio-temporel, et peut par exemple prendre la forme de latitude et longitude ou de coordonnées UTM.
- Nécessité de faire interagir analyse statistique et cartographie

Données spatiales

- Données spatiales ou géoréférencées : données pour lesquelles une information géographique est attachée à chaque unité statistique. L'information géographique est en général la position de l'unité sur une carte ou dans un référentiel spatio-temporel, et peut par exemple prendre la forme de latitude et longitude ou de coordonnées UTM.
- Nécessité de faire interagir analyse statistique et cartographie
- Un traitement statistique de telles données qui ignorerait cet aspect ou l'intègrerait de façon inadéquate resulterait en une perte d'information, des erreurs de spécifications, des estimations non convergentes et non efficaces, des erreurs de prédiction.

Données spatiales

- Données spatiales ou géoréférencées : données pour lesquelles une information géographique est attachée à chaque unité statistique. L'information géographique est en général la position de l'unité sur une carte ou dans un référentiel spatio-temporel, et peut par exemple prendre la forme de latitude et longitude ou de coordonnées UTM.
- Nécessité de faire interagir analyse statistique et cartographie
- Un traitement statistique de telles données qui ignorerait cet aspect ou l'intègrerait de façon inadéquate resulterait en une perte d'information, des erreurs de spécifications, des estimations non convergentes et non efficaces, des erreurs de prédiction.

Divers courants

La statistique spatiale rassemble divers courants (géostatistique, économétrie spatiale, semis de points)

- données de nature différente
- problématiques et outils spécifiques
- mais des points communs

Domaines d'application

Domaines scientifiques privilégiés d'application de la statistique spatiale

- la géologie
- la séismologie
- la météorologie
- l'économie
- la géographie
- l'épidémiologie
- secteur industriel : l'industrie pétrolière
- secteur tertiaire : géomarketing

Domaines d'application

Exemple en prospection pétrolière : prédire la quantité de pétrole potentielle en un lieu donné en fonction de prélèvements effectués en certains points répartis sur une zone pour optimiser l'emplacement des forages.

Domaines d'application

Exemple en prospection pétrolière : prédire la quantité de pétrole potentielle en un lieu donné en fonction de prélèvements effectués en certains points répartis sur une zone pour optimiser l'emplacement des forages.

Exemple en économie urbaine : l'ajustement de modèles hédoniques qui expliquent le prix d'une transaction en fonction des caractéristiques du bien immobilier mais aussi des caractéristiques socio-économiques ou autres de leur lieu d'implantation permet de mieux comprendre ce qui influence le marché immobilier et de proposer des modèles pour créer des indices de prix.

Domaines d'application

Exemple en environnement : la production de cartes de prédictions de niveaux de pollution utilise les outils de la géostatistique.

Domaines d'application

Exemple en environnement : la production de cartes de prédictions de niveaux de pollution utilise les outils de la géostatistique.

Exemple en hydrologie : la géostatistique permet de distinguer entre les changements de la qualité de l'eau dus à des sources locales de pollution et ceux dus à la diversité régionale des propriétés géologiques des nappes phréatiques.

Domaines d'application

Exemple en épidémiologie : produire des cartes de niveau de risque lors d'une épidémie

Domaines d'application

Exemple en épidémiologie : produire des cartes de niveau de risque lors d'une épidémie

Exemple en géomarketing : définir des zones de chalandise, prédire les flux de clients d'une zone géographiqe donnée vers un magasin donné.

Champ aléatoire

Pour une localisation s, une caractéristique X_s est mesurée : on la considère comme une réalisation $X(s, \omega)$ d'une variable aléatoire X_s. Le champ aléatoire $X(s, \omega)$ est l'objet mathématique qui permet de modéliser ces observations.

L'indice s varie dans une partie $\mathcal{D}$ de $\mathbb{R}^d$. La dimension d varie de 1 à 3 dans les applications courantes.

On imagine donc que, pour un lieu s donné, il existe un univers de réalisations possibles (pour chaque ω) de la caractéristique X_s mais dans la réalité on observe généralement **une seule réalisation** de X_s et pour un **nombre fini de sites** s. Pluralité de données due à une pluralité de lieux mais non à une pluralité de réalisations sauf si on est dans le cas d'observations répétées.

Séries temporelles et champs aléatoires

Le champ aléatoire pour $d = 1$ correspond à la série temporelle, alors que pour $d = 2$, il correspond au champ spatial. Mais les méthodes de séries temporelles ne se résument pas à un cas particulier de la statistique spatiale. Inversement, la statistique spatiale n'est pas une simple généralisation des séries temporelles.
Elles partagent cependant deux caractéristiques : la dépendance et l'hétérogénéité.

La dépendance :

- ce qui se passe aujourd'hui est nécéssairement influencé par ce qui s'est passé hier et dans une moindre mesure par un passé lointain : c'est le phénomène de dépendance temporelle.
- dépendance spatiale : les variables X_s et X_t sont d'autant plus corrélées que la distance entre s et t est petite. On parle d'autocorrélation spatiale.

Séries temporelles et champs aléatoires

L'hétérogénéité :

- dans le cas des séries temporelles, l'hypothèse de répartitions marginales identiques est remise en question dans la mesure où le phénomène peut présenter une évolution en moyenne résultant en une non stationarité.
- de même le champ spatial peut présenter une hétérogénéité spatiale : la répartition marginale de X_s varie avec s.

Mais à la différence des séries temporelles, les notions de passé et de futur n'ont pas leur pendant en spatial et il n'y a pas d'ordre naturel dans $\mathbb{R}^d$.

Avantages modélisation spatiale

Quels sont les avantages d'une modélisation adaptée aux données spatiales ?

- éviter les biais d'estimation des paramètres
- éviter inefficacité
- éviter biais de prédiction
- modéliser les effets de débordements (spillovers)

Illustration : biais et variance d'estimation

Géographie : région Midi-Pyrénées découpée en 283 pseudo-cantons
Voisinage : une unité spatiale est voisine d'une autre si les unités spatiales partagent une frontière commune
On simule X selon $\mathcal{N}(\mu = 40, \sigma = 10)$
On simule Y selon $Y = \rho WY + \beta X + \epsilon$, où ϵ est un bruit blanc spatial et WY désigne le vecteur des moyennes de la variable Y dans le voisinage de chaque unité spatiale

Illustration : biais et variance d'estimation

Le biais d'estimation du coefficient β est donné par
$(X'X)^{-1}X'(I-\rho W)^{-1}X-1$
La différence entre variance estimée dans le modèle OLS et le modèle LAG est donnée par $(X'X)^{-1}X'((I-\rho W)'(I-\rho W))^{-1}X-1$

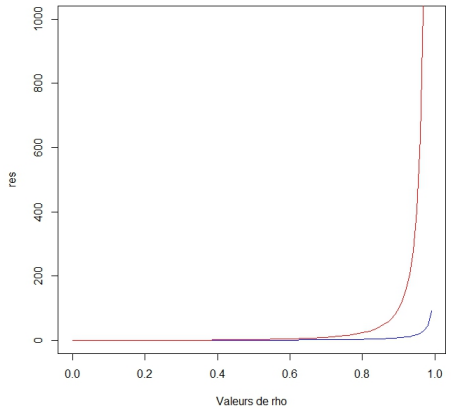
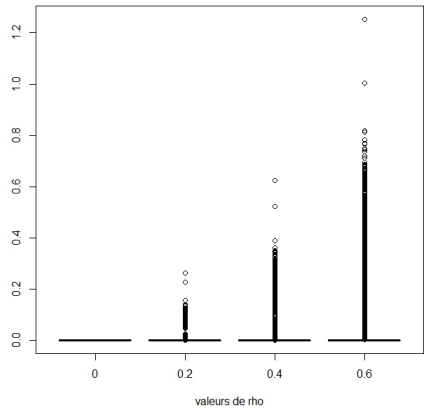


Illustration : hétéroscédasticité

Distribution des éléments de la partie triangulaire supérieure de la matrice de variance de Y dans ce modèle, donnée à facteur d'échelle près par :

$$((I - \rho W)'(I - \rho W))^{-1}$$



Trois grands types

Les grands types de données spatiales sont

- les données ponctuelles ou de type géostatistique
- les données surfaciques ou de type économétrie spatiale
- les données de type semis de points.

Autres types

- images (pixels)
- données bilocalisées ou flux

Données ponctuelles ou de type géostatistique

La position observée est déterministe.
La position varie continuement dans l'espace, même si en pratique on ne l'observe que de façon discrète en des points non nécéssairement sur grille régulière.

Exemples : mesures de pluviométrie en des stations météo, concentration en polluants en des stations de mesure.

Données surfaciques ou de type économétrie spatiale

La position observée est déterministe mais la donnée géographique est de nature surfacique. Les données économiques sont souvent diffusées sur des découpages administratifs d'un territoire.

Exemples : taux de chomage d'une commune, prix moyen des maisons d'un quartier.

Données de type semis de points ou processus ponctuels spatiaux

La position observée est aléatoire et à chaque position peut être attachée (ou non) une ou des caractéristiques appelées marques.

Quelques exemples

- la disposition de certaines espèces végétales dans une forêt,
- les adresses de patients affectés d'une certaine maladie dans une région,
- la répartition de cellules dans un tissu biologique,
- les emplacements des épicentres de secousses sismiques enregistrées,
- la localisation de trésors archéologiques retrouvés sur un site

Les logiciels d'analyse spatiale

- Cartographie : les SIG ou Geographic Information System : ARCINFO, MAPINFO, ARCVIEW (version légère de ARCINFO), SAS/GIS, GEOCONCEPT, CARTE ET BASE, ASTEROP, GRASS
- Liens entre GIS et boites à outils statistiques : SAS avec SAS/GIS, S+, peut être lié à ARCVIEW et à ARCINFO grâce à S+Gislink, SAGE (Haining, Wise, Ma), avec ARCINFO, SPACESTAT (Anselin, Bao)(langage GAUSS), avec ARCVIEW, MANET (Unwin, Hofman), CDV avec TCL/TK (Dykes), XLISP-STAT (Brundson).
- Boite à outils Matlab de spatialeconometrics.com (Le Sage),
- Les packages de R :GeoXp (Toulouse), spdep, geoR, spatstat, etc.

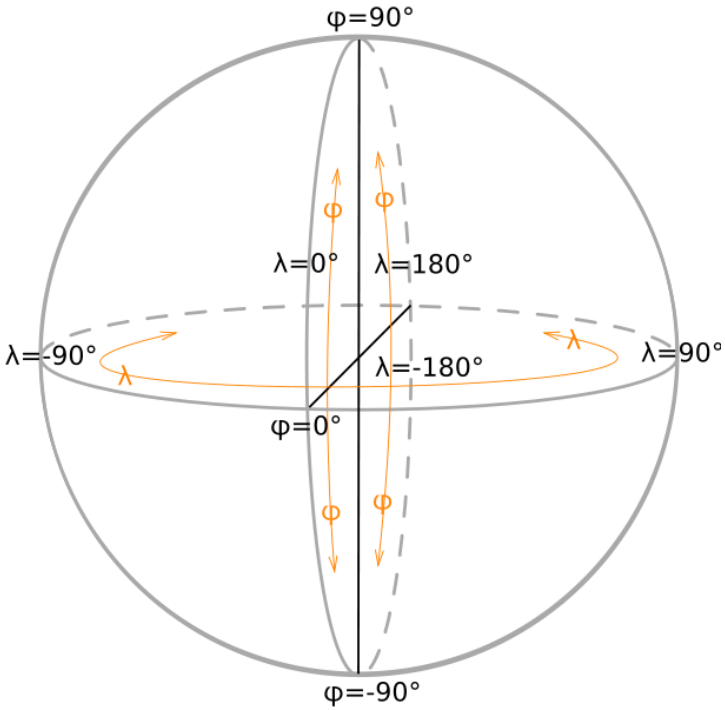
Coordonnées et projections

Pour dessiner une carte il faut

- un système de coordonnées : des axes et une origine
- un système de projection cartographique

Une projection est une correspondance entre les coordonnées planimétriques X et Y d'un point, mesurées sur une grille régulière, et sa latitude ϕ et longitude λ. Au besoin, l'altitude du point est mesurée au dessus (du géoïde ou) du niveau zéro des mers local.

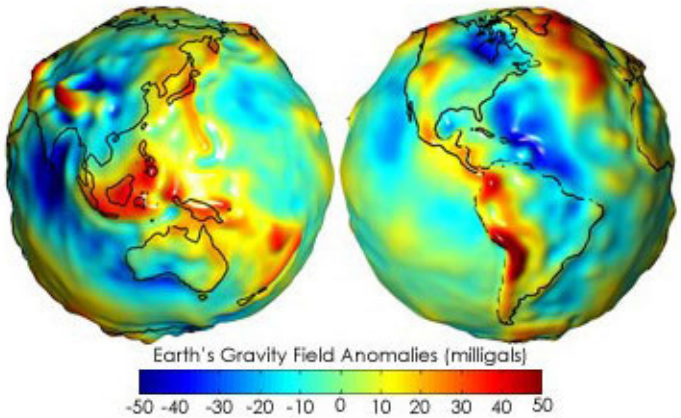
Latitude et longitude



Latitude : mesure de l'angle ϕ par rapport à l'équateur.
Longitude : mesure de l'angle λ par rapport au méridien de référence.
Différentes unités : degrés-minutes-secondes, degrès-décimaux, radians, grades.

Datum géodésique

La surface réelle de la terre est patatoïde ou géoïde ; on l'approxime par un ellipsoïde. Exemple : ellipsoïde de Clarke.
La donnée de cet ellipsoïde et de la projection constitue ce que l'on appelle un "datum géodésique"ou CRS (coordinate reference system). Les coordonnées d'un point sont mesurées sur l' ellipsoïde de révolution de référence, l'altitude du point est égale à la hauteur au dessus de cet ellipsoïde, ses coordonnées planimétriques sont sa latitude et sa longitude.



Datum géodésique

Il faut connaître les datum les plus classiques :

- European Datum (ED) 50 : système européen unifié, avec comme projection courante la projection UTM.
- World Geodetic System (WGS84) : système mondial mis au point par le Département de la Défense des Etats Unis et utilisé par le GPS, avec comme projection courante la projection UTM.

Conversions

Il est courant que l'on récupère des données géoréférencées dans un certain système alors que le fond de carte dont on dispose est codé dans un autre système. Il faut alors recourir à un convertisseur, par exemple Convers.

`http://vtopo.free.fr/convers.htm`

ou le package **p**roj4 de R.

Divers types de projection

La représentation de la surface terrestre sur un plan (la feuille de papier) nécessite la définition d'une projection. La projection est la méthode de réduction de la distorsion due à la rotondité de la terre appliquée sur une surface plate. On distingue plusieurs sortes de projections

- conique : le sommet du cône est dans l'axe des pôles et la tangence avec la terre se fait suivant un parallèle,
- cylindrique : la tangence avec la terre se fait suivant l'équateur,
- azimutale : la projection se fait sur un plan tangent en un point ou sécant en un cercle.

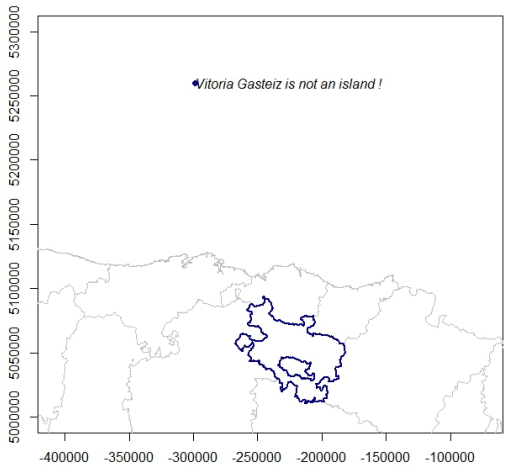
Divers types de projection

Les projections les plus courantes sont :

- la projections de Mercator : projection cylindrique (cylindre tangent à la terre le long de l'équateur), utilisation limitée à des latitudes inférieures a 70°.
- la projection de Mercator Transverse Universelle : projection cylindrique (cylindre tangent à la terre le long d'un méridien choisi), limitée à 3° d'amplitude de part et d'autre du méridien d'origine, pour minimiser les déformations en limite de fuseau. La terre est ainsi divisée en 60 fuseaux de 6°. Utilisée par le GPS.
- les projections Lambert : projections coniques (Lambert I et Lambert I Carto (Nord), Lambert II et Lambert II Carto (Centre), Lambert III et Lambert III Carto (Sud), Lambert IV et Lambert IV Carto (Corse), Lambert Grand Champ, Lambert 93)

Exemple : Vitoria

Si on utilise les contours de la province d'Alava dans l'ellipsoide du WGS84 et avec la projection Lambert Conformal Conic, et simultanément les coordonnées de la ville de Votoria dans l'ellipsoide du WGS84 et avec la projection de Mercator, on obtient



La fonction **p**roj4string du package **m**aptools permet de préciser le CRS.
Le package **p**roj4 permet les conversions d'un système à l'autre.

Diverses classes de données spatiales : package sp

La classe (nature) des objets spatiaux en R (réponse à la question 'class') dépends de la structure initiale des unités spatiales importées ainsi que de package utilisé pour les importer.
Avec le package **sp**, on peut fabriquer des objets de classes suivantes :

- les `SpatialPolygonsDataFrame`, si les unités spatiales sont définies pas des contours, comme des limites territoriales (une commune, un canton, un pays, un IRIS, etc).
- les `SpatialPointsDataFrame`, si les unités spatiales sont définies par des points comme c'est souvent le cas en géostatistique.
- les `SpatialPixelsDataFrame` ou `SpatialGridDataFrame`, si les unités spatiales correspondent à des pixels (diffèrent entre eux par la façon dont les informations sont stockées).
- les `SpatialLinesDataFrame`, si les objets sont des segments

Diverses classes de données spatiales : package spatstat

Avec le package **spatstat**, on peut fabriquer des objets de classes suivantes :

- les `ppp` pour les semis de points (Point Patterns)
- les `owin` pour les fenêtres
- les `im` pour les images pixelisées

Principaux types de formats géographiques

Les objets spatiaux sont créés dans R par importation de fichiers de divers formats

- format vectoriel : ESRI shapefile (importé avec la fonction `readShapePoly` ou `readShapeSpatial` du package **maptools**)
- format vectoriel : MAPINFO (importé avec la fonction `readOGR` du package **rgdal**))
- format raster pour les images (importé avec la fonction `readAsciiGrid` du package **maptools** si format Acii initial, ou avec

Format vectoriel : ESRI shapefile

ESRI=*Environmental Systems Research Institute*
Un ESRI shapefile est formé de :

- un fichier principal (.shp) qui contient toute l'information liée à la géométrie des objets décrits qui peuvent être : des points, des lignes ou des polygones ;
- un fichier (.shx) qui stocke l'index de la géométrie ;
- un fichier dBASE (.dbf) pour les données attributaires (ou données statistiques) ;
- des fichiers facultatifs comme un fichier sur les datums/projections (.prj).

Importation d'un shapefile en SpatialPolygonsDataFrame

Le code ci-dessous importe un fichier `shapefile` qui contient les contours géographiques et un certain nombre d'informations (taux de criminalité, taux de chômage, etc.) des districts de la ville de Columbus aux Etats-Unis.

```
>library(spdep)
>columbus <- readShapePoly(system.file("etc/shapes/columbus.shp",
 package="spdep")[1])
>class(columbus)
[1] "SpatialPolygonsDataFrame"
attr(,"package")
[1] "sp"
> dim(columbus)
[1] 49 20
> head(columbus@data)
>plot(columbus)
```

Les objets de type SpatialPolygonsDataFrame

Pour accéder et connaître la structure des variables d'intérêt d'un objet de type `SpatialPolygonsDataFrame` :

```
str(columbus@data)
```

Enfin, pour afficher les contours géographiques :

```
plot(columbus,axes=TRUE)
title("Neighbourhoods in Columbus")
```

Les objets de type SpatialPolygonsDataFrame

On peut également représenter des couleurs différentes selon une variable d'intérêt. Par exemple, pour représenter les districts du "centre" en rouge et les districts périphériques en "bleu", on utilisera le code suivant :

```
CP<-as.numeric(as.factor(columbus@data$CP))
col.map<-c("royalblue2","red3")
plot(columbus,col=col.map[CP])
legend("topleft", legend = c("0","1"), cex = 0.8,
title = "Core-periphery dummy ",fill=col.map[1:2])
```

Construction d'un objet de type SpatialPointsDataFrame

Le jeu de données contient des informations sur les prix de l'immobilier dans les grandes villes de France.
Le code ci dessous construit d'abord un objet de type `SpatialPoints` qui contient les coordonnées géographiques des observations :

```
library(GeoXp)
data(immob)
immob.sp = SpatialPoints(cbind(immob$longitude, immob$latitude))
class(immob.sp)
```

Ensuite, on associe à cet objet un jeu de caractéristiques des points afin de construire un objet de type `SpatialPointsDataFrame` dont la représentation peut se faire avec la fonction `plot` :

```
immob.spdf = SpatialPointsDataFrame(immob.sp, immob)
class(immob.spdf)
plot(immob.spdf)
```

Construction d'un objet de type SpatialPixelsDataFrame

L'exemple ci-dessous montre un exemple de création et d'affichage d'objet de type `SpatialPixelsDataFrame`.

```
data(meuse.grid)
m = SpatialPixelsDataFrame(points = meuse.grid[c("x", "y")],
data = meuse.grid)
class(m)
plot(m)
```

Format vectoriel : MAPINFO

Le format MIF/MID est le format d'import-export de MapInfo, les formats natifs de MapInfo étant les formats .DAT/.ID/.MAP/.TAB.

Les données sont réparties dans deux fichiers ASCII : le fichier MID contient les attributs alphanumériques, à chaque fichier MID étant associé un fichier MIF. Chaque ligne du fichier MID est associée à un objet graphique du fichier MIF.

Format vectoriel : MAPINFO

Le fichier MIF contient essentiellement les données graphiques et un en-tête décrivant les paramètres suivants :

- un numéro de version (A)
- le caractère servant de séparateur des attributs alphanumériques (B),
- le système de coordonnées (C).
- le type de projection (C).
- les paramètres de transformation des coordonnées (C),
- la colonne des attributs qui sert d'index,
- le nombre de colonnes des attributs alphanumériques c'est à dire le nombre de champs définis dans la table (D),
- le nom des colonnes des attributs ainsi que leur type (caractère, numérique) et leur longueur (E).

Importation d'un fichier MAPINFO : exemple

```
>library(rgdal)
> xy <- readOGR("departements_region.mif",
"departements_region")
OGR data source with driver: MapInfo File
Source: "departements_region.mif", layer:
"departements_region"
with 98 features and 7 fields
Feature type: wkbPolygon with 2 dimensions
> class(xy)
[1] "SpatialPolygonsDataFrame"
attr(,"package")
[1] "sp"
```

Suite : carte choroplèthe de la population

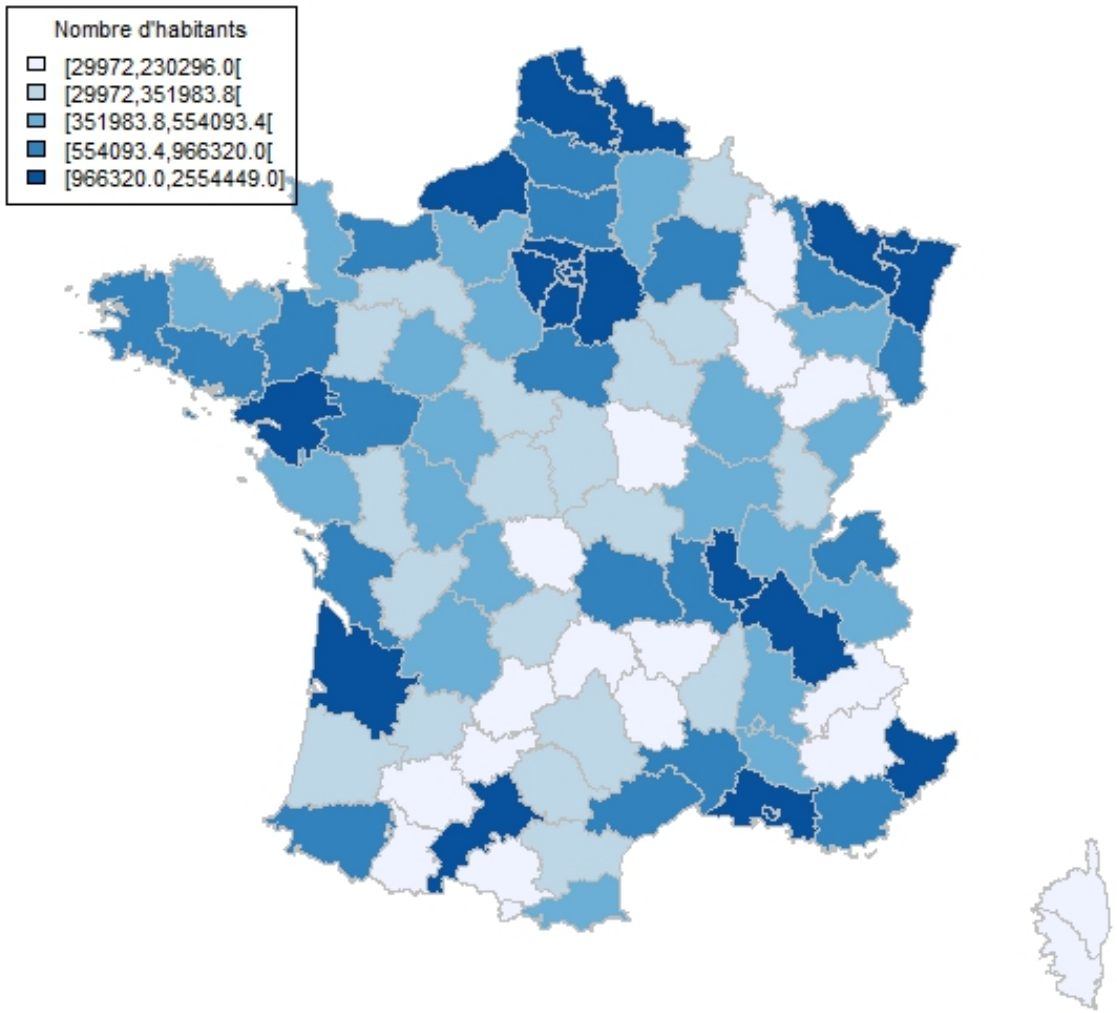
```
> plotclr <- c("#EFF3FF", "#BDD7E7", "#6BAED6", "#3182BD",
 "#08519C")

> breaks<-quantile(xy@data$PSDC,c(0,0.2,0.4,0.6,0.8,1))

> plot(xy,col=plotclr[findInterval(xy@data$PSDC, breaks,
all.inside=TRUE)], border='grey')

> legend("topleft", legend = c("[29972,230296.0[",
"[29972,351983.8[", "[351983.8,554093.4[", "[554093.4,966320.0[",
"[966320.0,2554449.0]"),
title = "Nombre d'habitants",fill=plotclr,cex=0.7)
```

Suite : carte choroplèthe de la population



Format raster

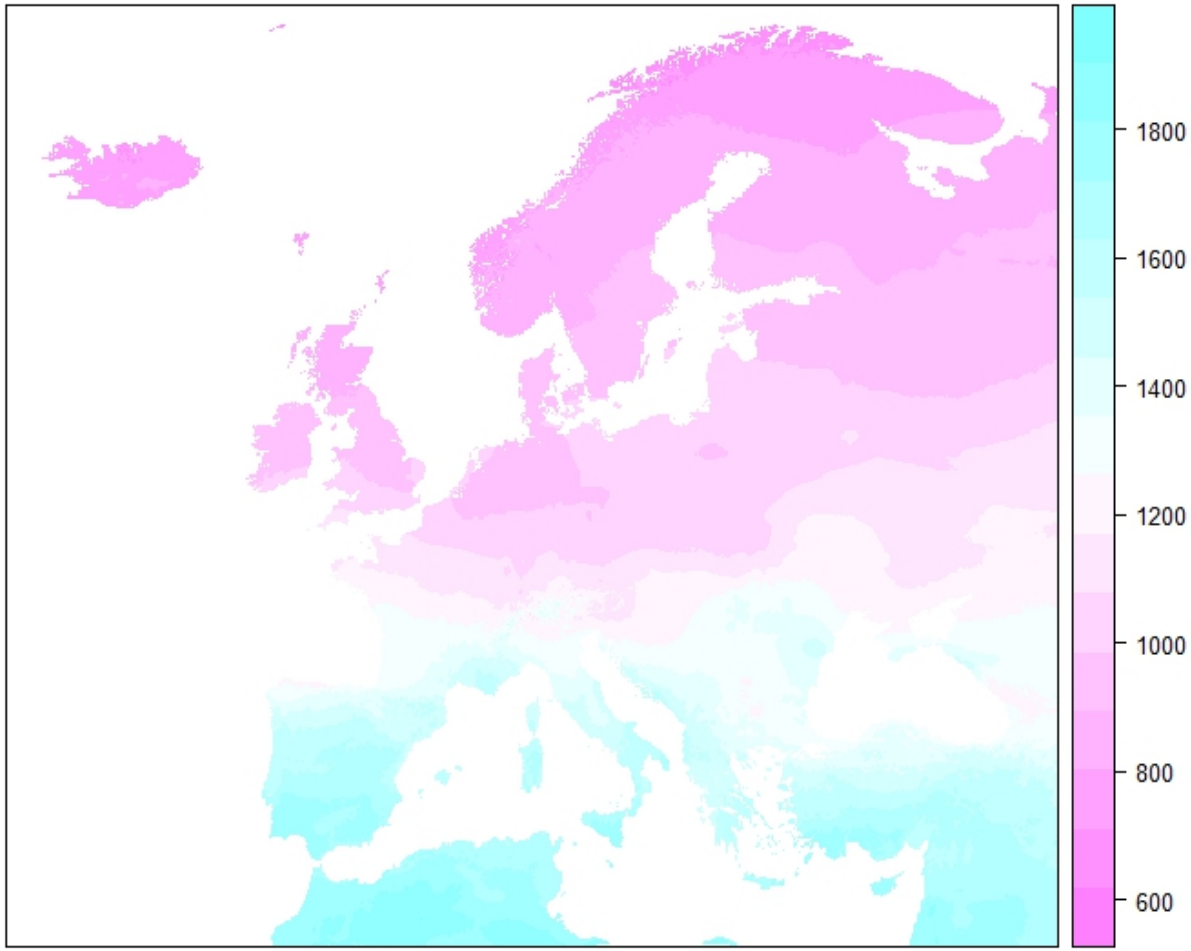
Importation d'un fichier au format Ascii .asc avec la fonction `readAsciiGrid` du package **maptools**

```
> gr <- readAsciiGrid("pvgis_g13year00.asc")

> proj4string(gr)=CRS("+proj=longlat +ellps=WGS84")

> class(gr)
[1] "SpatialGridDataFrame"
attr(,"package")
[1] "sp"
> spplot(gr,axes=TRUE)
```

Format raster : exemple ensoleillement en Europe



Pour aller plus loin

Note : une fois le jeu de données importées dans R, il vaut mieux le sauver au format .Rdata

```
save.image(file = "Departements.RData")
```

Pour aller plus loin sur manipulation d'objets spatiaux en R :

`http://geostat-course.org/system/files/monday_slides.pdf`

Analyse statistique des données spatiales II

Christine Thomas-Agnan

Toulouse School of Economics

30 octobre 2012

Les grands types

Rappelons les trois grands types de données géoréférencées :

- les données ponctuelles ou de type géostatistique
- les données surfaciques ou de type économétrie spatiale
- les données de type semis de points.

Un exemple de données de type surfacique : Columbus

Le jeu de données économiques de Luc Anselin sur la ville de Columbus (Ohio, US) en 1980 se trouve dans le package **s**pdep au format .Rdata et dans le package maptools au format .shp. La ville de Columbus est découpée en 49 quartiers pour lesquels on dispose de 18 attributs parmi lesquels nous avons choisi

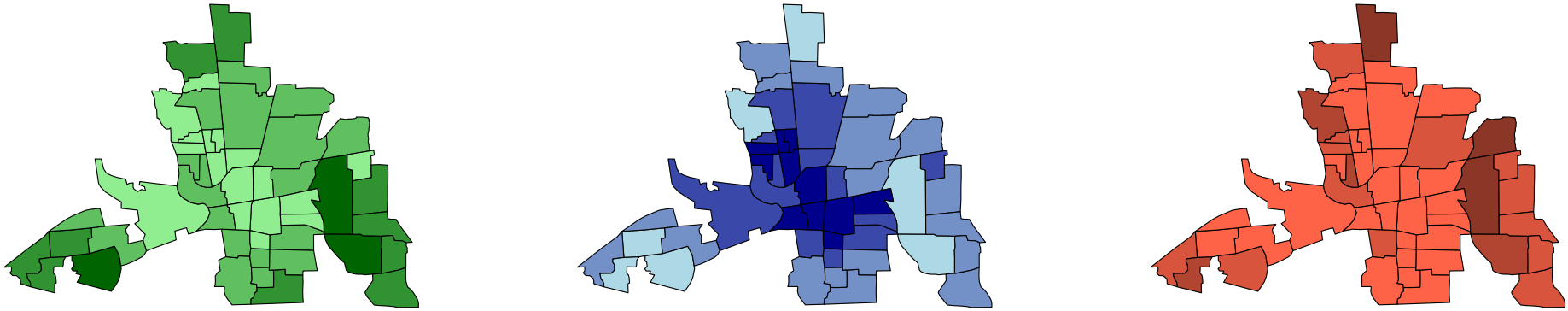
- HOVAL valeur immobilière en $ 1000
- INC revenu moyen des ménages en $ 1000
- CRIME nombre de cambriolages et vols de voitures pour 1000 habitants

On va chercher à expliquer la valeur immobilière par la criminalité dans les quartiers et le revenu des ménages.

Un exemple de données de type surfacique : Columbus

```
library(classInt)
q5 <- classIntervals(columbus@data$INC , n=4, style="equal")
plot(columbus, col=findColours(q5, c("lightgreen", "darkgreen")))
```

Left : INC, center : CRIME, right : HOVAL



Un autre exemple de données de type surfacique : North Carolina SIDS data

Lire

`http://cran.r-project.org/web/packages/spdep/vignettes/sids.pdf`

```
library(spdep)
nc_file <- system.file("etc/shapes/sids.shp", package = "spdep")
llCRS <- CRS("+proj=longlat +datum=NAD27")
nc <- readShapeSpatial(nc_file, ID = "FIPSNO", proj4string = llCRS)
```

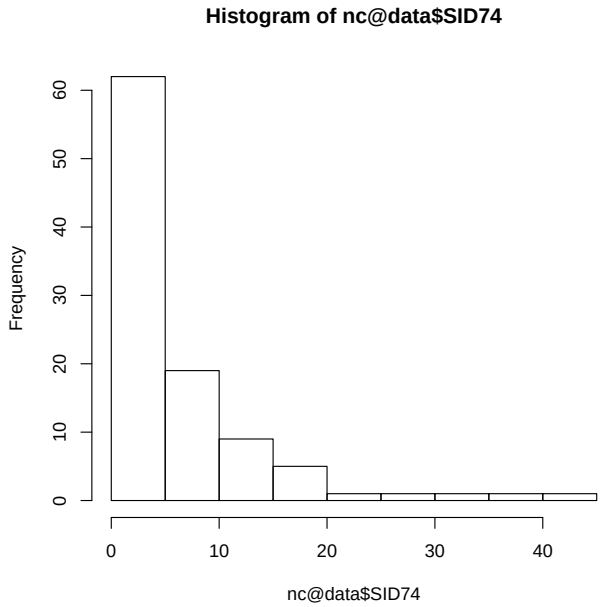
l'objet 'nc' contient les variables ainsi que les polygones.

Un autre exemple de données de type surfacique : North Carolina SIDS data

Le jeu de données est au sujet du 'Suden Infant Death Syndrome' : mort subite du nourrisson. Nous retenons les attributs suivants, pour chaque conté de Caroline du nord :

```
BIR74 births, 1974-78
SID74 SID deaths, 1974-78
NWBIR74 non-white births, 1974-78
BIR79 births, 1979-84
SID79 SID deaths, 1979-84
NWBIR79 non-white births, 1979-84
```

Un autre exemple de données de type surfacique : North Carolina SIDS data

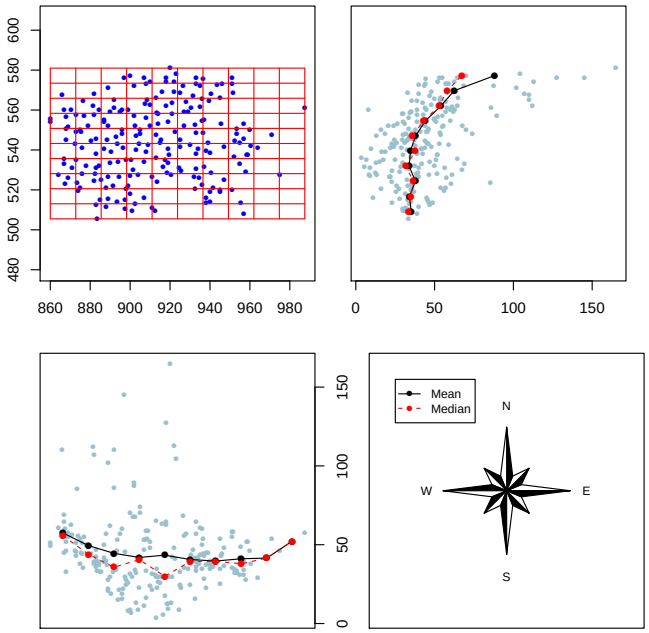


Un exemple de données de type géostatistique : Baltimore

Le jeu de données Baltimore se trouve dans le package **s**pdep au format .Rdata. Il contient les caractéristiques de 211 transactions immobilières sur des maisons dans la ville de Baltimore (Maryland) en 1978. Nous avons choisi de conserver les attributs suivants

- PRICE le prix de la maison
- NROOM le nombre de pièces
- NBATH le nombre de salles de bain
- PATIO : 1 si patio, 0 sinon
- FIREPL : 1 si cheminée, 0 sinon
- AC : 1 si climatisation, 0 sinon
- BMENT : 1 si cave, 0 sinon
- NSTOR : nombre d'étages
- AGE : age du batiment
- LOTSZ : surface du terrain (en centaine de pieds carrés)
- SQFT : surface de l'intérieur (en centaine de pieds carrés)

Un exemple de données de type géostatistique : Baltimore

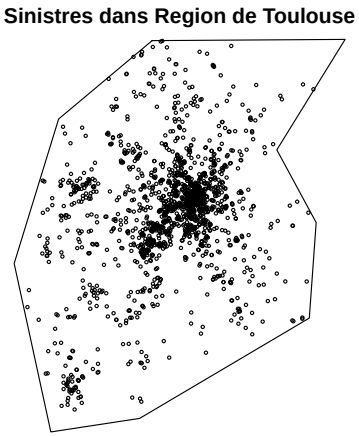


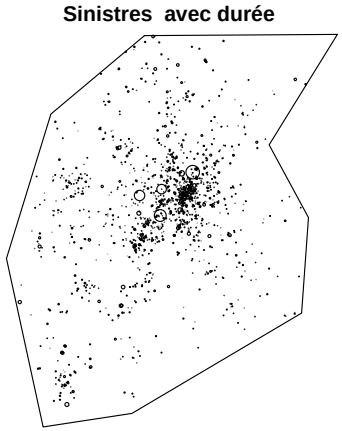
Un exemple de semis de points : Pompiers de Toulouse

Cette base provient du SDIS 31 (Service Départmental d'Incendie et de Secours). Elle contient les localisations et caractéristiques d'un échantillon de sinistres durant le mois de janvier de l'année 2004 sur une zone autour de la ville de Toulouse. La variable M contient la durée du sinistre en "minutes sur le lieu du sinistre" multipliée par le nombre de pompiers mobilisés.

Un exemple de semis de points : Pompiers de Toulouse

```
library(spatstat)
load("Pompiers_janvier+region.Rdata")
PP=ppp(sinistres_janvier$X,sinistres_janvier$Y,window=Region)
marks(PP)<-sinistres_janvier$M
plot(PP,main="Sinistres  avec durée ")
PPu=unmark(PP)
plot(PPu,main="Sinistres dans Region de Toulouse",cex=0.4)
```





Notations

Une seule notation commune pour données ponctuelles ou surfaciques : champ X_s observé en des localisations $s_1, \cdots, s_n$

- si données ponctuelles : X_s désigne la variable aléatoire de la caractéristique X au point s
- si données surfaciques, X_s désigne la variable aléatoire de la caractéristique X dans l'unité spatiale dont le représentant est s

La loi du champ X_s est caractérisée par

- les lois marginales de X_s pour chaque localisation s
- les lois conjointes de vecteurs $X_{s_1}, \cdots, X_{s_n}$ pour un ensemble fini de localisations $s_1, \cdots, s_n$

Nombre d'observations

On observe généralement une seule réalisation de X_s et ce pour un nombre fini de sites s sauf si on est dans le cas d'observations répétées : plusieurs données mais une seule réalisation ! !
Solution : puiser des forces dans la continuité spatiale du phénomène et dans la corrélation entre lieux voisins pour rendre cette inférence possible.

Décomposition classique

X_s champ aléatoire à valeurs réelles admettant un moment d'ordre un fini : $\mathbb{E}(X_s) < \infty$.
Décomposition classique en deux parties

$$X_s = \mathbb{E}(X_s) + (X_s - \mathbb{E}(X_s))$$

Le terme déterministe $\mathbb{E}(X_s)$ s'appelle la **tendance** et modélise les variations à grande échelle du phénomène décrit par ce champ. Le terme aléatoire $(X_s - \mathbb{E}(X_s))$ s'appelle la **fluctuation** et modélise les variations du champ à petite échelle. Notons que la fluctuation a une moyenne nulle.

Part d'arbitraire dans la décomposition

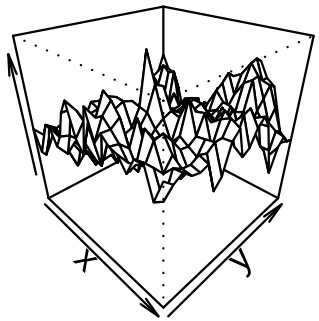
Dans la pratique, cette décomposition en deux termes pour un phénomène observé une fois n'est pas unique et c'est le choix du modélisateur d'affecter certains aspects à la partie aléatoire ou à la partie déterministe : une coupe transversale ne permet pas de distinguer entre hétérogénéité et autocorrélation.

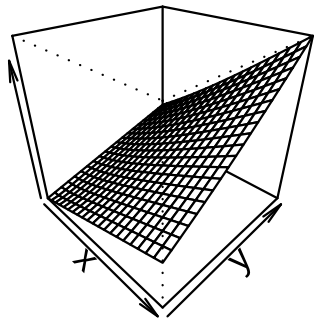
Décomposition classique

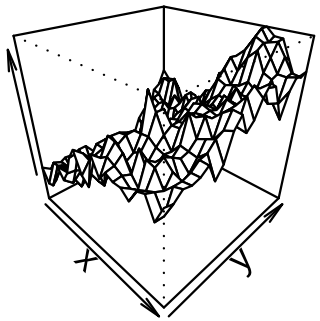
On dit qu'**il y a une tendance** lorsque $\mathbb{E}(X_s)$ est non constante dans l'espace : on dit aussi que la moyenne est non stationnaire.
Pour comprendre ce découpage, il est bon de penser à une montagne : le détail de la variation de l'élévation mesuré avec précision constitue le champ ; on peut penser à l'allure de la montagne vue d'avion telle qu'elle se découpe sur l'horizon comme à une tendance ; la différence entre l'élévation précise et cette tendance représente alors les accidents de terrain visibles de près.

Illustration

Champ (droite), tendance (centre) et fluctuation (gauche)







Hétérogénéité : définition

La répartition marginale du champ aléatoire X_s varie avec la localisation s.
On dit qu'**il y a une tendance** lorsque $\mathbb{E}(X_s)$ est non constante dans l'espace (moyenne non stationnaire).
L'hétérogénéité spatiale sera prise en compte par l'usage de variables explicatives pour modéliser la tendance. Certaines de ces variables peuvent être spatiales de nature comme, par exemple, la distance à certains lieux d'intérêt pour le problème.
Il n'est pas suffisant de prendre en compte ces variables dans la moyenne pour évacuer totalement la structure spatiale du problème qui peut rester présente à l'ordre deux.

Autocorrélation : intuition

Everything is related to everything else but closer things more so.
Si la tendance est spécifique au moment d'ordre un d'un champ, l'autocorrélation concerne le moment d'ordre deux que l'on supposera exister dans ce paragraphe : on dit alors que le champ est du **second ordre**.

Autocorrélation : intuition

Everything is related to everything else but closer things more so.
Si la tendance est spécifique au moment d'ordre un d'un champ, l'autocorrélation concerne le moment d'ordre deux que l'on supposera exister dans ce paragraphe : on dit alors que le champ est du **second ordre**. Pour les données spatiales, une corrélation peut se produire entre X_s et X_t du fait de leur proximité géographique.
De façon qualitative, on parle d'autocorrélation spatiale **positive** pour une variable lorsqu'il y a regroupement géographique de valeurs similaires de la variable.

Autocorrélation : intuition

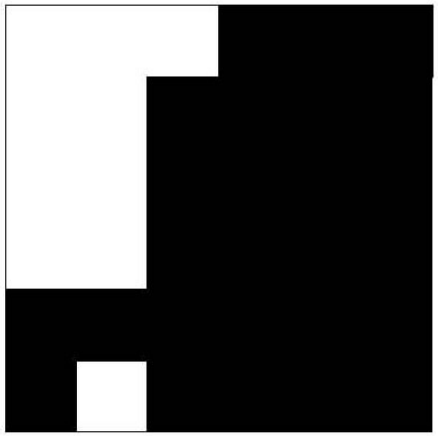
Everything is related to everything else but closer things more so.
Si la tendance est spécifique au moment d'ordre un d'un champ, l'autocorrélation concerne le moment d'ordre deux que l'on supposera exister dans ce paragraphe : on dit alors que le champ est du **second ordre**. Pour les données spatiales, une corrélation peut se produire entre X_s et X_t du fait de leur proximité géographique.
De façon qualitative, on parle d'autocorrélation spatiale **positive** pour une variable lorsqu'il y a regroupement géographique de valeurs similaires de la variable. De même, on parle d'autocorrélation spatiale **négative** pour une variable lorsqu'il y a regroupement géographique de valeurs dissemblables de la variable.

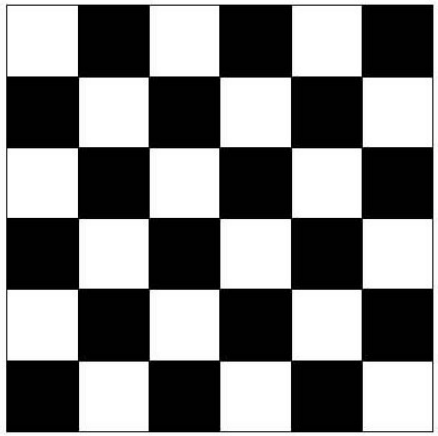
Autocorrélation : intuition

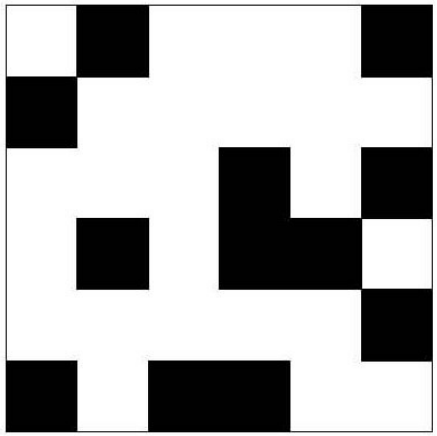
Everything is related to everything else but closer things more so.
Si la tendance est spécifique au moment d'ordre un d'un champ, l'autocorrélation concerne le moment d'ordre deux que l'on supposera exister dans ce paragraphe : on dit alors que le champ est du **second ordre**. Pour les données spatiales, une corrélation peut se produire entre X_s et X_t du fait de leur proximité géographique.
De façon qualitative, on parle d'autocorrélation spatiale **positive** pour une variable lorsqu'il y a regroupement géographique de valeurs similaires de la variable. De même, on parle d'autocorrélation spatiale **négative** pour une variable lorsqu'il y a regroupement géographique de valeurs dissemblables de la variable. Enfin, on parle **d'absence d'autocorrélation** pour une variable lorsqu'il n'y a pas de relation entre la proximité géographique et le degré de ressemblance des valeurs de la variable.

Autocovariance : Illustration

Prenons pour illustrer cette notion l'exemple d'un champ dichotomique à valeurs 0 ou 1 représentées respectivement par les couleurs blanche et noire et constant sur les carrés d'une grille régulière.







Modélisation de l'autocorrélation

Deux approches selon le type

- pour des données de type géostatistique, l'autocorrélation se modélise par la fonction d'autocovariance ou le **variogramme**,
- pour les données de type latticiel il se modélise par l'intermédiaire des **matrices de voisinage** et se mesure par les **indices de Moran et Geary**

Notions de stationnarité

La structure de covariance d'un champ du second ordre est définie par la fonction d'autocovariance

$$R(s,t) = \mathbb{C}ov(X_s, X_t)$$

Pour modéliser un tel champ, une des hypothèses simplificatrices que l'on est souvent amené à faire sur sa structure de covariance est celle de la stationnarité.
La **stationnarité stricte** ou forte d'un champ suppose que la loi du vecteur $X_{s_1}, \ldots, X_{s_k}$ est invariante par translation quel que soit le nombre de points k et quelles que soient leurs positions $s_1, \ldots s_k$ i.e. $X_{s_1}, \ldots, X_{s_k}$ a même loi que $X_{s_1+h}, \ldots, X_{s_k+h}$ quel que soit $h \in \mathbb{R}^d$.

Caractérisation mathématique des fonctions d'autocovariance

Les fonctions d'autocovariance peuvent être caractérisées par la propriété mathématique suivante :
Une fonction $R(s,t)$ de $\mathbb{R}^2$ à valeurs dans $\mathbb{R}$ est une fonction d'autocovariance d'un champ aléatoire réel du second ordre si et seulement si elle est **de type positif** c'est à dire que quels que soit l'entier k, quels que soient les k sites $s_1, \ldots, s_k$ et les réels $a_1, \ldots, a_k$, on a

$$\sum_{i=1}^{k} \sum_{j=1}^{k} a_i a_j R(s_i, s_j) \geq 0.$$

Notions de stationnarité

Un champ aléatoire X_s à valeurs réelles du second ordre est dit **stationnaire au second ordre** ou au sens faible s'il existe un vecteur $\mu \in \mathbb{R}$ et une fonction
$R : \mathbb{R}^d \mapsto \mathbb{R}$ dite fonction d'autocovariance tels que

$$\mathbb{E}(X_s) = \mu \tag{1}$$

$$\mathbb{C}ov(X_s, X_{s+h}) = R(h) \tag{2}$$

Notons que dans ce cas, la fonction d'autocovariance est une fonction d'une variable au lieu de deux. Il est clair que la stationnarité forte implique la stationnarité faible. Dans le cas gaussien, ces deux notions sont équivalentes puisque les moments d'ordre un et deux déterminent la distribution.

Caractérisation mathématique des fonctions d'autocovariance stationnaires

Une fonction $R(s)$ de $\mathbb{R}$ à valeurs dans $\mathbb{R}$ est une fonction d'autocovariance d'un champ aléatoire réel stationnaire du second ordre si et seulement si elle est **de type positif** ce qui signifie dans ce cas que la fonction de deux variables $(s,t) \mapsto R(s-t)$ est de type positif. Notons que le vocabulaire "de type positif" est le même mais qu'il s'applique dans un cas à une fonction de deux variables et dans l'autre à une fonction d'une variable.

Stationnarité intrinsèque

La stationnarité est souvent une hypothèse trop forte dans les applications et une façon de l'affaiblir est de considérer la **stationnarité intrinsèque**. On n'exige pas l'existence d'un moment d'ordre un pour le champ lui-même mais seulement pour les accroissements du champ et l'on demande que

$$\mathbb{E}(X_{s+h} - X_s) = 0$$
$$\mathbb{V}\mathrm{ar}(X_{s+h} - X_s) = 2\gamma(h) = \mathbb{E}(X_{s+h} - X_s)^2$$

La fonction γ s'appelle alors le **semi-variogramme** et 2γ le **variogramme**.

Stationnarité et stationnarité intrinsèque

Dans le cas où le champ est stationnaire (donc nécessairement intrinsèquement stationnaire), il existe la relation suivante entre variogramme et fonction d'autocovariance

$$\begin{aligned}\mathbb{V}\mathrm{ar}(X_{s+h} - X_s) &= \mathbb{V}\mathrm{ar}(X_{s+h}) + \mathbb{V}\mathrm{ar}(X_s) - 2\mathbb{C}\mathrm{ov}(X_s, X_{s+h}) \\ &= 2\sigma^2 - 2R(h) \\ &= 2\gamma(h)\end{aligned}$$

Caractérisation mathématique des fonctions variogramme

Une fonction $\gamma(t)$ de $\mathbb{R}$ vers $\mathbb{R}$ est le variogramme d'un champ aléatoire intrinsèquement stationnaire si et seulement si $-\gamma$ est conditionellement de type positif d'ordre 1 i.e. pour tout entier k, pour tout ensemble de k sites $s_1, \ldots, s_k$ et tout choix de réels $a_1, \ldots, a_k$, on a

$$-\sum_{i=1}^{k}\sum_{j=1}^{k} a_i a_j \gamma(s_i - s_j) \geq 0,$$

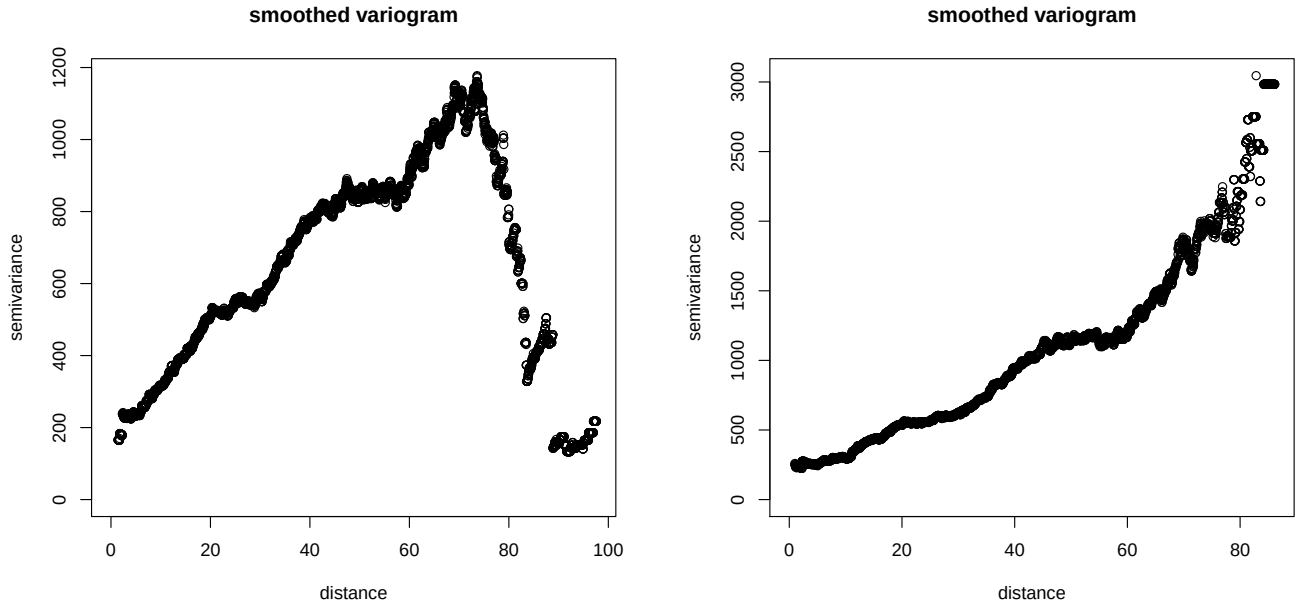
dès que $a_1, \ldots, a_k$ satisfont la condition $\sum_{i=1}^{k} a_i = 0$.
On parle alors d'un **variogramme valide**.

Notion d'isotropie

Un champ intrinsèquement stationnaire est **isotrope** si son variogramme $\gamma(h)$ ne dépend que de la norme de h. Dans ce cas la fonction $\| h \| \longmapsto E(X_{s+h} - X_s)^2 = \gamma_0(\| h \|)$ est appellée variogramme **omnidirectionnel** isotrope.
On parle d'**anisotropie** lorsque l'hypothèse d'isotropie n'est pas vérifiée. On peut alors représenter une fonction variogramme univariée pour chaque direction : **variogramme directionnel**.
Si les lignes de niveau du variogramme sont des ellipses, on dit qu'il y a anisotropie géométrique. On peut alors se ramener à une configuration d'isotropie par une rotation composée par une affinité (A). Alors $\gamma(h) = \gamma_0(\| Ah \|)$.

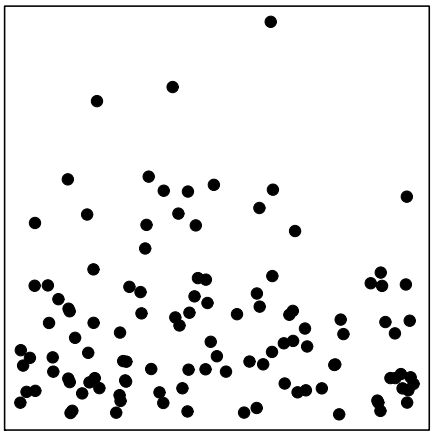
Illustration isotropie

Variogram directionnel lissé pour la variable PRICE dans le jeu de données Baltimore : à gauche angle $\pi/2$ et à droite angle $\pi/4$



Homogénéité d'un semis de points

La notion d'homogénéité est une notion d'ordre un : il s'agit de savoir si le nombre moyen de points par unité de surface est constant au travers du domaine. On parle aussi de façon équivalente de stationnarité.



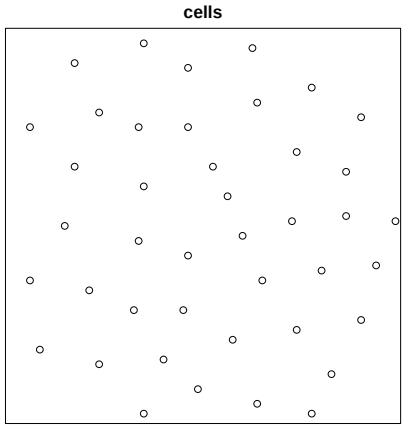
Interaction dans un semis de points

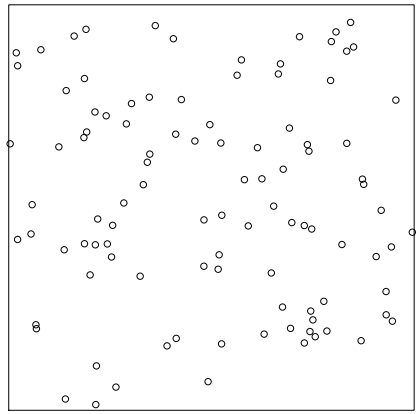
La notion d'interaction est une notion d'ordre deux : il s'agit de savoir si le nombre (aléatoire) de points $N(A)$ dans une partie de l'espace A est dépendant ou indépendant (de façon probabiliste) du nombre de points $N(B)$ dans une autre partie B disjointe de A. Les phénomènes qui présentent de l'attraction ou de la répulsion entre les points comportent une dépendance entre $N(A)$ et $N(B)$.
Par exemple

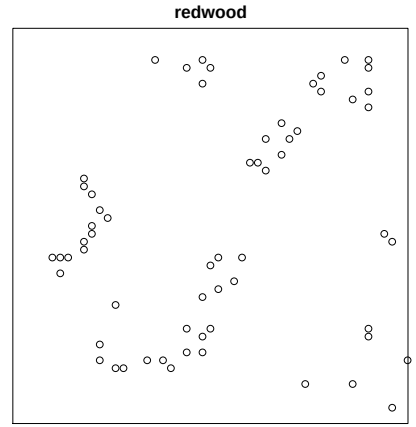
- les positions d'animaux sur un territoire présentent de la **répulsion en raison de la compétition pour la nourriture**
- les positions de personnes atteintes d'une maladie épidémique vont au contraire montrer de **l'attraction** en raison de la contagion

Interaction dans un semis de points : exemples

Gauche : régulier, Centre : Homogène, Droite : Agrégé

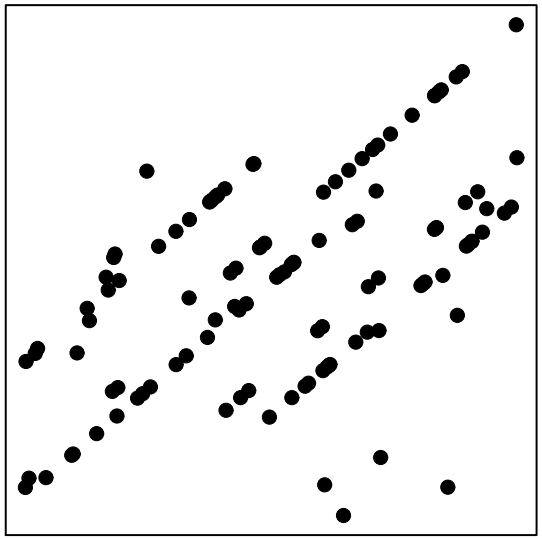






Isotropie dans un semis de points

On dit qu'un semis de points est **isotrope** lorsque toutes ses caractéristiques sont invariantes par rotation. Un exemple non isotrope



Analyse statistique des données spatiales III

Christine Thomas-Agnan

Toulouse School of Economics

31 octobre 2012

Les packages de R

- pour les données surfaciques : le package “spdep” par R. Bivand
- pour les données ponctuelles : les packages “gstat”, “geoR” et “geoRglm”
- pour les semis de points : le package “SpatStat” de A. Baddeley et R. Turner

Variogramme isotrope : effet de pépite

Effet de pépite : remarquons que $\gamma(0) = 0$. On dit que le processus est continu en moyenne quadratique si $\lim_{h \longrightarrow 0} \gamma(h) = 0$.
Cette condition équivaut à la continuité de la fonction d'autocovariance.
Si $\lim_{h \longrightarrow 0} \gamma(h) = c_0 \neq 0$ alors c_0 est appellé **effet pépite** (nugget effect) et témoigne d'une discontinuité dans le processus.

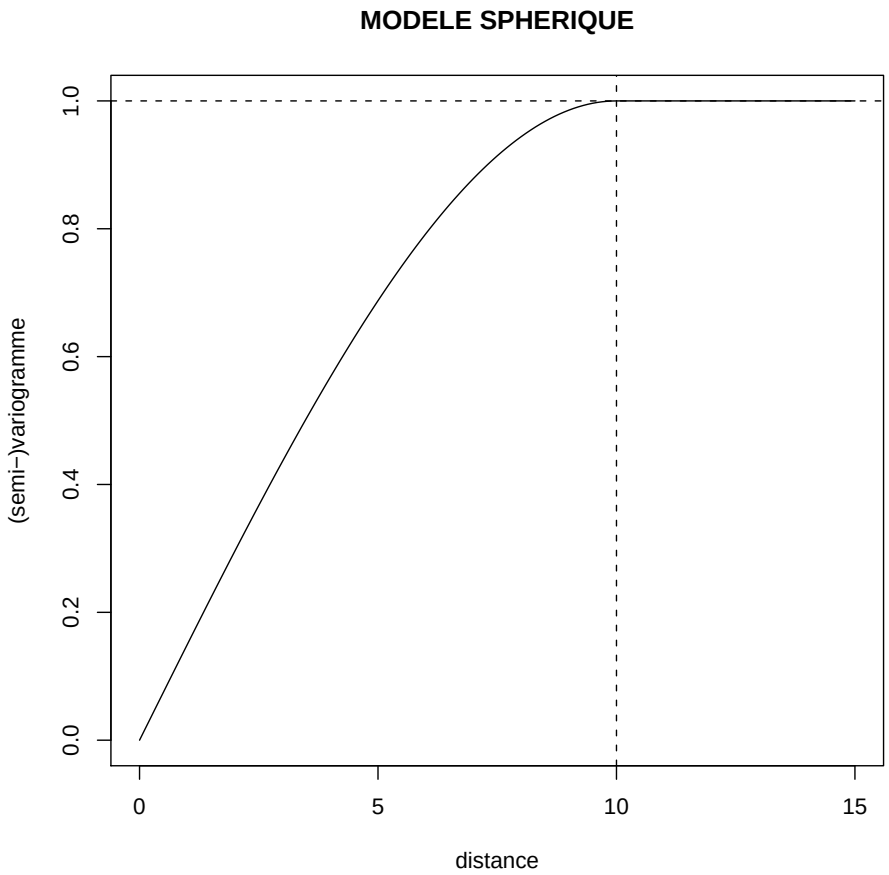
Variogramme isotrope : seuil et portée

Seuil : Si X_s est stationnaire et si $R(h) \longrightarrow 0$ quand $h \longrightarrow +\infty$ alors $\gamma(h)$ tends vers $R(0)$ appellé **seuil** (sill) du semi-variogramme.

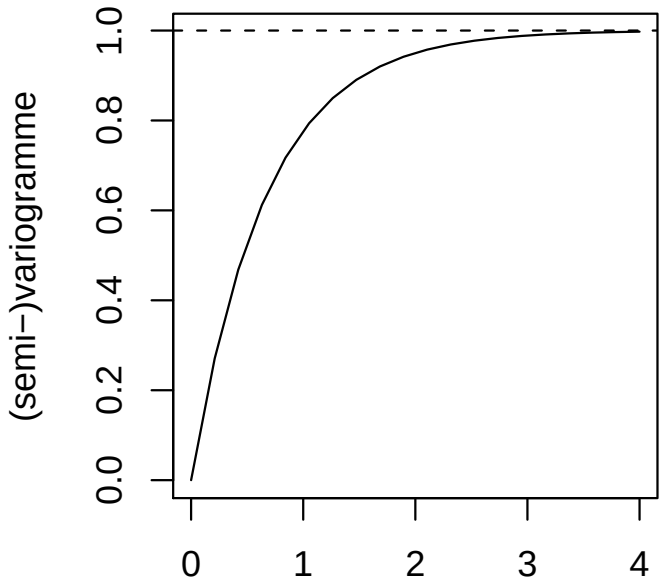
Portée : la plus petite valeur de $\| r \|$ telle que $\gamma(r(1+\epsilon)) = R(0)$ quel que soit $\epsilon > 0$ est appellée la **portée** (range) dans la direction r.

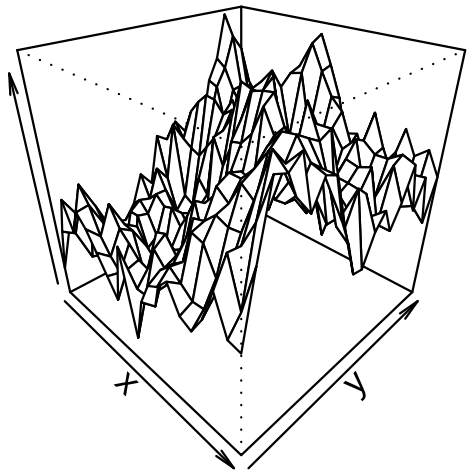
Portée pratique : la plus petite valeur de $\| r \|$ telle que $\gamma(r) = 0.95R(0)$ est appellée la **portée pratique** dans la direction r.

Illustration : seuil et portée

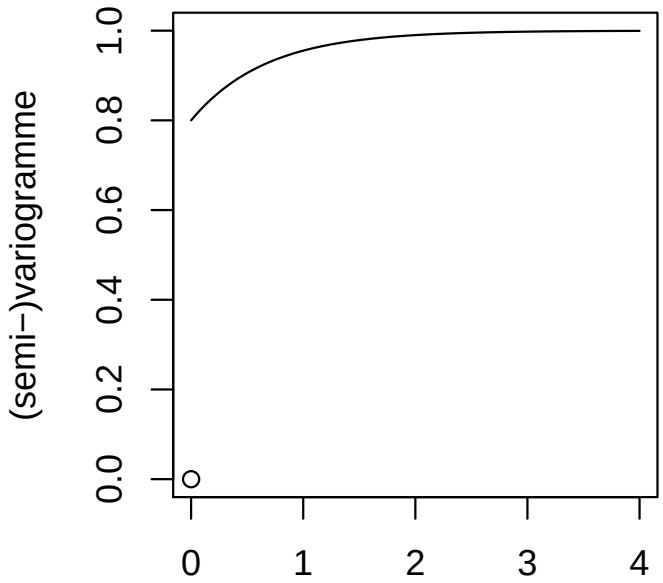


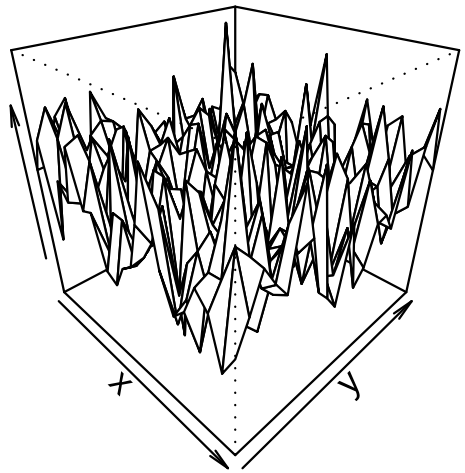
Exemples de variogrammes isotropes : exponentiel sans effet de pépite





Exemples de variogrammes isotropes : exponentiel avec effet de pépite





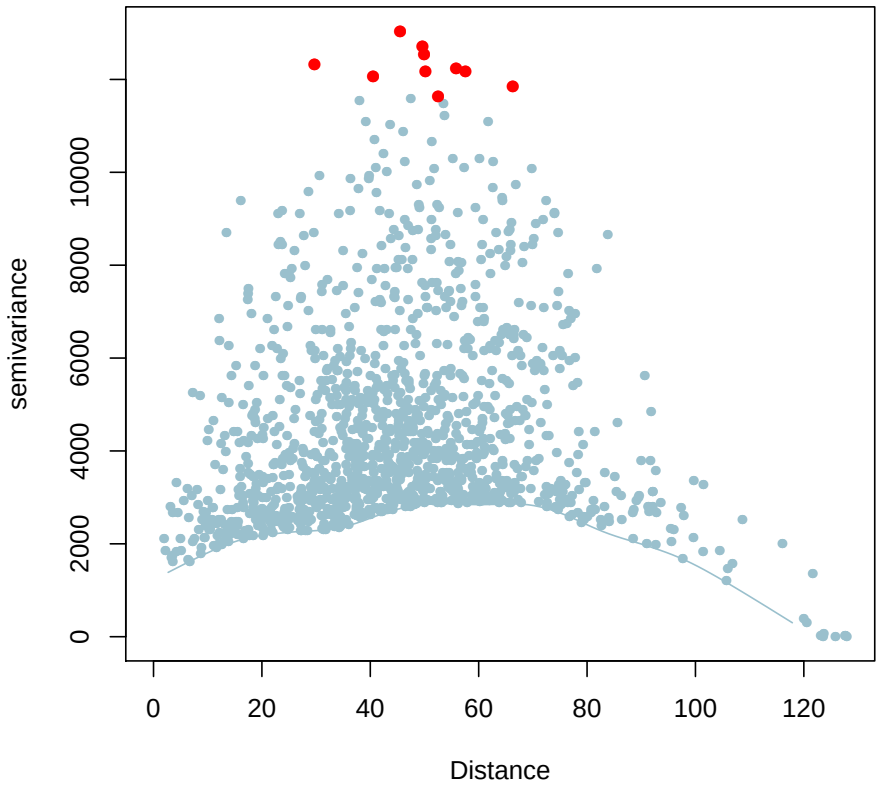
Nuage de variogramme

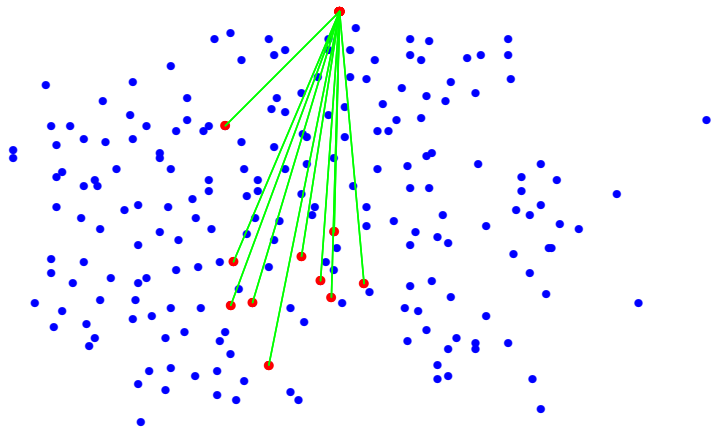
Soit X un champ centré intrinsèquement stationnaire.

- **Cas isotrope : omnidirectionnel**
  Soit h_{ij} la distance entre deux unités géographiques s_i et s_j. Le **"Nuage de variogramme"** est le nuage de points d'abscisse h_{ij} et d'ordonnée $\frac{1}{2}(X_{s_i} - X_{s_j})^2$.
  Dans le **"Nuage de variogramme"** normalisé, les ordonnées sont $\frac{(X_{s_i} - X_{s_j})^2}{\beta_{ij}}$, où β_{ij} est la moyenne des $(X_{s_i} - X_{s_j})^2$ pour les couples distants de h_{ij}.
- **Cas non isotrope : directionnel**
  un graphique pour chaque direction e : les points d'abscisse h ont pour ordonnées $\frac{1}{2}(X_{s_i+he} - X_{s_i})^2$ chaque fois qu'il existe un j tel que $s_i + he = s_j$.

Nuage de variogramme : exemple isotrope

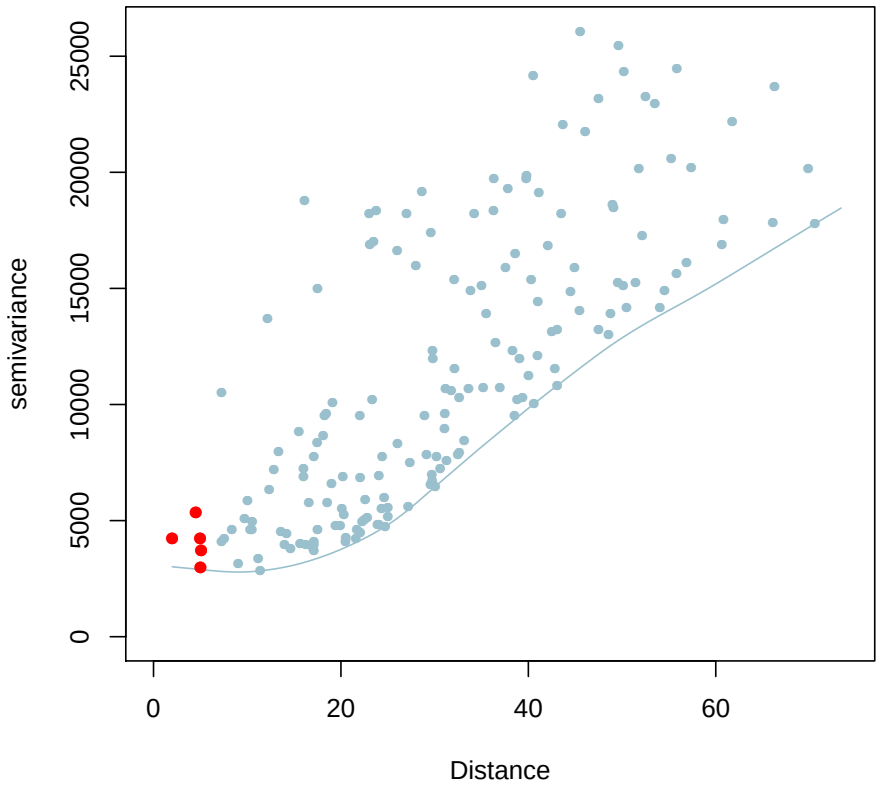
Données Baltimore, nuage de variogramme omnidirectionnel.

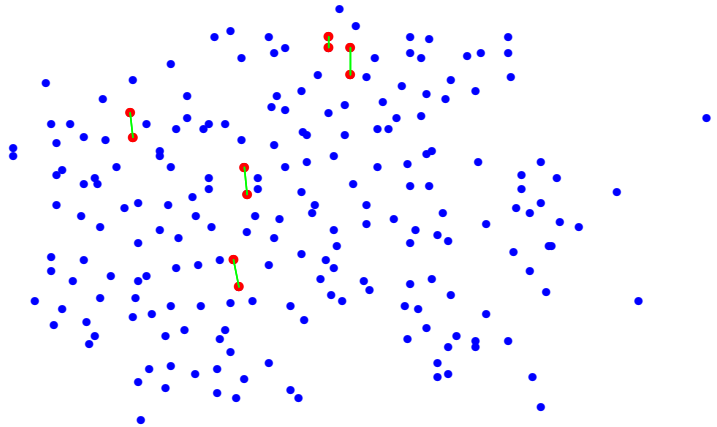




Nuage de variogramme : exemple non isotrope

Données Baltimore, nuage de variogramme unidirectionnel, direction $\pi/2$.





Matrices de poids

La matrice de poids est la version spatiale de l'opérateur retard en séries temporelles.
Pour n sites géographiques, une matrice de poids W est de taille $n \times n$ et son élément w_{ij} indique l'intensité de la proximité entre la zone i et la zone j (elle spécifie la topologie du domaine).
Par convention $w_{ii} = 0$.
W n'est pas nécéssairement symétrique. Si W quelconque, $(W + W')/2$ est symétrique.

Matrices de poids : normalisation

On dit qu'une matrice de poids est **normalisée** si

$$\sum_{j=1}^{n} w_{ij} = 1.$$

Utilité : cette contrainte permet de rendre les paramètres spatiaux comparables entre divers modèles ; cette contrainte a une conséquence sur le vecteur spatialement décalé (voir plus loin).
On peut normaliser une matrice W en W^* en divisant chaque ligne par son total.
Attention : si W est symétrique, sa normalisée W^* n'est plus symétrique.
Attention : si W est normalisée, sa symétrisée $(W + W')/2$ n'est plus normalisée

Variable spatialement décalée

Si X est une variable et W une matrice de poids, la variable spatialement décalée associée à X est WX.

Variable spatialement décalée

Si X est une variable et W une matrice de poids, la variable spatialement décalée associée à X est WX.

Si W est binaire, le terme i de WX est la somme des valeurs de X associées aux voisins du site i.

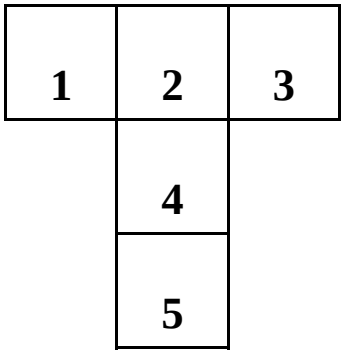
Variable spatialement décalée

Si X est une variable et W une matrice de poids, la variable spatialement décalée associée à X est WX.

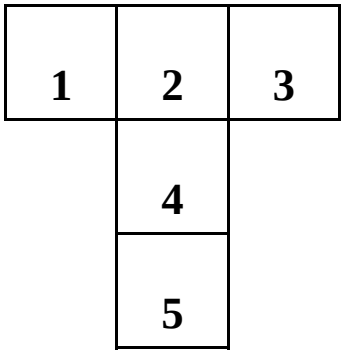
Si W est binaire, le terme i de WX est la somme des valeurs de X associées aux voisins du site i.

Si W est normalisée, le terme i de WX est la moyenne (pondérée par la proximité) des valeurs de X sur les voisins du site i.
Noter que même si X ne présente pas d'autocorrélation spatiale, WX va en présenter.

Petit exemple

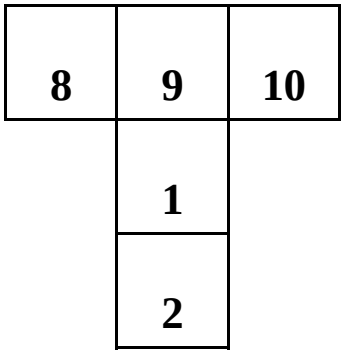


Petit exemple



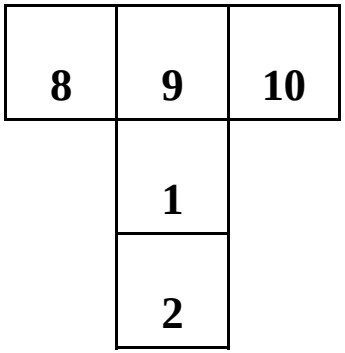
Si $\mathbf{z} = (8 \;\; 9 \;\; 10 \;\; 1 \;\; 2)^T$,

Petit exemple



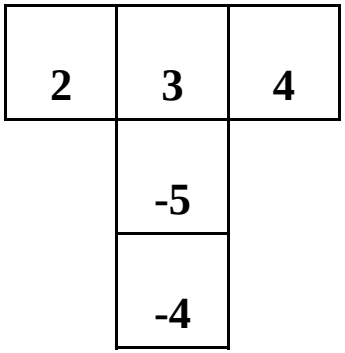
Si $\mathbf{z} = (8 \;\; 9 \;\; 10 \;\; 1 \;\; 2)^T$,

Petit exemple



Si $\mathbf{z} = (8 \;\; 9 \;\; 10 \;\; 1 \;\; 2)^T$, $\bar{\mathbf{z}} = (6 \;\; 6 \;\; 6 \;\; 6 \;\; 6)^T$

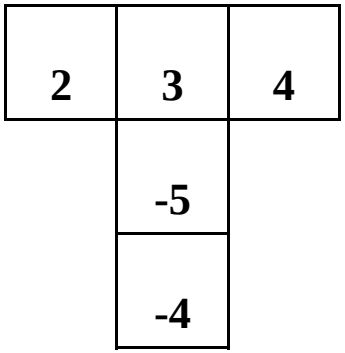
Petit exemple



Si $\mathbf{z} = (8 \;\; 9 \;\; 10 \;\; 1 \;\; 2)^T$, $\bar{\mathbf{z}} = (6 \;\; 6 \;\; 6 \;\; 6 \;\; 6)^T$

$\mathbf{z} - \bar{\mathbf{z}} = (2 \;\; 3 \;\; 4 \;\; -5 \;\; -4)^T$

Petit exemple

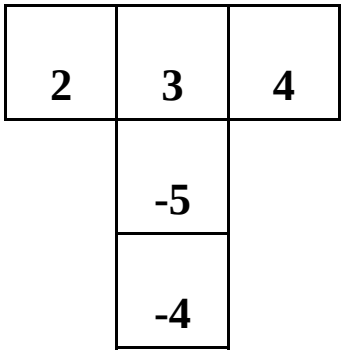


Si $\mathbf{z} = (8 \;\; 9 \;\; 10 \;\; 1 \;\; 2)^T$, $\bar{\mathbf{z}} = (6 \;\; 6 \;\; 6 \;\; 6 \;\; 6)^T$

$\mathbf{z} - \bar{\mathbf{z}} = (2 \;\; 3 \;\; 4 \;\; -5 \;\; -4)^T$

$W \times (\mathbf{z} - \bar{\mathbf{z}})$

Petit exemple

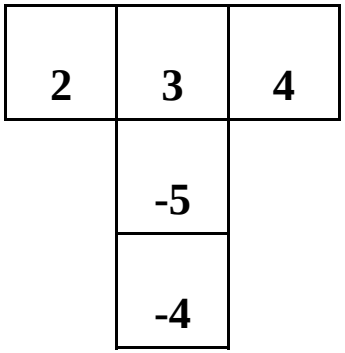


Si $\mathbf{z} = (8 \;\; 9 \;\; 10 \;\; 1 \;\; 2)^T$, $\bar{\mathbf{z}} = (6 \;\; 6 \;\; 6 \;\; 6 \;\; 6)^T$

$\mathbf{z} - \bar{\mathbf{z}} = (2 \;\; 3 \;\; 4 \;\; -5 \;\; -4)^T$

$$W \times (\mathbf{z} - \bar{\mathbf{z}}) = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1/3 & 0 & 1/3 & 1/3 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 0 & 1/2 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix} \times \begin{pmatrix} 2 \\ 3 \\ 4 \\ -5 \\ -4 \end{pmatrix}$$

Petit exemple



$$\text{Si } \mathbf{z} = (8 \;\; 9 \;\; 10 \;\; 1 \;\; 2)^T, \;\; \bar{\mathbf{z}} = (6 \;\; 6 \;\; 6 \;\; 6 \;\; 6)^T$$

$$\mathbf{z} - \bar{\mathbf{z}} = (2 \;\; 3 \;\; 4 \;\; -5 \;\; -4)^T$$

$$W \times (\mathbf{z} - \bar{\mathbf{z}}) = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1/3 & 0 & 1/3 & 1/3 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 0 & 1/2 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix} \times \begin{pmatrix} 2 \\ 3 \\ 4 \\ -5 \\ -4 \end{pmatrix} = \begin{pmatrix} 3 \\ 1/3 \\ 3 \\ -1/2 \\ -5 \end{pmatrix}$$

Formats des matrices de poids sous R

Le type 'matrix' n'est pas optimal pour stocker une matrice de voisinage (ce sont plutôt des matrices creuses). Par exemple, pour les données SIDS

```
class(nc)
```

Le package spdep utilise plusieurs classes (types) de fichiers pour cela

- la classe 'nb'
- la classe 'listw'
- la classe 'knn'

Il existe des fonctions de conversion d'un type à l'autre : 'knn2nb', 'mat2listw','listw2mat','nb2listw','nb2mat'.

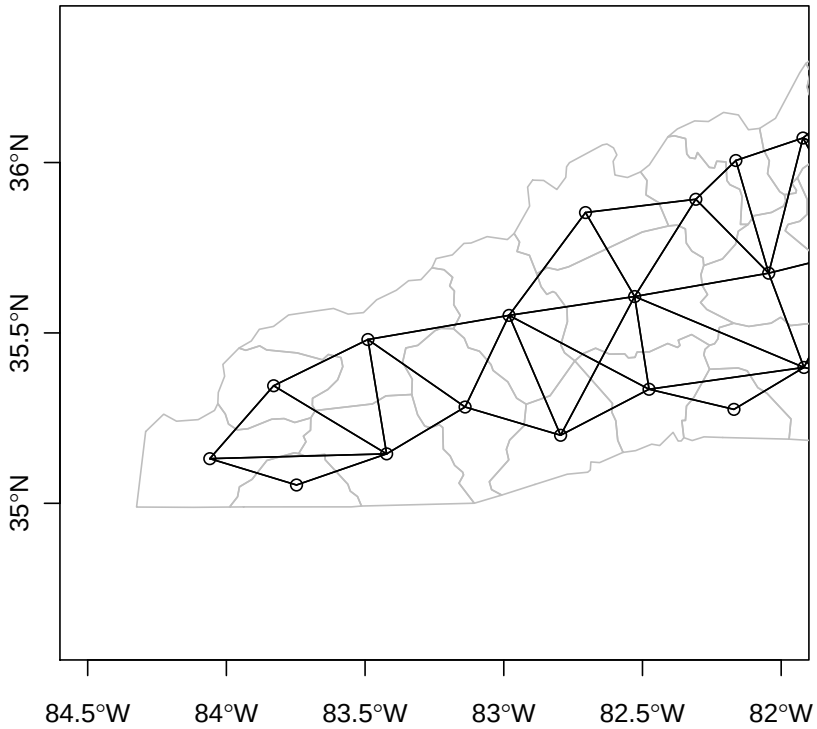
La classe nb

La fonction 'poly2nb' permet de construire une matrice de voisinage basé sur le principe suivant : les unités i et j sont voisines si elles partagent une frontière commune. Avec les données sids :

```
wc.nb=poly2nb(nc)
class(wc.nb)
is.symmetric.nb(wc.nb)
str(wc.nb)
plot(nc, border='grey',xlim=c(-84.5,-82),ylim=c(35,36),
axes=TRUE)
coord=coordinates(nc)
plot(wc.nb,coord,add=TRUE)
```

La classe nb

Voisinage basé sur les frontières communes pour les données SIDS



Matrices de contiguité

Plus généralement les matrices de contiguité sont basées sur le partage d'une frontière ou d'un sommet de polygone

1	2	3
4	0	5
6	7	8

- "rook" : au moins une frontière commune

0 voisin de 2, 7, 4, 5

- "bishop" : au moins un sommet commun

0 voisin de 1, 3, 6, 8

- "queen" : au moins une frontière ou un sommet commun

0 voisin de 1, 2 , 3, 4 , 5, 6, 7, 8

Dans spdep, les fonctions 'queencell' et 'rookcell' permettent de construire certaines de ces matrices pour des unités disposées sur une grille.

Matrices basées sur un seuil de distance

- $w_{ij} = 1(d(s_i, s_j) \leq seuil)$
- $w_{ij} = \frac{C}{d(s_i,s_j)^\alpha}$
- $w_{ij} = \exp(-\alpha d(s_i, s_j))$

remarque : dans certains cas, $d(s_i, s_j)$ peut être autre chose que la distance géographique, par exemple $d(s_i, s_j) = \mid x_i - x_j \mid$, où x_i désigne une caractéristique socio-économique pertinente.

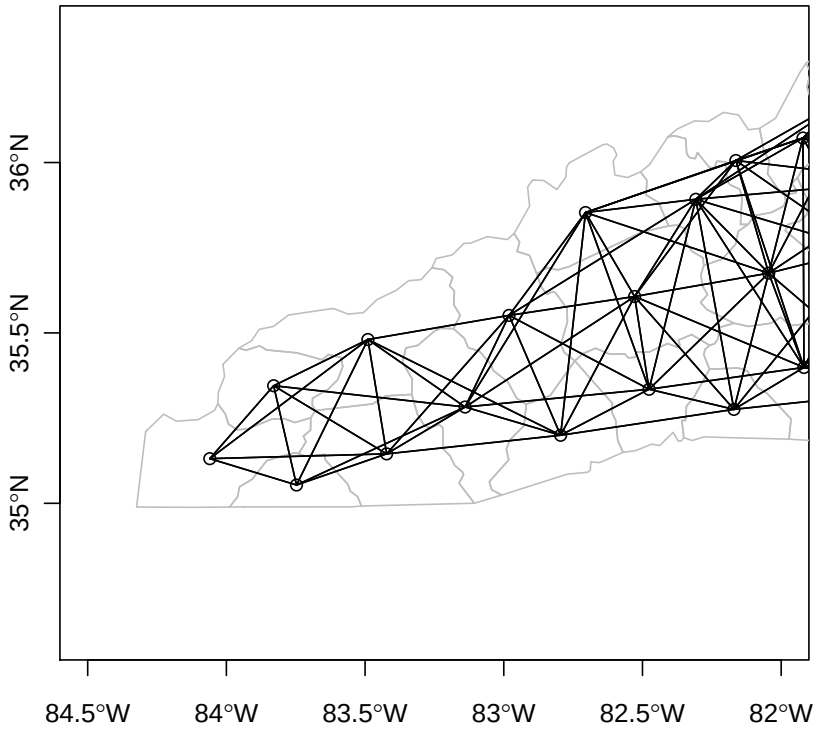
Matrices basées sur un seuil de distance

Pour les données SIDS, voici la contruction d'une matrice basée sur un seuil de 75km, l'option LONGLAT=TRUE permet d'utiliser une distance kilométrique alors que les coordonnées sont exprimées en degrés.

```
wd.nb=dnearneigh(coord,0,75,longlat=TRUE)
class(wd.nb)
plot(nc, border='grey',xlim=c(-84.5,-82),ylim=c(35,36),
axes=TRUE)
coord=coordinates(nc)
plot(wd.nb,coord,add=TRUE)
```

Matrices basées sur un seuil de distance

Voisinage basé sur un seuil de distance de 75km pour les données SIDS



Matrices basées sur un seuil de distance

Exemple de code pour construire une matrice basée sur l'inverse de la distance

```
wd.nb.2=dnearneigh(coord,0,1000,longlat=TRUE)
dlist <- nbdists(wd.nb.2, coord)
dlist <- lapply(dlist, function(x) 1/x)
wd.list<-nb2listw(wd.nb.2, glist=dlist)
```

Matrices basées sur un nombre de plus proches voisins

Matrice basée sur le plus proche voisin : $w_{ij} = 1$ si et seulement si s_j est le plus proche voisin de s_i.
Matrice basée sur les k plus proches voisins : étant donné un entier k, pour un site i, les indices j tels que $w_{ij} = 1$ sont ceux de son plus proche voisin, de son deuxième plus proche voisin, etc... jusqu'à son k-ième plus proche voisin.

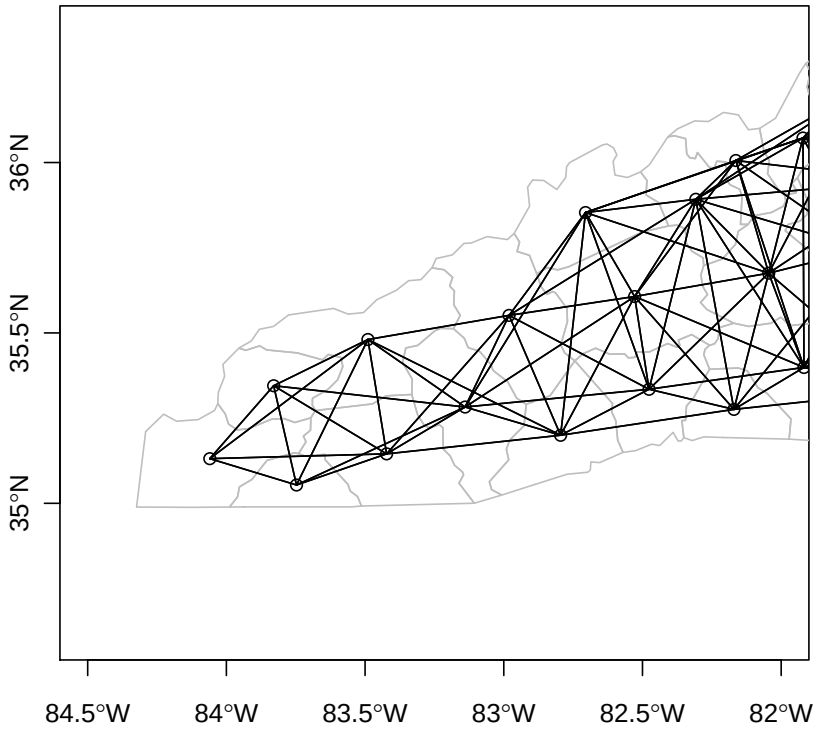
La classe knn

Exemple de code pour construire et représenter une matrice basée sur un nombre de plus proches voisins égal à 4

```
wv.knn=knearneigh(coord, k=4, longlat = TRUE)
class(wv.knn)
str(wv.knn)
plot(nc, border='grey',xlim=c(-84.5,-82),ylim=c(35,36),
axes=TRUE)
plot(knn2nb(wv.knn), coord, add=TRUE)
```

Matrices basées sur un nombre de plus proches voisins

Voisinage basé sur les quatre plus proches voisins pour les données SIDS



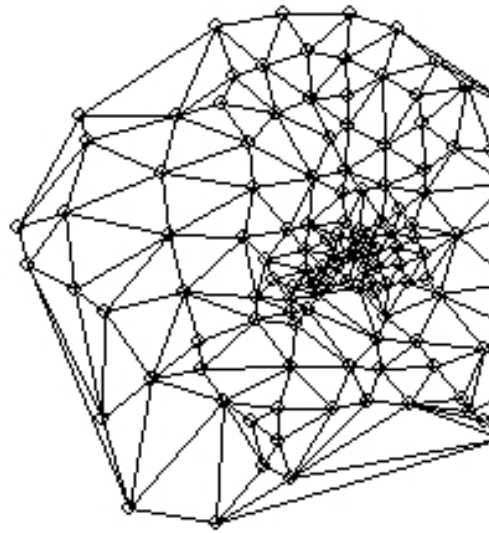
Variable spatialement décalée avec spdep

Ne pas faire de produit matriciel

```
nc$SID74_lag.B=lag.listw(nb2listw(knn2nb(wv.knn), style="B"),
nc$SID74)
nc$SID74_lag.W=lag.listw(nb2listw(knn2nb(wv.knn), style="W"),
nc$SID74)
```

Matrices basées sur triangulation de Delaunay

Triangulation de Delaunay : unique triangulation telle que le cercle circonscrit à trois sommets quelconques ne contient aucun autre sommet.
Permet de construire une matrice : deux sites sont voisins si le segment les



Matrices basées sur triangulation de Delaunay : syntaxe

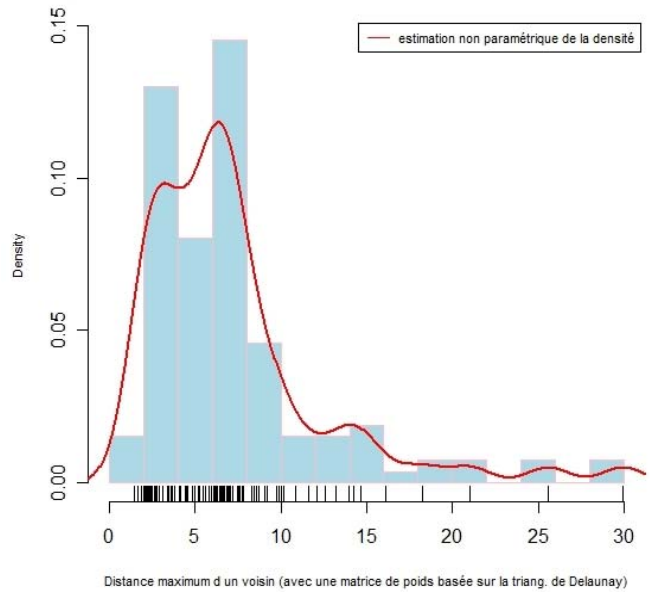
```
w.tri=tri2nb(coord)
class(w.tri)
plot(nc, border='grey',axes=TRUE)
plot(w.tri, coord, add=TRUE)
```

Analyse de la matrice de Delaunay

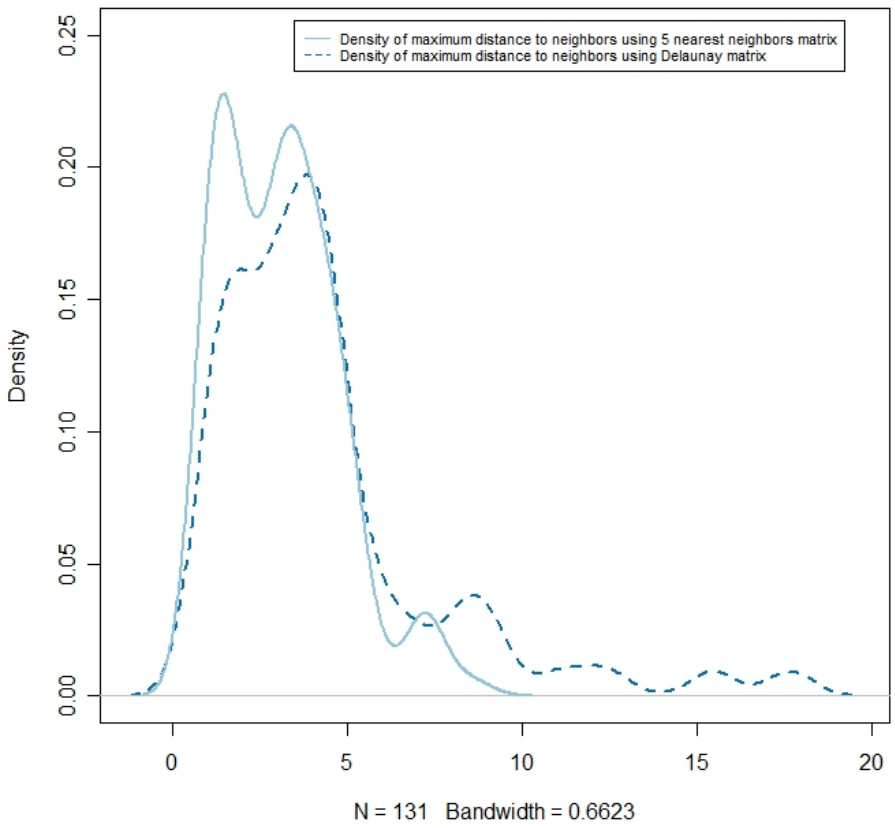
```
summary(w.tri)
```

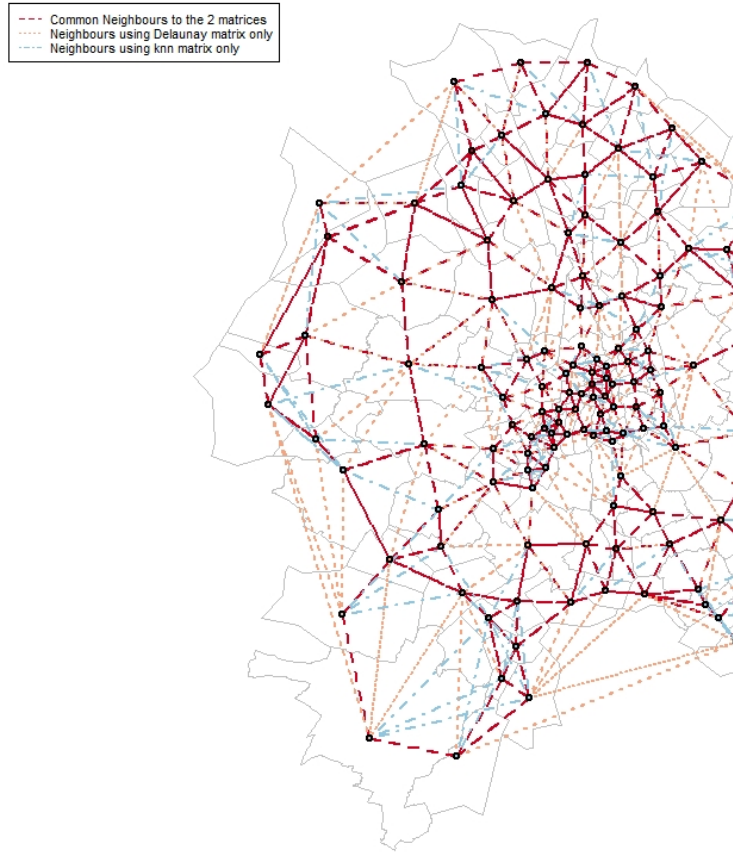
```
Neighbour list object:
Number of regions: 131
Number of nonzero links: 752
Percentage nonzero weights: 4.382029
Average number of links: 5.740458
Link number distribution:

 3  4  5  6  7  8  9
 1 15 43 38 27  6  1
1 least connected region:
68 with 3 links
1 most connected region:
65 with 9 links
```



Comparaison avec matrice des 5 plus proches voisins





La classe listw

Lorsqu'on utilise une matrice de poids dans un modèle de régression spatial, on a besoin de la mettre au format listw. Il faut alors préciser les options style="B" ou style= "W"

- B matrice binaire
- W normalisation des lignes

Pour extraire la partie nb d'un objet listw, on utilisera la commande

```
$neighbours
```

. La commande

```
print
```

appliquée à un objet de type listw donne aussi des statistique utiles.

La classe listw : petit exemple

```
t=c(1,2,3,4)
u=c(3,2,5,1)
plot(t,u)
co=cbind(t,u)
W.knn=knearneigh(co,k=2,longlat=TRUE)
plot(knn2nb(W.knn), co, add=TRUE)
W.nb=knn2nb(W.knn)
W.listw1=nb2listw(W.nb,style="B")
str(W.listw1)
W.listw1$neighbours[]
W.listw1$weights[]
W.listw2=nb2listw(W.nb,style="W")
W.listw2$neighbours[]
W.listw2$weights[]
```

Indice de Moran : définition

Pour une matrice de poids W vérifiant $w_{ii} = 0$ et un champ $X_{s_i} = X_i, i = 1, \ldots n$, le **"I" de Moran** est défini par :

$$I = \frac{n}{1'W1} \frac{X'WX}{X'X} = \frac{\frac{\sum_{i,j} w_{ij}(X_i - \bar{X})(X_j - \bar{X})}{\sum_{i,j} w_{ij}}}{\frac{\sum_i (X_i - \bar{X})^2}{n}}$$

C'est le rapport d'une sorte de covariance entre unités contigues à la variance $\mapsto$ sorte de coefficient d'autocorrélation.

Indice de Moran : propriétés

- I est indépendant des unités dans lesquelles X est exprimé
- I est invariant à une symétrisation de la matrice W, (i.e. $W \longrightarrow (W + W')/2$)
- Attention : le **I** de Moran dépend du choix de la matrice W, et peut être affecté par le niveau d'aggrégation (effet d'échelle) ainsi que par la forme des unités spatiales

Indice de Moran : interprétation

Si X est centré, les valeurs de X de même signe et géographiquement proches contribuent positivement à **I**.

- les valeurs positives et fortes de **I** indiquent une autocorrélation spatiale positive
- les valeurs négatives et fortes de **I** indiquent une autocorrélation spatiale négative
- les valeurs proches de 0 indiquent une absence d'autocorrélation

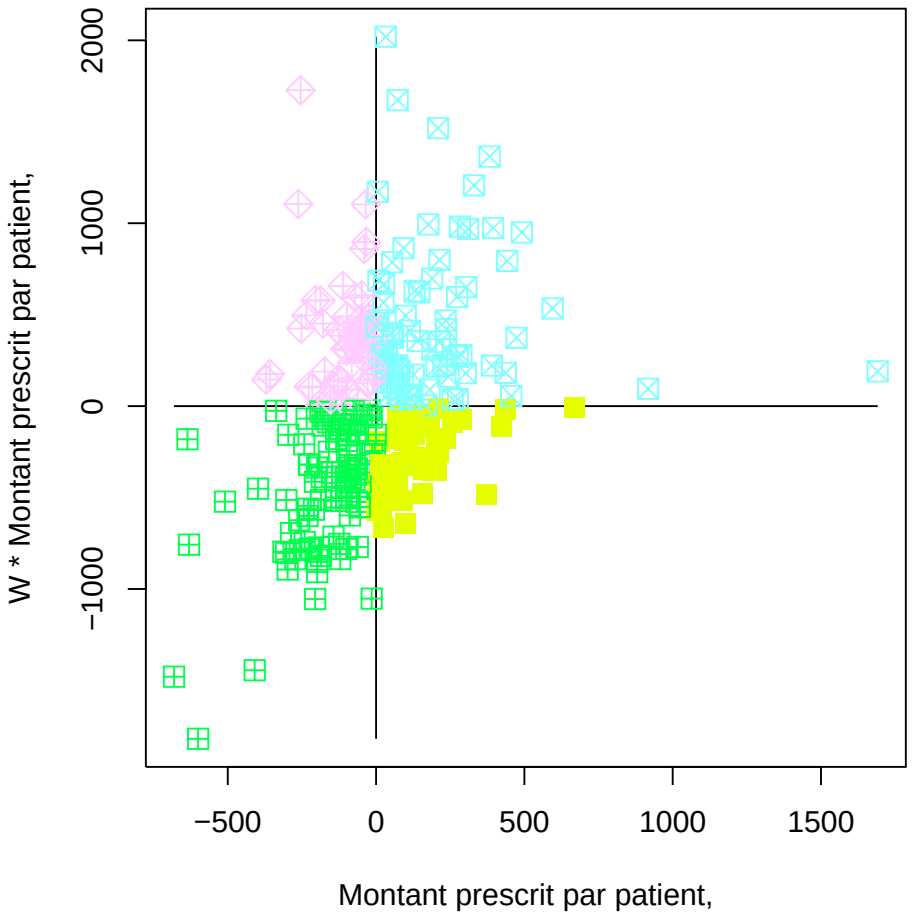
Moran scatterplot

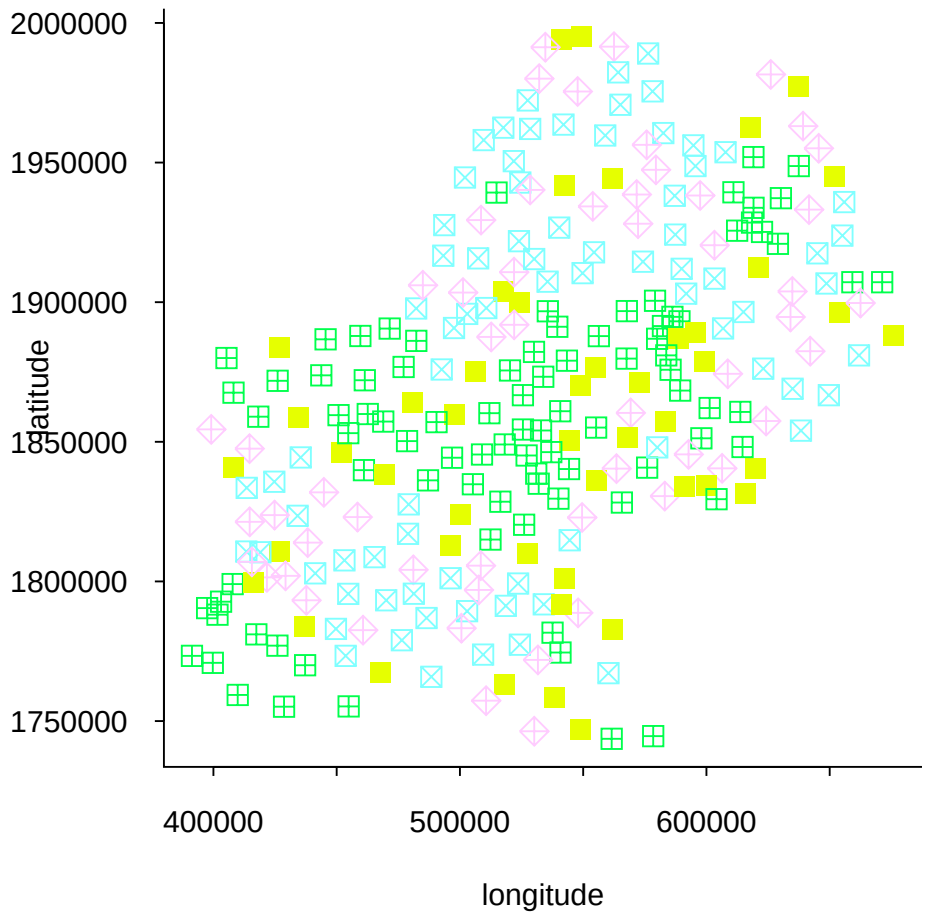
Le **"Moran scatterplot"** est un nuage de points de WX contre X, où X est centrée et W normalisée.
Les deux propriétés X centrée et W normalisée impliquent que la moyenne empirique de WX est égale à $\bar{X}$ et donc à 0.
On peut superposer au nuage la droite de régression qui passe donc par l'origine. La pente de celle-ci est égale à l'indice de Moran.
Utilisation :

- détecter des points aberrants
- aprécier le degré d'autocorrélation
- non linéarité $\mapsto$ plusieurs régimes d'association spatiale.

Remarque : il est intéressant de normaliser X avant de faire le graphique pour pouvoir ainsi comparer plusieurs moran plots entre eux.

Moran scatterplot : exemple





Le C de Geary

Le **C** de Geary est défini par

$$C = \frac{n-1}{2\sum_{i,j} w_{ij}} \frac{\sum_{i,j} w_{ij}(X_{s_i} - X_{s_j})^2}{\sum_i (X_{s_i} - \bar{X})^2}$$

Les valeurs faibles de **C** indiquent une autocorrélation spatiale positive et les valeurs fortes de **C** une autocorrélation spatiale négative.
C est indépendant des unités dans lesquelles X est exprimé.
L' indice de Geary ressemble à la statistique de Durbin Watson en séries temporelles. Pour comparaison, la statistique de Durbin-Watson pour une série temporelle centrée est donnée par
$DW = \sum_{t=2}^{n}(X_t - X_{t-1})^2 / \sum_{t=1}^{n} X_t^2$.

Lien entre l'indice de Moran et l'indice de Geary

$$
\begin{aligned}
G &= \frac{n-1}{2n}\left[\frac{\sum_{i\neq j} w_{ij}(X_i-\bar{X})^2 + \sum_{i\neq j} w_{ij}(X_j-\bar{X})^2}{\sum_i (X_i-\bar{X})^2}\right] \\
&- \frac{n-1}{n} I
\end{aligned}
$$

Mesures locales d'autocorrélation spatiale (LISA)

Sous les mêmes conditions que pour le I de Moran global, pour un site i, on définit un indice de Moran local par :

$$I_i = (X_i - \bar{X}) \sum_{j \neq i} w_{ij}(X_j - \bar{X})$$

Le numérateur du I de Moran global est alors la somme des I_i. Si de plus le champ est standardisé, alors le I de Moran global est la moyenne des I_i.

LISA : valeurs extrêmes

Les valeurs extrêmes de I_i indiquent une agglomération locale de valeurs semblables : on considère que seules les valeurs éloignées de plus de deux écarts types sont interprétables.
Si l'autocorrélation globale est positive, on distingue les cas suivants :

- I de Moran local élevé et positif : agrégat local de valeurs extrêmes avec voisins similaires ; on parle de "Hot- spot" si dans quadrant supérieur droit et "Cold spot" si dans le quadrant inférieur gauche
- I de Moran local élevé et négatif : "High-Low" (resp : Low-High) valeurs basses avec valeurs voisines similaires et fortes (resp : valeurs fortes avec valeurs voisines similaires et basses) : ces deux dernières catégories correspondent à des atypiques locaux.

Statistiques "join counts" pour variable dichotomique

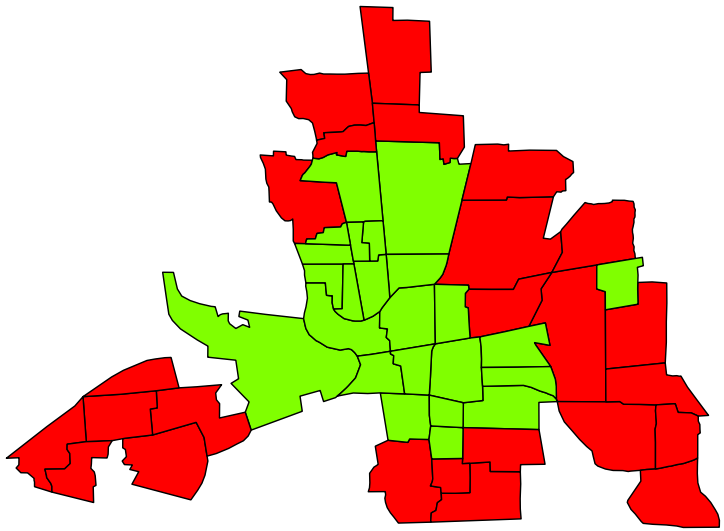
Si X_i a deux modalités 0 et 1 avec : $P(X_i = 1) = p$, on introduit les statistiques suivantes appellées "join counts"

$$BB = \frac{1}{2} \sum_{i,j} w_{ij} X_i X_j$$

$$BW = \frac{1}{2} \sum_{i,j} w_{ij} (X_i - X_j)^2$$

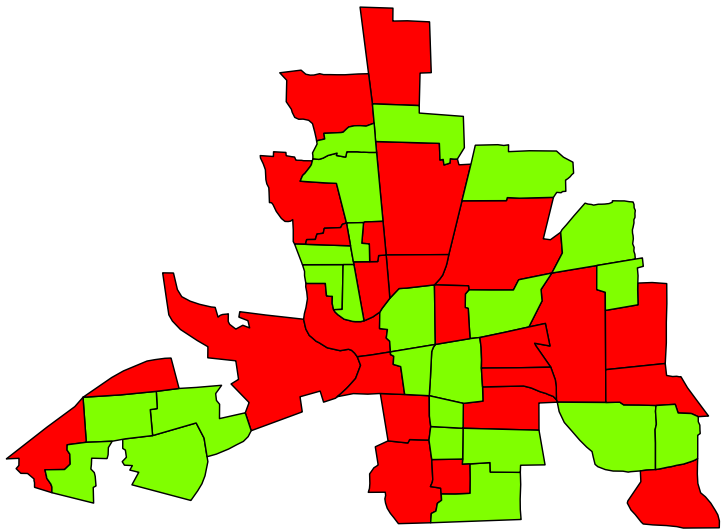
Statistiques "join counts" : exemple

```
joincount.multi(HICRIME, list)
          Joincount Expected Variance z-value
low:low      34.000   29.337   18.638  1.0802
high:high    52.000   26.990   17.648  5.9534
high:low     29.000   58.673   26.041 -5.8149
```



Statistiques "join counts" : exemple

```
 s.HICRIME<-sample(HICRIME)
joincount.multi(s.HICRIME, list)
          Joincount Expected Variance z-value
low:low      27.000   29.337   18.638 -0.5413
high:high    24.000   26.990   17.648 -0.7117
high:low     64.000   58.673   26.041  1.0438
```



Les domaines d'application des semis de points

Domaines classiques : épidémiologie, écologie, foresterie.
Quelques exemples

- la disposition de certaines espèces végétales dans une forêt,
- les adresses de patients affectés d'une certaine maladie dans une région,
- la répartition de cellules dans un tissu biologique,
- les emplacements des épicentres de secousses sismiques enregistrées,
- la localisation de trésors archéologiques retrouvés sur un site

Phénomènes à modéliser

Répartition alétoire de points dans $\mathbb{R}^2$ avec un nombre de points aléatoire.
Inhomogénéité spatiale. Des zones ont en moyenne plus de points que les autres.
Interaction spatiale. La compétition pour la nourriture ou l'espace peut engendrer de la répulsion entre les points. Au contraire, si l'on observe l'occurence de maladies épidémiques, on va avoir de l'aggrégation.
Difficulté. une seule réalisation $\Rightarrow$ confusion entre hétérogénéité et interaction.
Des agrégats apparents peuvent être engendrés soit par une inhomogénéité spatiale soit par de l'interaction entre les points.
Questions classiques : tester l'hypothèse CSR, détecter régularité ou agrégation, ajuster un modèle, détecter agrégats.

Modèle mathématique : processus ponctuel

Configuration de n points de $E \subset \mathbb{R}^2$: ensemble de n points non ordonné $x = \{x_1, \cdots, x_n\}$.
Espace des configurations (ou **espace exponentiel**) : l'espace N_{lf} des sous-ensembles x localement finis de E, c'est à dire tels que le nombre de points de x contenus dans tout borné de E est fini, muni d'une tribu $\mathcal{N}_{lf}$.
Tribu $\mathcal{N}_{lf}$ sur N_{lf} : exemples d'évenements "il y a au plus 50 points dans la configuration", "les points de x sont distants d'au moins r", "il n'y a aucun point dans B", etc.
Un processus ponctuel est dit **marqué** lorsqu'à chaque position est associée une variable aléatoire dite "marque" : par exemple, s'il s'agit d'arbres dans une forêt, la marque peut être la taille ou le diamètre de l'arbre.

Modèle mathématique : processus ponctuel

Deux définitions pour un processus ponctuel X :

- une variable aléatoire à valeurs dans l'espace N_{lf} muni de $\mathcal{N}_{lf}$
- un ensemble aléatoire X de points X_i de E tel que le nombre de points de E tombant dans A soit une variable aléatoire finie, pour tout borélien borné A de E.

Un PP X est **simple** si le nombre de points de E tombant dans $\{x\}$ pour tout $x \in E$ est presque surement égal à 0 ou 1.

Loi d'un processus ponctuel

La loi de probabilité induite sur N_{lf} muni de $\mathcal{N}_{lf}$ est la loi de X.
Pour un borélien B de $\mathbb{R}^2$, on notera $N_X(B) = \sum_{x_i \in x} \mathbf{1}(x_i \in B)$ le nombre de points d'une configuration appartenant à B : pour tout B, $N(B)$ est une variable aléatoire.
La loi d'un processus ponctuel est définie par les probabilités $\mathbb{P}(X \in Y)$, pour tout $Y \in \mathcal{N}_{lf}$: cette famille contient en particulier la famille des probabilités fini-dimensionnelles $\mathbb{P}(N_X(B_1) = n_1, \ldots, N_X(B_k) = n_k)$ qui caractérisent entièrement la loi.

Loi d'un processus ponctuel

Nous adopterons ici une approche plus commode pour les applications (E borné) consistant à définir une densité jointe pour les variables N, nombre de points, et $X_1, \ldots, X_n$, localisations des N points (Cressie, 1993, p.622) : $f\big((x_1, \cdots, x_n), n\big)$. On a alors

$$\sum_{n=0}^{\infty} \int_{E^n} f\big((s_1, \cdots, s_n), n\big) ds_1 \cdots ds_n = 1.$$

De façon équivalente, on se donne

- la famille des probabilités $p_n = \mathbb{P}(N_X(E) = n)$, pour $n \geq 0$
- les densités g_n sur E^n des configurations à n points (invariantes par permutation)

Un exemple : le processus de Poisson homogène

Le processus de Poisson homogène PPP(λ) est le modèle de base en théorie des processus ponctuels car il formalise le concept de points répartis au hasard. Il est défini par les deux conditions suivantes pour E borélien borné :

- (i) il existe un réel $\lambda > 0$ tel que pour tout borélien A de $\mathbb{R}^2$, $N_X(A)$ suit une loi de Poisson de moyenne $\lambda \mid A \mid$, où $\mid A \mid$ désigne l'aire de A.
- (ii) sachant que $N_X(A) = n$, les n points du processus qui sont dans A forment un échantillon de la loi uniforme sur A.

Propriétés du Poisson homogène

Les deux conditions (i) et (ii) impliquent la condition (iii) suivante :
pour deux boréliens A et B, les variables aléatoires $N_X(A)$ et $N_X(B)$ sont indépendantes.
Le processus de Poisson homogène est stationnaire et isotrope.
On démontre que les probabilités fini-dimensionnelles de ce processus sont données par

$$\mathbb{P} \quad (N_X(B_1) = n_1, \ldots, N_X(B_k) = n_k) = \frac{\lambda^{n_1+\ldots+n_k} \mid B_1 \mid^{n_1} \ldots \mid B_k \mid^{n_k}}{n_1! \ldots n_k!} \exp(-\sum_{i=1}^{k} \lambda \mid B_k \mid).$$

Propriétés du Poisson homogène

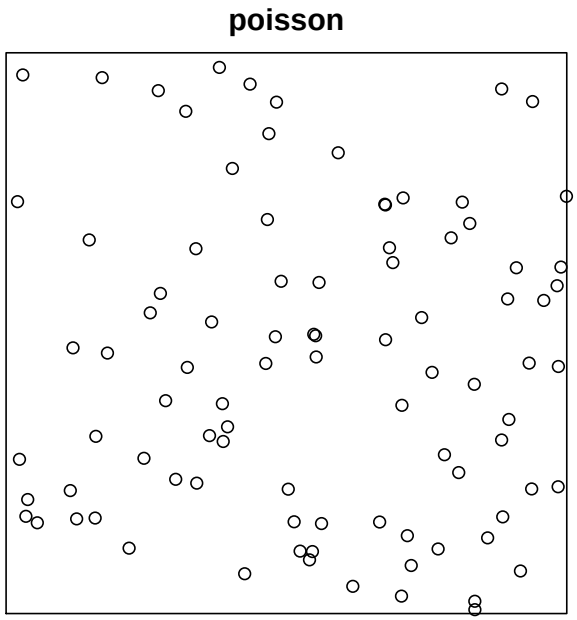
Soit A un borélien borné de E, conditionnellement à $\{N_X(A) = n\}$, les points $X_1, \cdots, X_n$ de X dans A sont indépendants et uniformément identiquement distribués

$$\mathbb{P}(X_i \in B) = \lambda \frac{aire(B)}{aire(A)},$$

pour tout borélien B inclus dans A.

Simulation d'un PPP homogène

```
window=owin(c(0,10),c(0,10))
poisson=rpoispp(1,win=window)
plot(poisson)
```



L'hypothèse CSR

CSR : complete spatial randomness

L'hypothèse CSR pour un PP est l'hypthèse que le PP est un processus de Poisson homogène. Elle contient donc deux sous-hypothèses :

- l'homogénéité de l'intensité
- l'absence d'interaction

Intensité

L'intensité est l'analogue pour le processus ponctuel de l'espérance pour une variable aléatoire.
La mesure d'intensité est une mesure sur les boréliens B de $\mathbb{R}^2$ vérifiant

$$\Lambda(B) = \mathbb{E}(N(B)),$$

de façon que $\Lambda(B)$ représente le nombre moyen de points du processus dans B.
Si le processus est stationnaire, cette mesure est proportionnelle à la mesure de Lebesgue et le facteur de proportionalité, λ, appellé intensité, représente le nombre moyen de points du processus par unité de surface.

Fonction d'intensité

Plus généralement, si Λ est absolument continue par rapport à la mesure de Lebesgue, il existe une fonction λ localement intégrable définie sur E telle que pour tout borélien B,

$$\Lambda(B) = \int_B \lambda(x)dx.$$

Cette fonction λ porte le nom de fonction d'intensité du processus ponctuel.

Intensité et stationnarité

Si le processus est stationnaire, la fonction d' intensité est constante. Inversement, si le fonction d'intensité est constante, le processus est dit stationnaire au premier ordre ou homogène (sinon, il est dit inhomogène).

Dans le cas du processus de Poisson homogène, la fonction d'intensité est constante égale au paramètre λ de la définition.

Estimation de l'intensité - cas homogène

Dans le cas d'un processus homogène d'intensité λ, un estimateur sans biais de l'intensité est donné par

$$\hat{\lambda} = \frac{N}{|\ W\ |},$$

où W est la fenêtre d'observation et $N = N(W)$ le nombre de points observés dans cette fenêtre. Il coincide en fait avec l'estimateur du maximum de vraisemblance dans le cas où le processus est un Poisson homogène.

```
    > summary(poisson)
Planar point pattern: 91 points
Average intensity 0.91 points per square unit

Window: rectangle = [0, 10]x[0, 10]units
Window area =  100 square units
```

Le processus de Poisson inhomogène

Le processus de Poisson homogène ayant une intensité constante ne peut servir à modéliser des phénomènes présentant une forte hétérogénéité spatiale.
Etant donné une fonction d'intensité λ, on peut définir le processus de Poisson X de mesure d'intensité λdx par les deux conditions suivantes

- (i) le nombre de points $N(A)$ de X dans tout borélien A de $\mathbb{R}^2$, suit une loi de Poisson de moyenne $\lambda(A)$,
- (ii) les nombres de points de X dans k boréliens $A_1, \ldots, A_k$ disjoints de $\mathbb{R}^2$ sont k variables aléatoires indépendantes.

Le processus de Poisson inhomogène

Ainsi défini, ce processus n'est pas stationnaire sauf si l'intensité est constante.
Conditionnellement à $N = n$, les n points $X_1, \ldots, X_n$ sont alors i.i.d..
Il existe une relation directe entre la fonction d'intensité du processus ponctuel, $\lambda(.)$, et la densité d-dimensionnelle $f(.)$ de toute localisation X_i conditionnellement à N :

$$\forall s \in E,\ f(s) = \frac{\lambda(s)}{\int_E \lambda(s)\nu(ds)}.$$

Simulation d'un PPP inhomogène

Pour simuler un processus de Poisson inhomogène dans spatstat, on utilise à nouveau la fonction rpoispp, en précisant en input l'intensité soit comme une fonction des coordonnées soit comme une image (le deuxième argument précise le maximum de l'intensité).

```
window=owin(c(0,10),c(0,10))
poisson_inhom=rpoispp(function(x,y){10*exp(-3*x)+10*exp(-3*y)}
,20,win=window)
plot(poisson_inhom)
poisson_inhom2=rpoispp(Z)
plot(poisson_inhom2)
```

Estimation de l'intensité - cas inhomogène

Dans le cas inhomogène, on peut utiliser un estimateur non paramétrique, introduit par Diggle (1985)donné par

$$\hat{\lambda}_h(s) = \frac{\sum_{i=1}^{N} h^{-d} K\left(\frac{s-X_i}{h}\right)}{\int_E h^{-d} K\left(\frac{s-u}{h}\right) du} \tag{1}$$

où le dénominateur est un terme de correction au bord nécessaire lorsque le domaine d'observation est limité et où K est une fonction noyau.

Estimation de l'intensité - cas inhomogène

L' estimateur de Diggle est, de même qu'un estimateur non paramétrique de densité, peu sensible au choix du noyau K. Le choix de la largeur de bande ou fenêtre h permettant de minimiser l'erreur quadratique moyenne intégrée

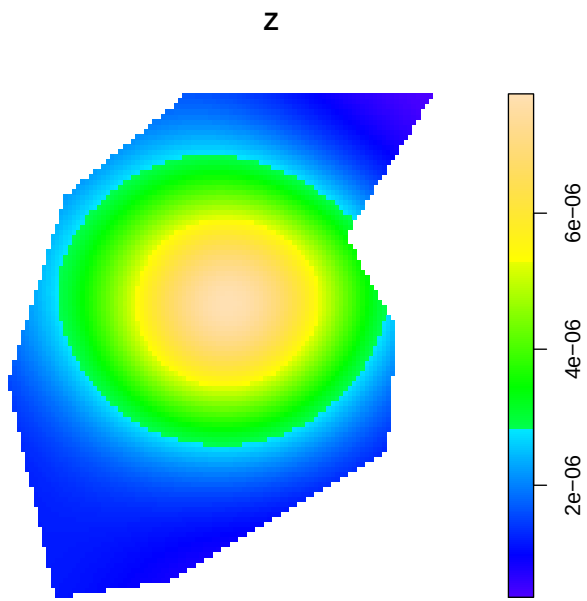
$$EQMI(h) = \mathbb{E}\{ \int_E (\hat{\lambda}_h(s) - \lambda(s))^2 ds \}$$

se fait selon des méthodes similaires au cas de l'estimation de densité.

Estimation de l'intensité - cas inhomogène

Dans spatstat, on peut évaluer cet estimateur avec un noyau gaussien par la fonction density.ppp, l'output est alors une image de classe im que l'on peut représenter avec plot. Avec les données Pompiers

```
h=5000
Z=density.ppp(PP,h, edge=TRUE)
plot(Z)
```



Interaction spatiale

Du fait de la propriété (ii), le processus de Poisson implique une **absence d'interaction** entre les évènements.
Les caractéristiques du second ordre vont permettre de mettre en évidence deux autres types de comportement. On distingue d'une part

- les processus pour lesquels les évènements ont tendance à s'attirer : **aggrégation**
- ceux pour lesquels les évènements ont tendance à se repousser : **régularité**.

Distance d'un point courant au plus proche voisin

Soit x un point de E qui ne figure pas nécéssairement dans une configuration du PP X.
Pour un processus ponctuel X homogène, on définit

$$F_x(r) = \mathbb{P}(d(x, \{x_1, \cdots, x_n\} \setminus \{x\}) \leq r).$$

Notons qu'en raison de l'homogénéité F_x ne dépends pas de x, c'est pourquoi nous le noterons plus simplement F.

Fonction F

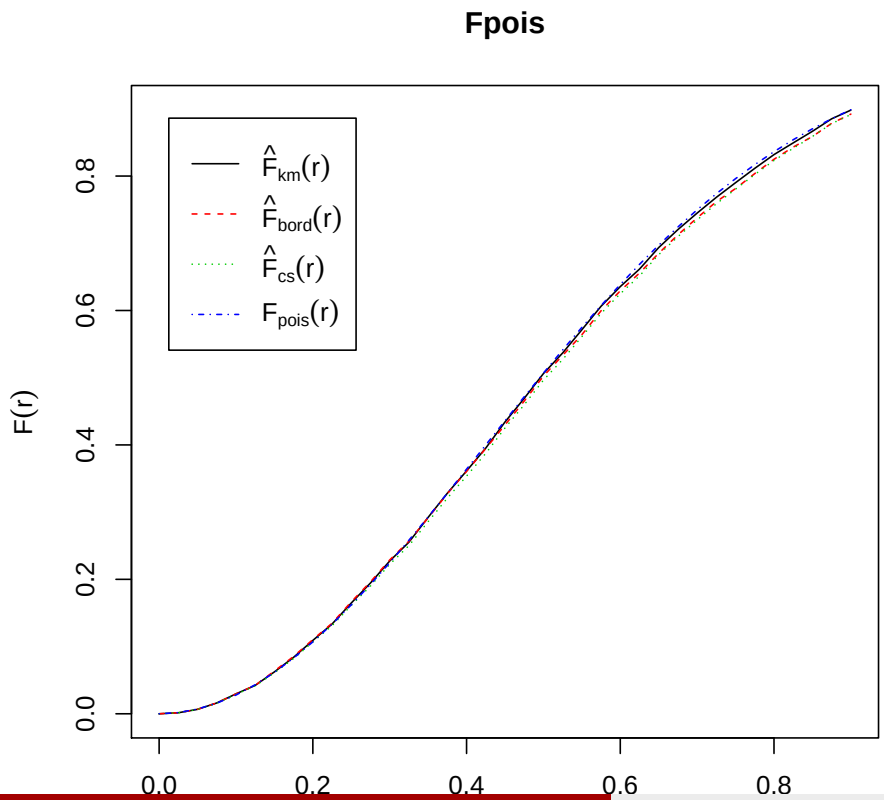
F est la fonction de répartition de la distance au plus proche voisin et peut aussi s'interpréter comme la mesure de "l'espace vide" (c'est pourquoi on l'appelle "empty space function" en anglais) dans le sens suivant :
$1 - F(r)$ est la probabilité qu'une boule de centre 0 (ou un quelconque point de E fixé) et de rayon r ne contienne aucun point de X.

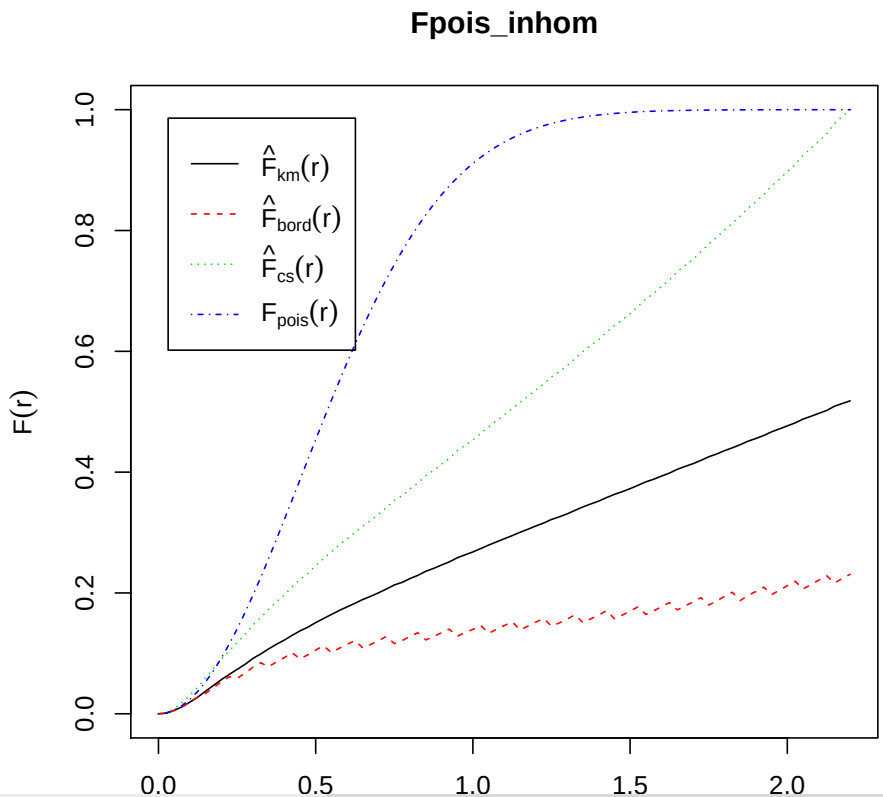
Sous l'hypothèse CSR d'homogénéité spatiale sur $\mathbb{R}^2$, la fonction F a la forme analytique suivante pour $x > 0$

$$F(x) = 1 - \exp(-\pi\lambda x^2).$$

Estimateur de F

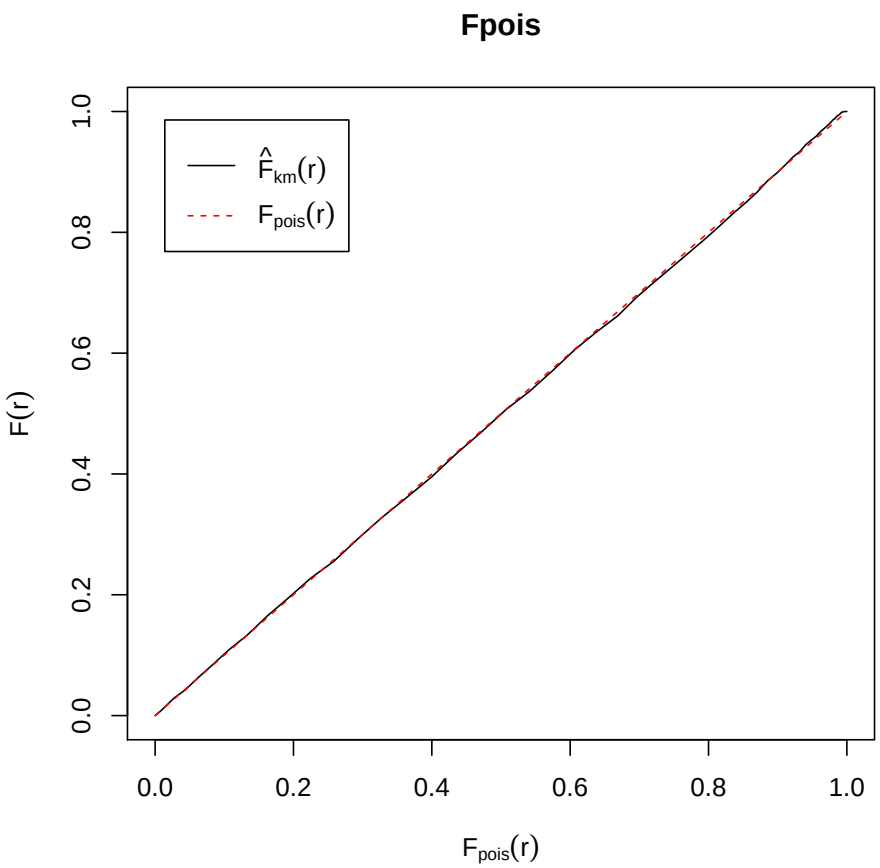
Pour estimer F, on utilise en général une grille fine de points définie sur E qui permet d'approximer les distances au plus proche voisin. A gauche, exemple simulé Poisson homogène, à droite, exemple simulé Poisson inhomogène.

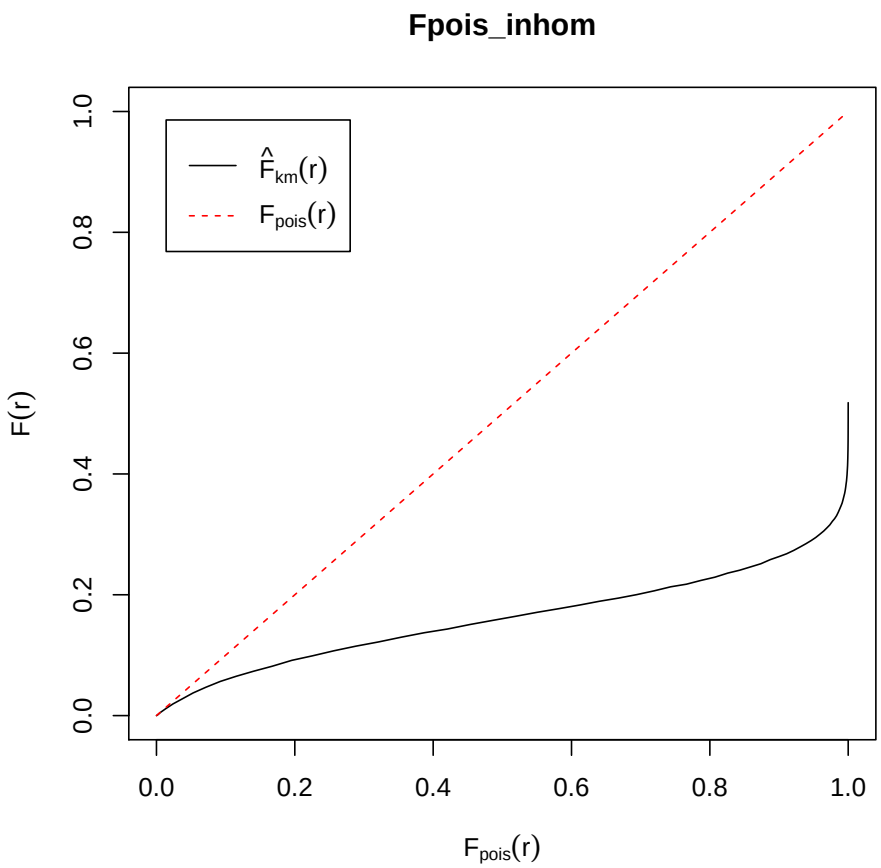




Estimateur de F : graphique alternatif

Sous forme de probability-probability plot :





Distance d'un point du PP à son plus proche voisin : fonction G

Si cette fois, on s'intéresse à la distance entre un point du PP et son plus proche voisin, on définit la fonction de répartition de ces distances G par

$$G(r) = \mathbb{P}(d(x, \{x_1, \cdots, x_n\} \setminus \{x\}) \leq r \mid x \in X).$$

Un estimateur classique de G est donné par la fonction de répartition empirique définie par

$$\hat{G}(r) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}(d(x_i, x_{j(i)}) \leq r),$$

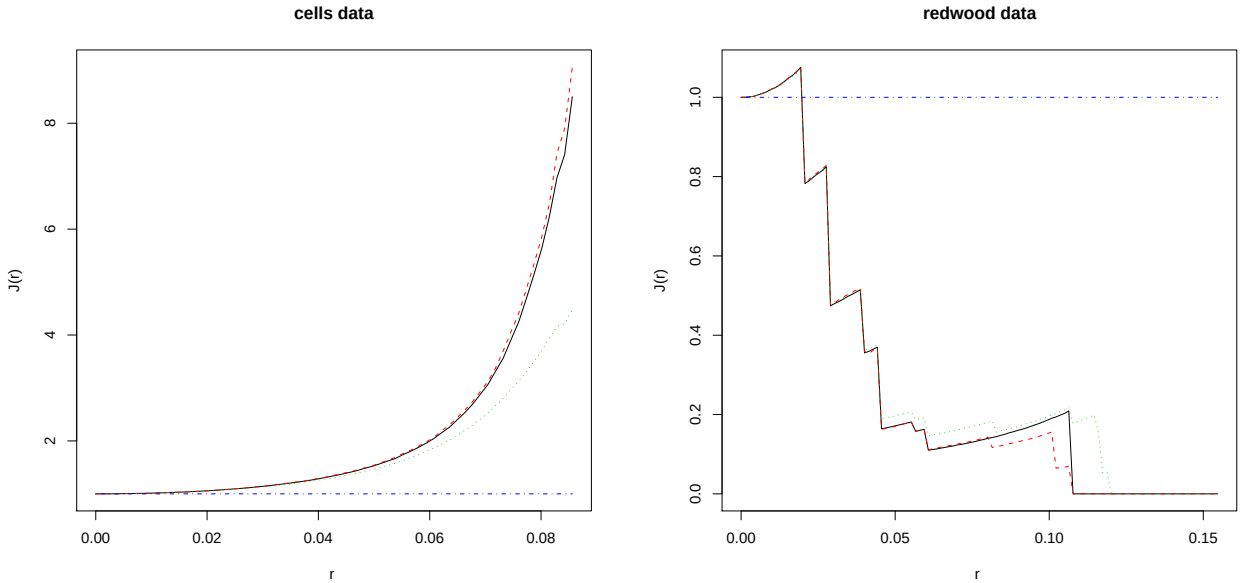
où $x_{j(i)}$ est le point de X le plus proche de x_i.

Fonction J

A partir de F et G, on peut définir la fonction J par

$$J(r) = \frac{1 - G(r)}{1 - F(r)}.$$

$J = 1$ correspond au cas d'un processus poissonnien.
$J > 1$ indique une tendance à la régularité et $J < 1$ à l'aggrégation.



Moment factoriel d'ordre 2

De même que l'on a introduit la mesure d'intensité pour le moment d'ordre 1, le rôle du moment d'ordre 2 est joué par la mesure de moment factoriel d'ordre 2, donnée pour tous boréliens B_1 et B_2 de $\mathbb{R}^2$ par

$$\alpha_2(B_1 \times B_2) = \mathbb{E}(N(B_1)N(B_2)) - \Lambda(B_1 \cap B_2).$$

Lorsque cette mesure est absolument continue par rapport à la mesure de Lebesgue, on note ρ_2 sa densité, appellée densité d'intensité d'ordre 2. Pour un PP stationnaire, la fonction $\rho_2(x, y)$ ne dépends que de $x - y$. Si de plus le PP est isotrope, elle ne dépends que de $\| x - y \|$.

Fonction de corrélation des paires

A partir de ρ_2, on définit la **fonction de corrélation des paires** g par

$$g(x,y) = \frac{\rho_2(x,y)}{\lambda(x)\lambda(y)}.$$

Fonction g et interaction :
Pour un PP de Poisson, on a $g(x,y) = 1$.
Si $g(x,y) > 1$, cela indique que pour ce PP, il est plus probable d'observer un couple de points en x et y que pour un PP de Poisson ayant la même intensité.
Si le PP est stationnaire et isotrope, g est une fonction de $r = \| x - y \|$; $g(r) > 1$ indique une tendance à l'aggrégation pour des points à distance r, et inversement, $g(r) < 1$ indique une tendance à la répulsion pour des points à distance r.

Fonction K de Ripley

Une façon alternative de caractériser les propriétés du second ordre est au travers de la fonction K de Ripley et de la fonction L qui lui est associée. Pour un PP stationnaire, introduisons la mesure κ, appellée mesure des moments réduits d'ordre deux, pour un borélien B par

$$\kappa(B) = \frac{1}{\lambda^2} \int_B \rho_2(x) dx.$$

Si de plus le PP est isotrope, en prenant pour B une boule $B(0, r)$ de centre l'origine et de rayon r, la fonction K de Ripley est définie par

$$K(r) = \kappa(B(0, r)).$$

$K(r)$ peut aussi s'interpréter comme le nombre moyen de points du PP dans une boule centrée en un des points du PP, horsmis le centre lui-même.

Fonction L

Pour un PP de Poisson homogène, $K(r) = \pi r^2$ et ceci engendre une autre méthode de comparaison avec un modèle de Poisson.
Pour faciliter la comparaison et aussi pour réduire la variance, il est d'usage de renormaliser la fonction K en définissant la fonction L par

$$L(r) = (\frac{K(r)}{\pi})^{1/2}.$$

Pour le PP de Poisson homogène, la fonction L est donc égale à r.
Lorsque $L(r) - r > 0$, cela indique un phénomène d'aggrégation pour des distances inférieures ou égales à r, et lorsque $L(r) - r > 0$, cela indique un phénomène de régularité pour des distances inférieures ou égales à r.

Relations entre g, ρ_2 et K

Pour un PP stationnaire et isotrope, les relations suivantes existent entre g, ρ_2 et K :

$$g(r) = \frac{\rho_2(r)}{\lambda^2} = \frac{K'(r)}{2\pi r}$$

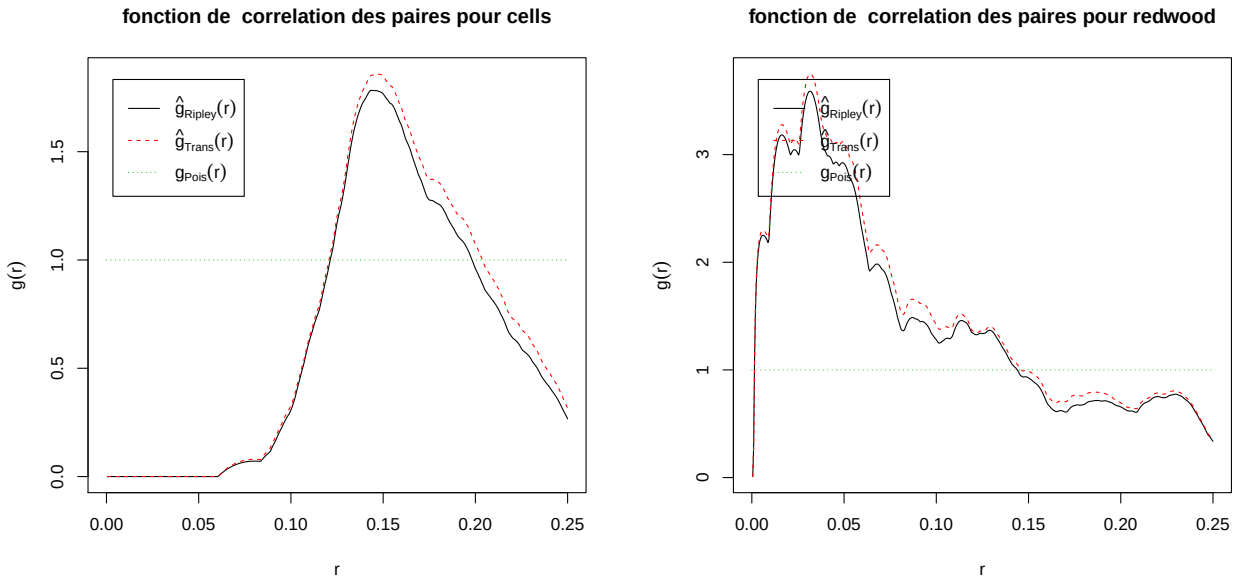
$$K(r) = \frac{2\pi}{\lambda^2} \int_0^r u\rho_2(u)du.$$

Estimateur de g

Pour estimer g, on peut commencer par estimer ρ_2 par un estimateur à noyau de la densité incluant une correction de bord (diverses corrections existent).
On peut alors en déduire un estimateur de la fonction de corrélation des paires en divisant par $\hat{\lambda}(x)\hat{\lambda}(y)$, où $\hat{\lambda}$ est l'estimateur de Diggle de l'intensité.

Estimateur de g

La figure suivante montre un estimateur de la fonction de corrélation des paires pour les données cells à gauche et redwood à droite.



Estimateur de K

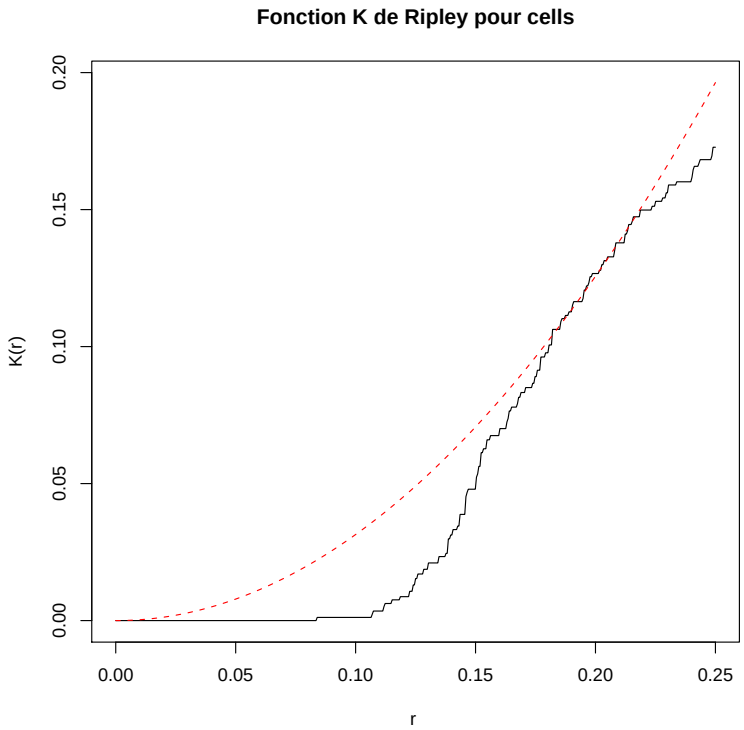
On peut estimer directement la fonction K par

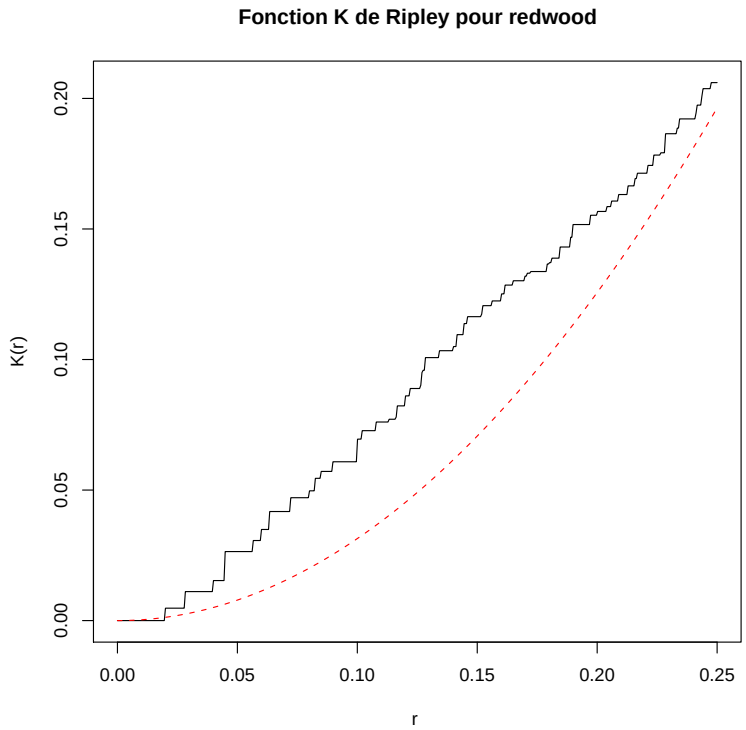
$$\hat{K}(r) = \sum_{x \in X, y \in W_{\ominus r}} \frac{1(x - y \in B(0, r))}{\hat{\lambda}(x)\hat{\lambda}(y)},$$

où $W_{\ominus r}$ désigne l'ensemble des points de la fenêtre W tels que la boule centré en ce point et de rayon r soit entièrement incluse dans W.
D'autres formules existent mais consistent essentiellement à faire d'autres corrections de bord.
Notons que les relations entre g et K permettent aussi de déduire un estimateur de g à partir d'un estimateur de K.

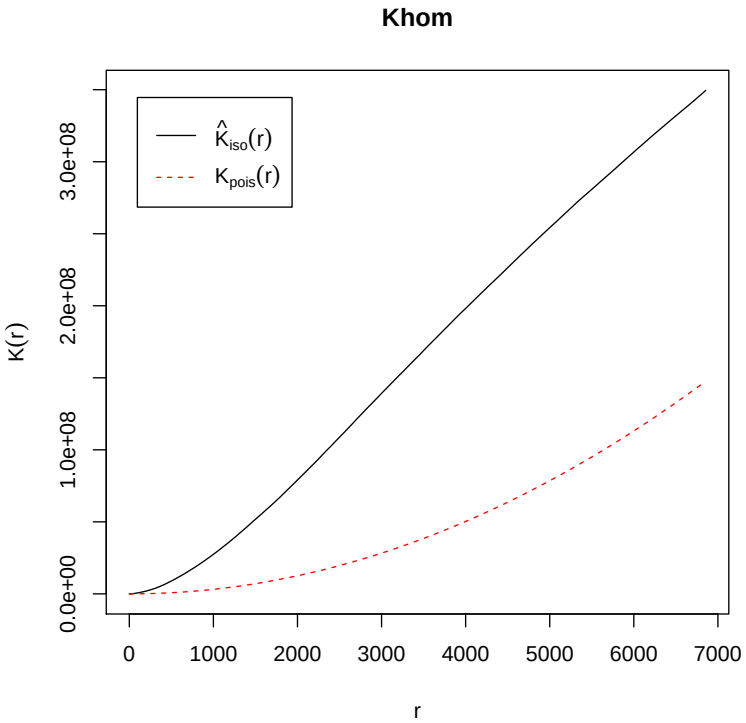
Estimateur de K

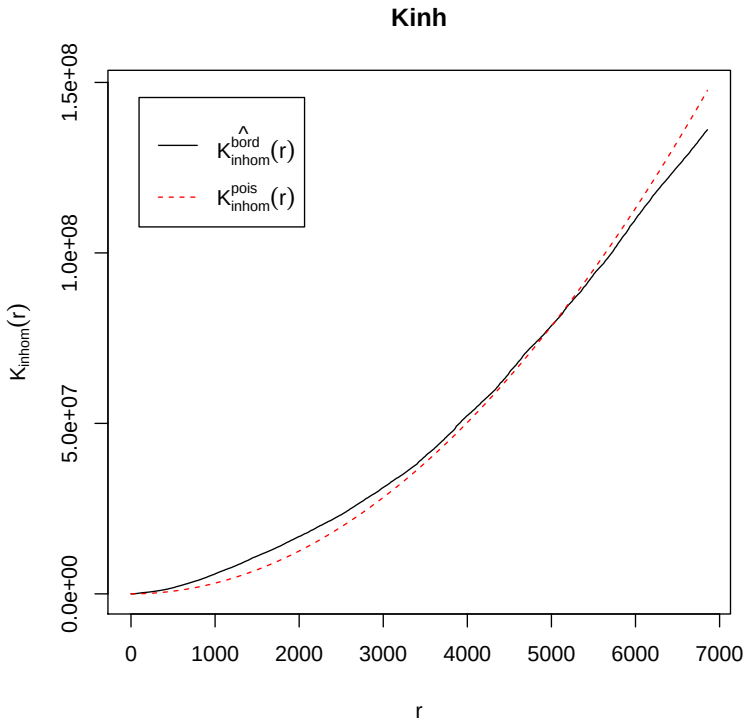
La figure suivante présente des estimateurs des fonctions de Ripley pour les données cells et redwood et l'on voit bien à nouveau la différence de comportement entre processus régulier et aggrégé.





Estimateur de K pour les données Pompiers





Analyse statistique des données spatiales IV

Christine Thomas-Agnan

Toulouse School of Economics

1er novembre 2012

Le package GeoXp

Le module ou "package" GeoXp du logiciel R, disponible sur le site CRAN, a été développé à l'Université des Sciences Sociales de Toulouse pour constituer un outil d'analyse exploratoire spatiale, complémentaire de divers autres packages de R, plus orientés vers la modélisation de données spatiales.
Depuis la version 1.5.0, les fonctions de GeoXp travaillent sur des objets de type `SpatialXXXDataFrame`, c'est à dire comportant en sus des variables d'intérêt, une information géographiques sur les unités spatiales.

Documentation sur GeoXp

Une fois le package installé et chargé dans la session de travail, l'utilisateur pourra consulter la notice en anglais disponible avec le package en exécutant la commande suivante :

```
vignette("presentation_geoxp")
```

Principe d'interactivité de GeoXp

GeoXp lie de façon dynamique des graphiques statistiques avec une carte
Nature des graphiques statistiques

- classiques : histogrammes, boîtes à moustaches, diagramme de dispersion, courbe Lorentz, etc.
- spécifiquement spatiaux : nuage de variogramme, diagramme de Moran

Principe d'interactivité de GeoXp

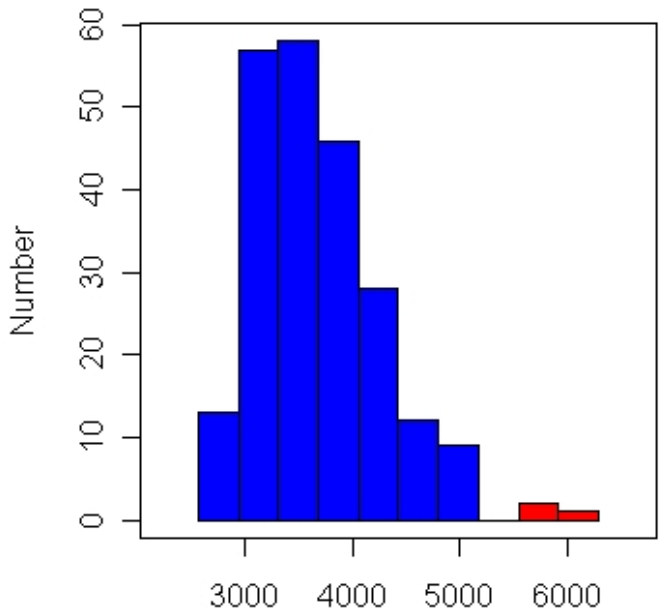
Lien dynamique bilatéral

- La sélection d'un point ou d'une zone sur la carte résulte en la mise en évidence des éléments correspondants du graphique statistique (changement couleur et/ou symbole)
- La sélection d'un élément du graphique statistique résulte en la mise évidence des points ou zones correspondantes sur la carte (changement couleur et/ou symbole)

La sélection peut se faire par point ou par polygone.
La mise en évidence de points ou zones sur un graphique se fait par un changement de couleur et/ou symbole.

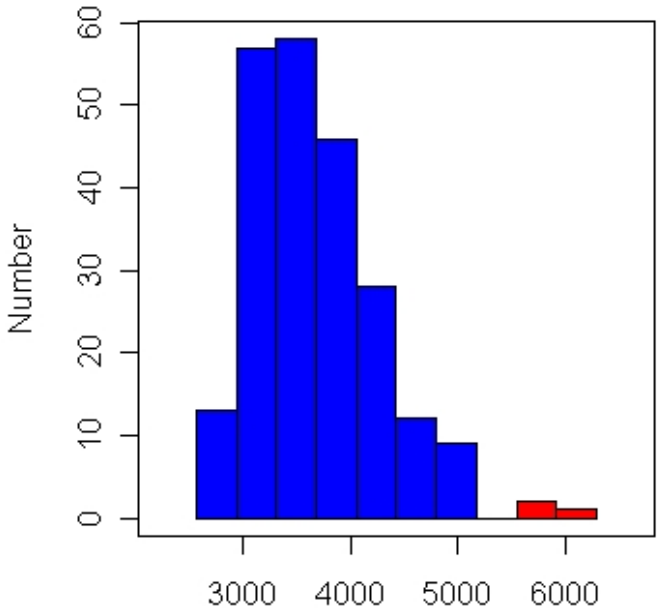
Exemple : histogramme

Coût par élève : sites sélectionnés par clic de souris sur les barres de l'histogramme et représentés en rouge sur la carte.

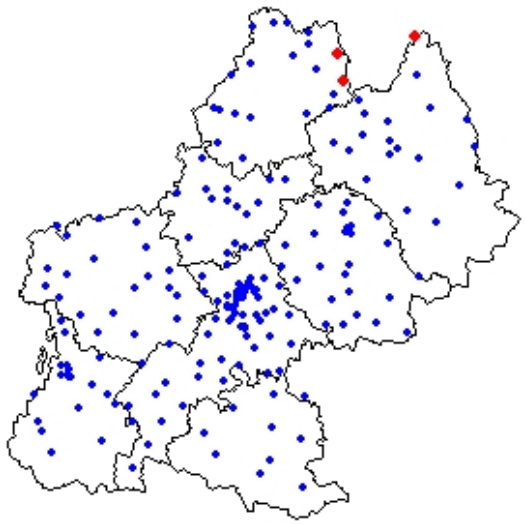


Exemple : histogramme

Coût par élève : sites sélectionnés par clic de souris sur les barres de l'histogramme et représentés en rouge sur la carte.



$\Rightarrow$



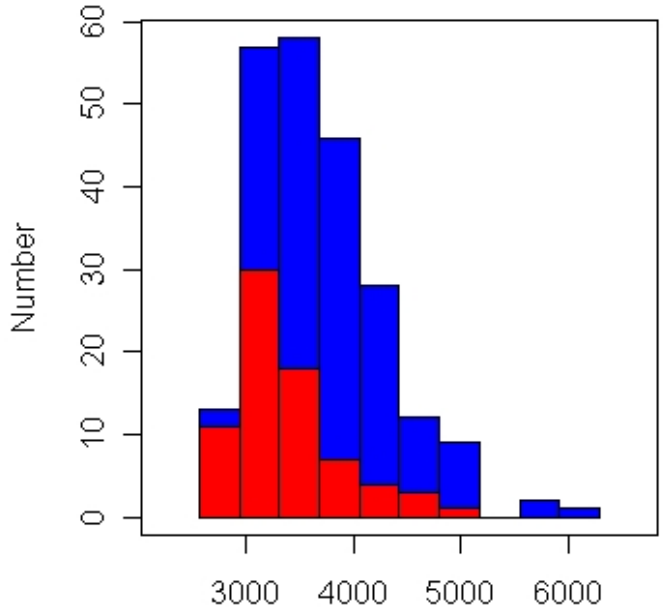
Exemple : histogramme

Coût par élève : sites sélectionnés point par point ou par polygone sur la carte et représentés en rouge sur l'histogramme.

Exemple : histogramme

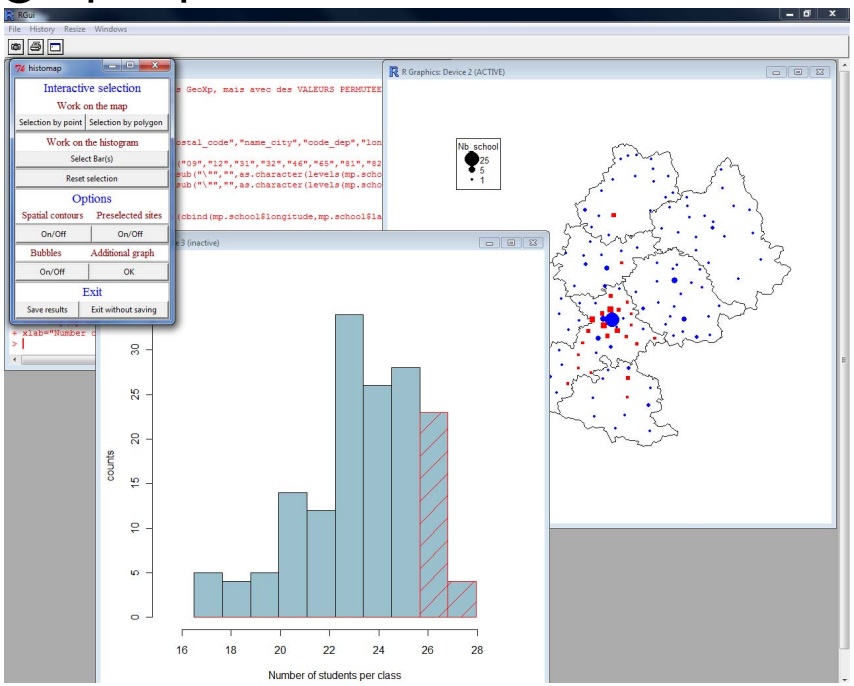
Coût par élève : sites sélectionnés point par point ou par polygone sur la carte et représentés en rouge sur l'histogramme.

$\Rightarrow$



Trois fenêtres

L'appel à une des fonctions GeoXp fait apparaître trois fenêtres : une fenêtre pour l'affichage du graphique statistique, une fenêtre pour l'affichage de la carte et une fenêtre "menu". L'utilisateur doit d'abord choisir dans le menu le graphique sur lequel il désire sélectionner (graphique statistique ou carte), ce qui a pour résultat de rendre ce graphique actif.



Exemple de syntaxe

Le premier argument renseigne un objet de type `SpatialXXXDataFrame` et le second argument le (ou les) nom de la variable à étudier.
Exemple d'appel :

```
histomap(immob.spdf,"prix.vente")
```

Les cartes de GeoXp

Les cartes produites par GeoXp sont rudimentaires car l'objectif n'est pas la cartographie mais l'analyse interactive entre carte et graphique statistique.

Néanmoins on peut améliorer l'aspect des cartes si l'objet est de type `SpatialPolygonsDataFrame` ou si l'on dispose d'un objet de type `SpatialPointsDataFrame` et d'un fond de carte. L'option d'affichage du fond de carte est accessible depuis le menu.

On peut donner un étiquetage ou **label** à chaque observation, par exemple le nom ou le code de la zone. Pour cela, il faut que `row.names(objet)` soit non nul. En utilisant l'option `identify=TRUE` des fonctions de GeoXp, les étiquettes des points sélectionnés apparaissent alors sur la carte à l'issue d'une sélection.

Options

Que ce soit pour les cartes ou les graphiques statistiques, il y a diverses options qui permettent de modifier leur apparence, par exemple

- la sélection sur une barre d'histogramme peut être représentée par une coloration différente (option `col=`),
- la sélection d'un point sur la carte peut être représentée par un symbole différent (option `pch=`).

La sélection

L'utilisateur choisit le type de sélection qu'il désire faire sur le graphique actif (par points, par polygone, barre, etc) et exécute ensuite cette sélection. Un click droit de la souris fait apparaître le bouton "stop" qui permet de terminer une sélection. L'utilisateur peut sélectionner des éléments de la carte avec la souris de deux façons différentes :

- soit un nombre fini de points non connu à l'avance,
- soit l'ensemble des points contenus dans un ou plusieurs polygones : l'utilisateur saisit alors les sommets du ou des polygones avec la souris et termine à nouveau par un click droit.

La sélection

Pour les sélections sur le graphique statistique, plusieurs cas :

- Dans le cas d'un histogramme pour une variable quantitative ou d'un diagramme en barre pour une variable qualitative, la selection permet de choisir une ou plusieurs barres de l'histogramme, non nécéssairement contigües.
- Dans le cas de la courbe de densité, la sélection porte sur un ou des intervalles sur l'axe des abscisses.
- Dans le cas des boîtes à moustaches, la sélection peut porter soit sur les points atypiques, soit sur un ou des quartiles.
- Dans le cas des diagrammes de dispersion, la sélection porte simplement sur un sous ensemble de points et peut se faire comme pour la carte soit sur un nombre fini de points non connu à l'avance soit sur l'ensemble des points contenus dans un polygone.

Quitter le menu et sauvegarder la sélection

La dernière ligne du menu contient les cases “Save results” et “Exit without saving ” qui permettent de quitter la fonction.

NB : tant que l'utilisateur n'aura pas cliqué sur une de ces cases, il ne pourra pas ouvrir une autre fonction de GeoXp.

Si l'utilisateur a choisi la case “Save results”, cela a pour effet de créer un objet de type numeric, appelé `last.select` et qui contient les indices des dernières unités spatiales sélectionnées qui peut être réutilisé dans des analyses ultérieures, par exemple pour caractériser la zone sélectionnée.

Corriger une sélection

L'utilisateur peut corriger toute sélection en cours sans avoir à redémarrer le processus à zéro, c'est à dire qu'il peut déselectionner un point sélectionné par erreur.

De même, il peut modifier une sélection par ajout ou soustraction après en avoir constaté les effets sur l'autre graphique sans avoir à sortir et rappeler la fonction.

Le graphique non actif est actualisé à mesure de la sélection.

Graphique supplémentaire

L'utilisateur a la possibilité de faire un graphique supplémentaire choisi parmi : histogramme, diagramme en barre, nuage de points

Mais avec une **interactivité unilatérale** : les sélections faites sur le premier graphique ou sur la carte se répercutent sur le graphique supplémentaire mais on ne peut pas sélectionner sur ce dernier.

Les variables proposées sont toutes celles contenues dans `objet@data`.

Label

On peut mettre un label ou étiquette sur les sites sélectionnés (nom ou code de la zone ou autre caractéristique). Pour cela on utilise l'option `identify=TRUE`.

```
histomap(immob.spdf,"prix.vente",identify=TRUE)
```

Cercles concentriques

On peut représenter les sites avec une taille proportionnelle à une variable choisie parmi les variables de type `numeric` incluses dans `objet@data`. Il suffit ensuite d'utiliser le bouton "bubbles".

Le choix de cette case a pour effet d'ouvrir une fenêtre tk qui vous demande si vous souhaitez afficher une légende sur la carte pour donner la correspondance entre taille des cercles et valeurs prises par la variable sélectionnée.

Sélection non interactive

On peut afficher une sélection supplémentaire non active qui sert de réperage avec le bouton "Preselected sites". Il faut utiliser l'option `criteria` qui contient un vecteur de booléen de la même taille que l'objet spatial. Par exemple, pour préselectionner les villes avec un prix de location moyen au m^2 supérieur à 12 euros :

```
histomap(immob.spdf,"prix.vente",
criteria=(immob.spdf@data$prix.location>12))
```

Couleurs et symboles

Pour les graphiques manipulant une variable qualitative, l'option `col=` permet de donner des couleurs différentes (sur le diagramme en tuyaux d'orgue et sur la carte) en fonction des modalités du facteur.
L'option `pch` a pour effet de d'afficher les unités spatiales sur la carte avec des symboles différents.
Par exemple :

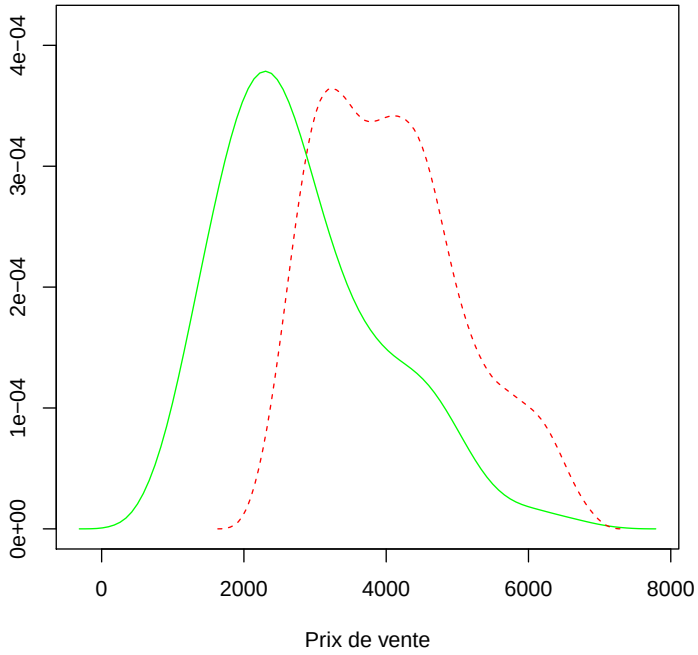
```
barmap(columbus,"CP",col=c("orange","violet"),pch=c(2,4))
```

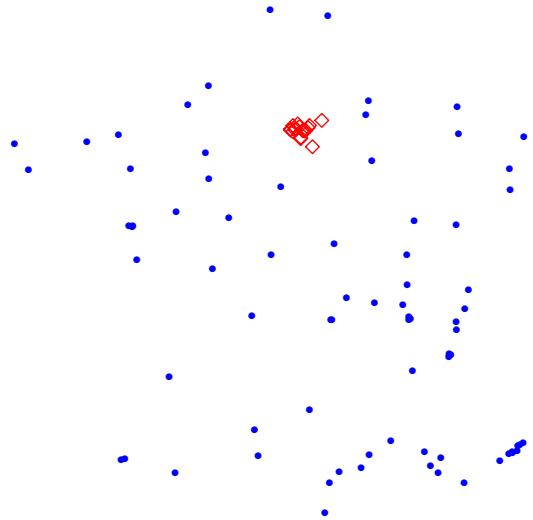
Le code ci-dessus a pour effet d'ouvrir une fenêtre tk qui vous demande si vous souhaitez afficher une légende sur la carte pour donner la définition des couleurs et symboles représentés.

Analyse d'une répartition

Pour décrire la répartition d'une variable quantitative de façon plus fine qu'avec une boîte à moustache, on peut utiliser un histogramme ou un estimateur à noyau de la densité.
L'avantage d'un estimateur continu de la densité sur l'histogramme est évident lorsque l'on veut comparer deux répartitions : par exemple celle d'une variable sur l'ensemble de la région avec celle de la même variable sur une sous-région. On peut superposer deux histogrammes en fréquence mais la superposition de deux courbes de densité reste plus lisible.

Analyse d'une répartition : exemple avec densitymap



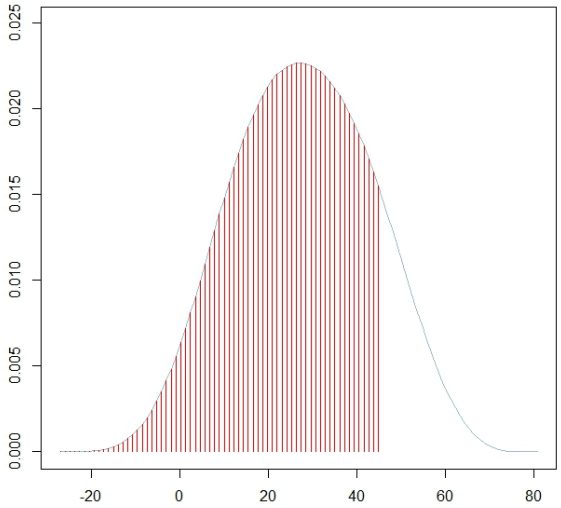


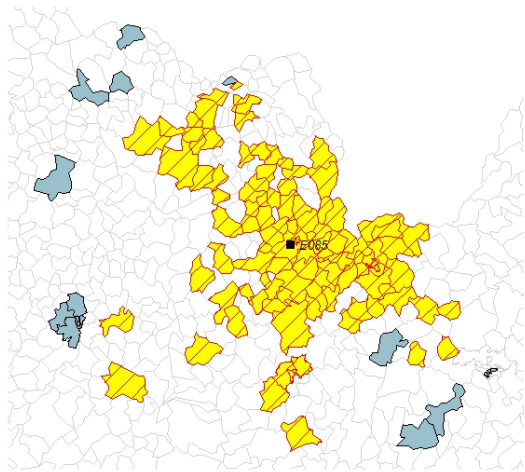
Analyse d'une répartition

Que ce soit avec l'histogramme ou avec la densité, il est intéressant pour une variable donnée d'explorer en particulier les queues de distribution, à droite et à gauche, pour déterminer si elles occupent une position particulière sur la carte. La sélection des queues de la distribution met potentiellement en lumière sur la carte des zones ayant un comportement atypique. Inversement, si l'on s'intéresse à une sous-région donnée, sa sélection sur la carte permet de comparer la sous-distribution de la variable dans cette zone avec la distribution globale. Le paramètre de lissage de l'estimateur à noyau de la densité est ajustable à l'œil avec une réglette. Pour une variable qualitative, le diagramme en barre remplace l'histogramme, mais l'utilisation et les objectifs sont similaires.

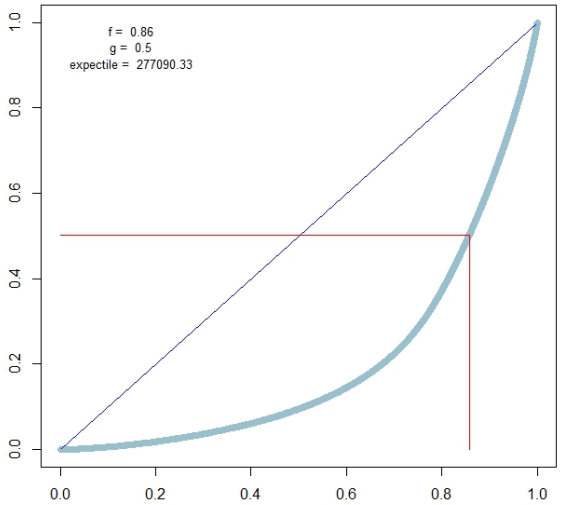
Application : Zone de chalandise basée sur les distances-temps

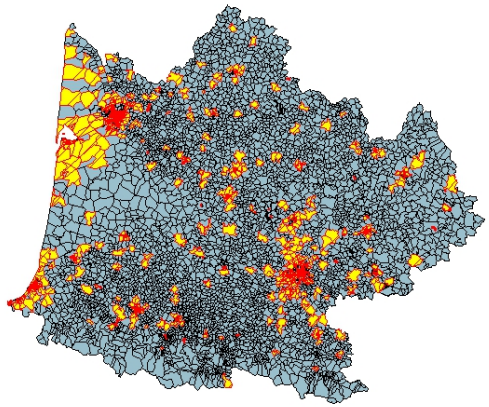
Zone de chalandise du magasin E085 basée sur les distances-temps.





Analyse de concentration : concentration du potentiel cumulé en géomarketing

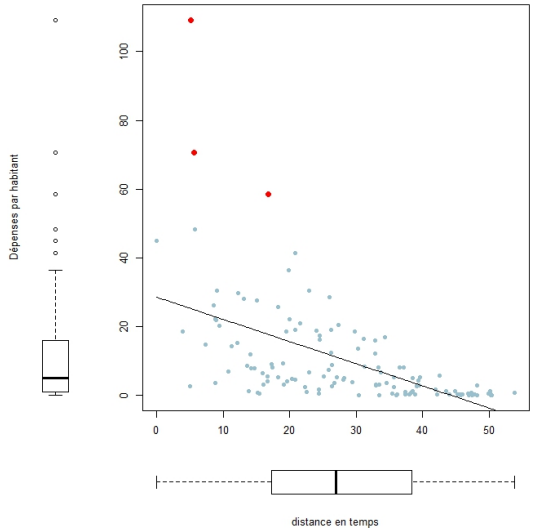


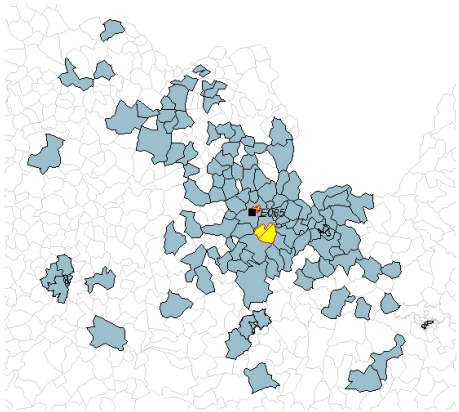


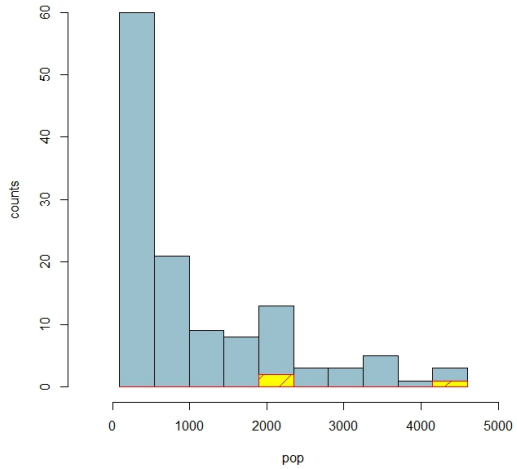
Potentiel : P_{ij} CA du magasin j provenant de l'iris i. Potentiel total de l'iris i : somme des potentiels sur tous les magasins (enseigne et concurrence).
Produits blancs : 86 % des iris aux plus faibles potentiels concentrent 50 % du potentiel total (correspondant à des dépenses de moins de 277090.33 euros sur la période d'interêt) $\longmapsto$ agglomérations toulousaines et bordelaises.

Analyse bivariée

Pour la magasin E085, sélection d'iris ayant un fort potentiel par habitant comparé à des iris à même distance de E085.

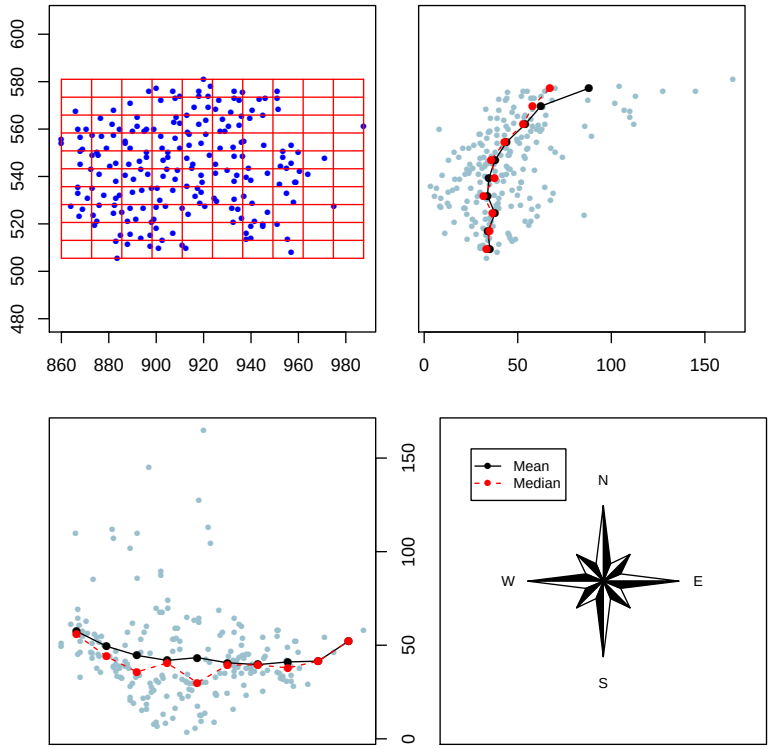






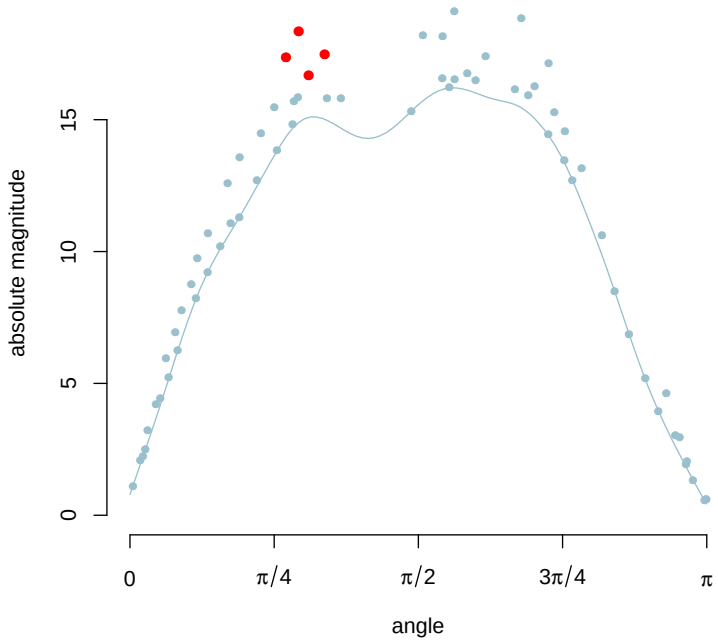
Analyse exploratoire d'une tendance directionnelle

Avec la fonction driftmap de GeoXp, tendance de la variable HOVAL des données columbus



Analyse exploratoire d'une tendance directionnelle

Avec la fonction angleplotmap de GeoXp, tendance de la variable latitude des données columbus



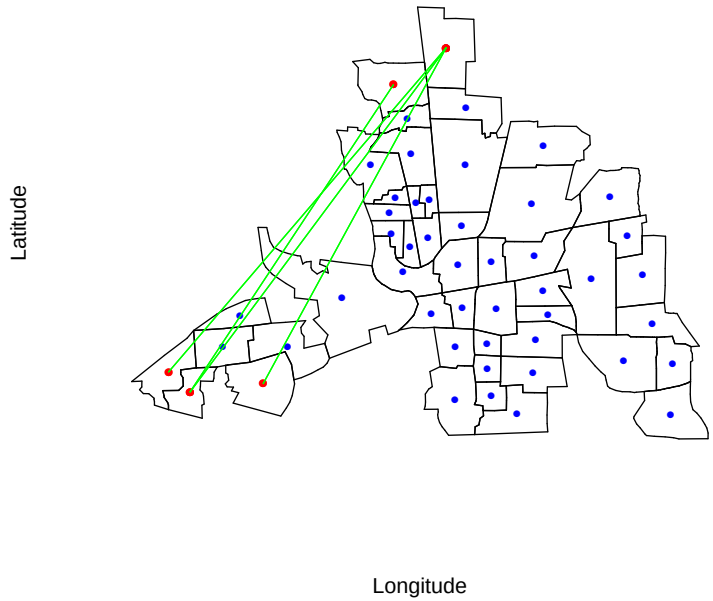


Diagramme de Moran

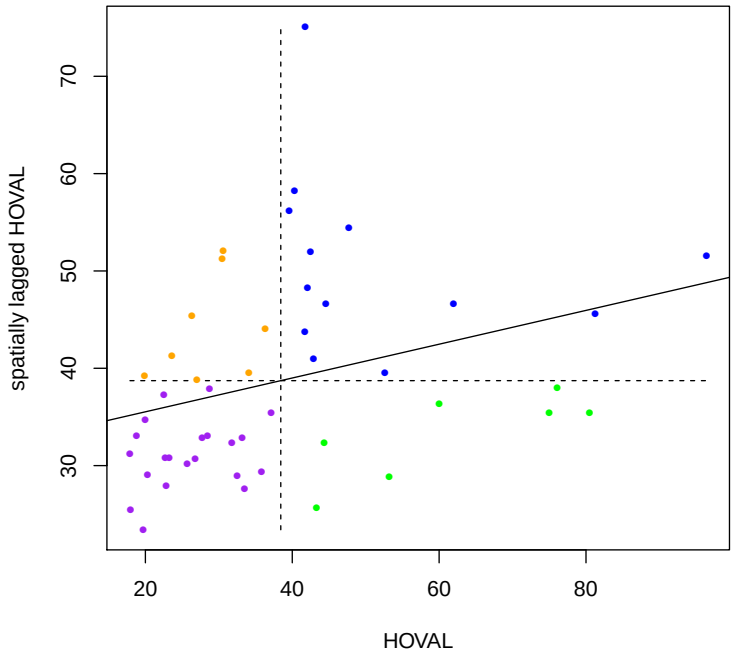
Le **"Diagramme de Moran"** est un nuage de points de WX contre X, où X est centrée et W normalisée.
On peut superposer au nuage la droite de régression qui passe par le point moyen. La pente de celle-ci est égale à l'indice de Moran.
Utilisation :

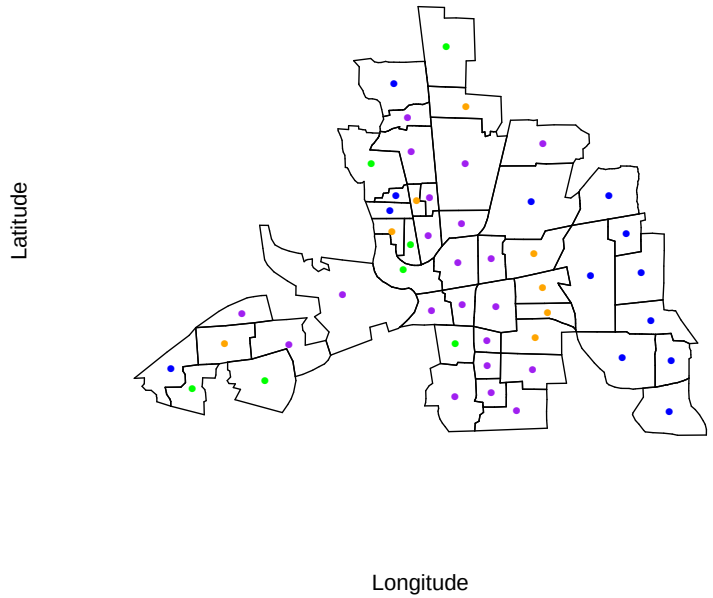
- détecter des points aberrants
- aprécier le degré d'autocorrélation
- non linéarité $\mapsto$ plusieurs régimes d'association spatiale.

Remarque : il est intéressant de normaliser X avant de faire le graphique pour pouvoir ainsi comparer plusieurs moran plots entre eux.

Diagramme de Moran

Diagramme de Moran de la variable HOVAL (données columbus) avec coloration des quatre quadrants





Graphique des voisinages

Pour une matrice de voisinage W et une variable X données, le **graphique des voisinages** consiste en un simple diagramme de dispersion où l'on porte pour tout site i, en abscisse la valeur X_i de la variable X au site i et en ordonnée les valeurs X_j de la variable X aux sites j voisins de i au sens de W, c'est-à-dire tels que $w_{ij} \neq= 0$.
Dans GeoXp, ce diagramme est lié à la carte grâce à la fonction **neighbourmap** et la sélection d'un point sur ce graphique provoque l'affichage du site correspondant sur la carte ainsi que de ses voisins au sens de W, reliées à i par un segment.

Analyse exploratoire d'une matrice de voisinage

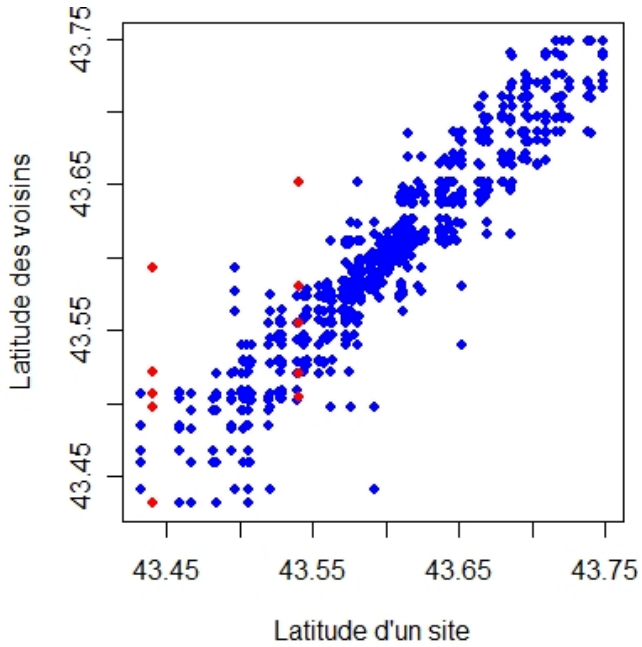
Si l'on utilise la fonction neighbourmap de GeoXp avec les variables géographiques (latitude ou longitude), ce diagramme permet d'explorer la matrice dans le sens suivant

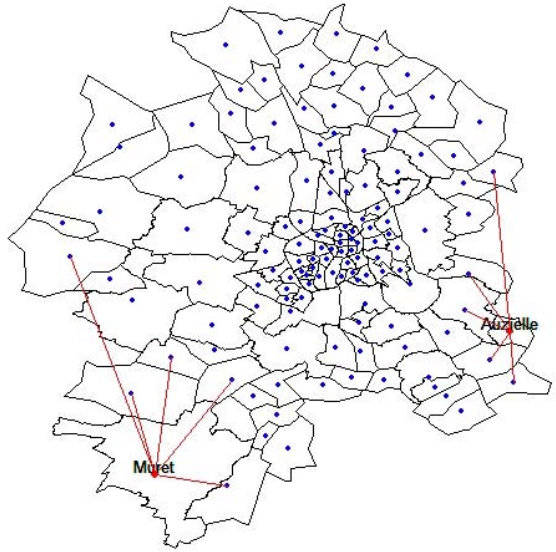
- visualiser qui est voisin de qui
- aprécier visuellement la taille des voisinages (si matrice de type knn)
- aprécier visuellement le nombre de voisins (si matrice de type distance)

```
library(GeoXp)
neighbourmap(nc, "east", wd.nb)
```

Diagramme des voisins pour la matrice de Delaunay

Variable : Latitude, voisins avec de grandes differences en latitude



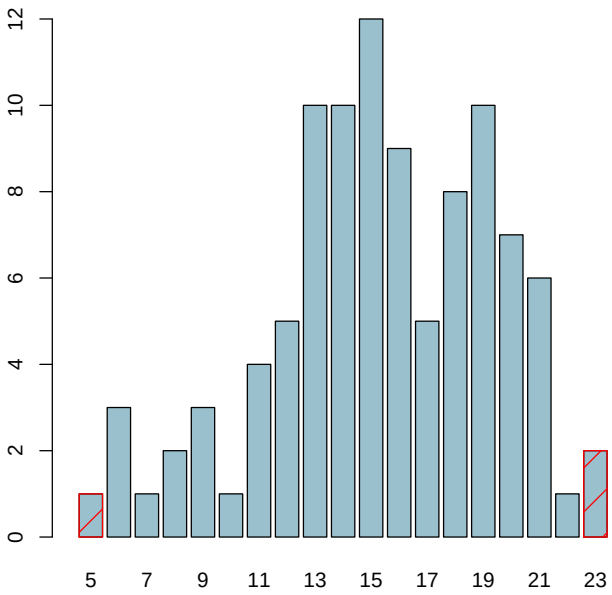


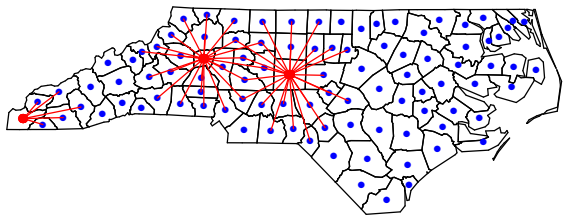
Analyse exploratoire d'une matrice de voisinage

La fonction 'barnbmap' de GeoXp réalise un diagramme en tuyaux d'orgue du nombre de voisins des sites, lié à une carte.

```
barnbmap(nc,wd.nb)
```

Pour données SIDS avec la matrice basée sur un seuil de distance



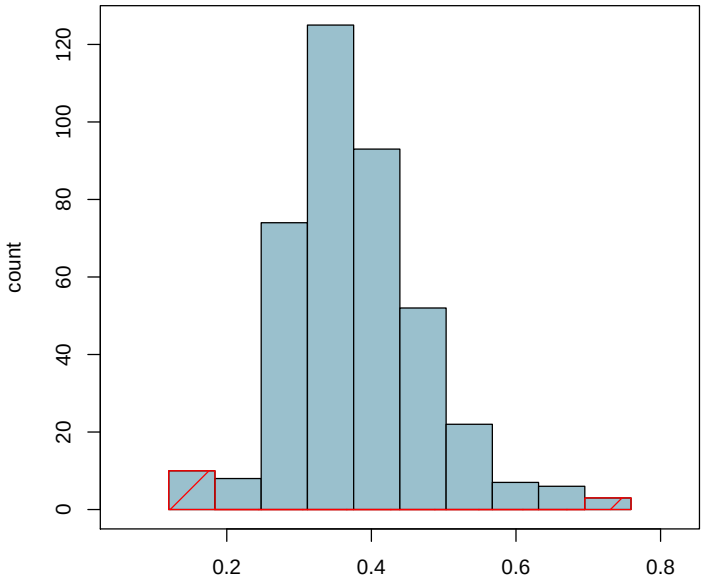


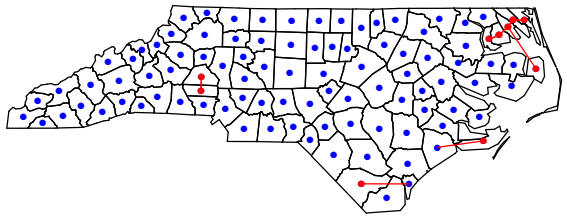
Analyse exploratoire d'une matrice de voisinage

De même, la fonction 'histnbmap' réalise un histogramme des distances aux voisins lié à une carte.

```
histnbmap(nc,knn2nb(wv.knn))
```

Pour données SIDS avec la matrice basée sur les quatre plus proches voisins

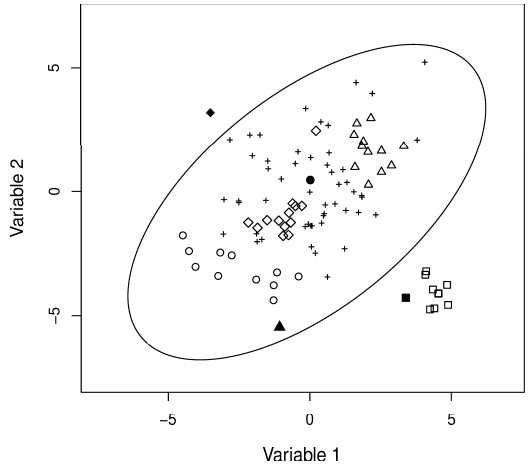


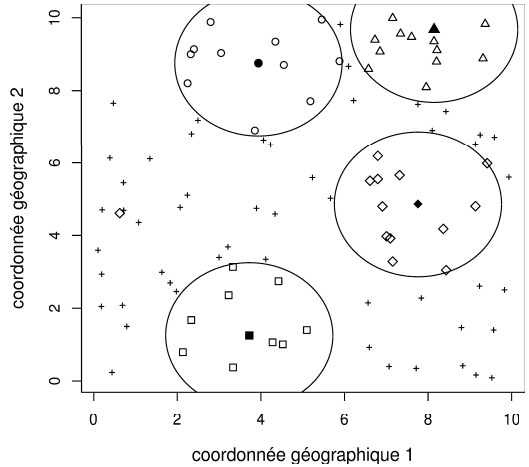


Valeurs atypiques

Dès que les données sont géoréférencées, il existe deux sortes de valeurs atypiques : les globales et les locales.
Un point est dit **aberrant global** pour la variable X si sa valeur pour X est extrême par rapport à l'ensemble de la distribution de X.
Un point est dit **aberrant local** pour la variable X si sa valeur pour X est extrême par rapport à l'ensemble de la sous-distribution des X sur ses voisins (pour une structure de voisinage donnée).
Un aberrant global est nécessairement un aberrant local, mais un aberrant local peut très bien ne pas être un aberrant global.
Pour détecter les aberrants locaux, on peut utiliser la fonction neighbourmap sur la variable d'intérêt. Ils apparaissent alors comme les points éloignés de la diagonale.

Valeurs atypiques illustration





Analyse statistique des données spatiales V

Christine Thomas-Agnan

Toulouse School of Economics

2 novembre 2012

Test de Moran pour variable surfacique continue

Il s'agit de tester l'hypothèse d'absence d'autocorrélation spatiale pour une variable brute X.

H_0 : absence d'autocorrélation spatiale

H_1 : présence d'autocorrélation spatiale

Il faut préciser $H_0 \hookrightarrow$ deux modèles différents

Test de Moran pour variable surfacique continue : test gaussien

- le modèle "free sampling" : $X_1, \cdots, X_n$ sont i.i.d. $\mathcal{N}(0, \sigma^2)$
  Ce test, dit "test gaussien", teste si l'échantillon observé est représentatif de la distribution d'un vecteur gaussien de composantes i.i.d.
  En pratique, on utilise la loi asymptotique de I sous H_0. Pour cela, on a besoin de normaliser d'abord l'indice en lui enlevant sa moyenne et en le divisant par son écart-type. Ensuite, on utilise la loi asymptotique $\mathcal{N}(0, 1)$ de l'indice normalisé pour calculer une p-valeur associée.

Les moments du I de Moran sous l'hypothèse nulle

Le calcul des moments du I de Moran utilise le **Théorème de Pitman et Koopmans**

Si $X_1, \cdots X_n$ sont i.i.d. $\mathcal{N}(0,1)$ et si
$H = h(X_1, \cdots, X_n)$ est une statistique indépendante de l'unité ($h(\lambda X_1, \cdots, \lambda X_n) = h(X_1, \cdots, X_n)$ quel que soit $\lambda > 0$), alors H est indépendante de $Q = \sum_{i=1}^n X_i^2$.

Dans le modèle free-sampling, on obtient

$$\mathbb{E}(I) = -\frac{1}{n-1}, \mathbb{E}(I^2) = \frac{n^2 S_1 - n S_2 + 3 S_0^2}{(n^2-1) S_0^2}$$

Test de Moran pour variable surfacique continue : test gaussien

```
> moran.test(columbus$HOVAL, nb2listw(col.gal.nb),
randomisation=FALSE)

        Moran's I test under normality

data:  columbus$HOVAL
weights: nb2listw(col.gal.nb)

Moran I statistic standard deviate = 2.066, p-value = 0.01941
alternative hypothesis: greater
sample estimates:
Moran I statistic       Expectation          Variance
      0.173645208      -0.020833333       0.008860962
```

Test de Moran pour variable surfacique continue : test gaussien

Expliquez cette expérience :

```
>S=sample(columbus$HOVAL, length(columbus$HOVAL))
>moran.test(S, nb2listw(col.gal.nb), randomisation=FALSE)


        Moran's I test under normality

data:  S
weights: nb2listw(col.gal.nb)

Moran I statistic standard deviate = -0.2821, p-value = 0.611
alternative hypothesis: greater
sample estimates:
Moran I statistic       Expectation          Variance
     -0.047385261      -0.020833333       0.008860962
```

Test de Moran pour variable continue : test de permutation

- le modèle "non free sampling" ou modèle de randomisation : conditionnellement à $X_i = x_i$, en l'absence d'autocorrélation spatiale les $n!$ permutations des réalisations $x_1, \cdots, x_n$ sont équiprobables. Ce test, dit "test de permutation", teste si l'échantillon observé est représentatif d'une allocation aléatoire uniforme des valeurs $x_1, \cdots, x_n$ aux n sites de la carte. Dans ce cas, notons que les lois marginales conditionnelles ne sont pas indépendantes.

  On a aussi $\mathbb{E}(I) = -\frac{1}{n-1}$ mais la formule de la variance est plus compliquée.

Test de Moran pour variable continue : test de permutation

En pratique, on tire au hasard T permutations, on calcule les indices de Moran pour chacune de T permutations, leur minimum I_{min} et maximum I_{max}. On compare alors la valeur observée de l'indice de Moran avec l'intervalle $[I_{min}, I_{max}]$.
On rejette H_0 si l'indice de Moran n'est pas dans cet intervalle.
Le "pseudo-niveau de signification" empirique du test est égal à $(L+1)/(T+1)$ où L est le nombre de fois parmi les T permutations que l'indice de Moran recalculé dépasse la valeur observée sur l'échantillon. (le $+1$ vient du fait qu'on compte l'observation et les T permutations).

Test de Moran pour variable continue : test de permutation

```
> moran.test(columbus$HOVAL, nb2listw(col.gal.nb),
randomisation=TRUE)

        Moran's I test under randomisation

data:  columbus$HOVAL
weights: nb2listw(col.gal.nb)

Moran I statistic standard deviate = 2.1001, p-value = 0.01786
alternative hypothesis: greater
sample estimates:
Moran I statistic       Expectation          Variance
      0.173645208      -0.020833333       0.008575953
```

Test d'autocorrélation basé sur l'indice de Geary

Du lien entre Moran et Geary, on déduit les formules des moments de l'indice de Geary

- free sampling

$$\mathbb{E}(G) = 1, \mathbb{V}ar(G) = \frac{(2S_1 + S_2)(n-1) - 4S_0^2}{2(n+1)S_02}$$

- non free sampling
  $\mathbb{E}(G) = 1, n(n-2)(n-3)S_0^2\mathbb{V}ar(G) = (n-1)S_1[n^2 - 3n + 3 - (n-1)b_2] - \frac{1}{4}(n-1)S_2[n^2+3n-6-(n^2-n+2)b_2] + S_0^2(n^2-3-(n-1)^2 b_2)$

Test d'autocorrélation basé sur l'indice de Geary

```
>geary.test(columbus$HOVAL, nb2listw(col.gal.nb),
randomisation=FALSE)
         Geary's C test under normality

data:  columbus$HOVAL
weights: nb2listw(col.gal.nb)

Geary C statistic standard deviate = 1.7972, p-value = 0.03615
alternative hypothesis: Expectation greater than statistic
sample estimates:
Geary C statistic       Expectation          Variance
       0.81754447        1.00000000       0.01030674

> geary.test(columbus$HOVAL, nb2listw(col.gal.nb),
randomisation=TRUE)
       Geary's C test under randomisation

data:  columbus$HOVAL
weights: nb2listw(col.gal.nb)

Geary C statistic standard deviate = 1.7083, p-value = 0.04379
alternative hypothesis: Expectation greater than statistic
sample estimates:
Geary C statistic       Expectation          Variance
       0.81754447        1.00000000       0.01140734
```

Test d'autocorrélation pour variable qualitative : test gaussien

Si X est qualitative avec k modalités :
- le modèle "free" : tirage aléatoire avec remise dans une population ayant k groupes de proportions $p_1, \cdots, p_k$ connues : les X_i sont indépendantes de loi multinomiale.
- le modèle "non free" : tirage aléatoire sans remise dans une population ayant k groupes d'effectifs connus $n_1, \cdots, n_k$: la loi du n-uplet $(X_1, \cdots, X_n)$ est la loi hypergéométrique conditionnelle aux effectifs de groupe observés.

En pratique, $p_1, \cdots, p_k$ doivent être estimées par les fréquences empiriques. Dans le cas "non free", notons que les lois marginales ne sont pas indépendantes.

Test "join counts" pour variable dichotomique

Le modèle "free sampling" suppose les X_i iid Bernouilli $\mathcal{B}(1,p)$

$$\mathbb{E}(BB) = \frac{1}{2} S_0 p^2$$

$$4\mathbb{V}ar(BB) = p^2(1-p)[S_1(1-p) + S_2 p]$$

$$\mathbb{E}(BW) = S_0 p(1-p)$$

$$4\mathbb{V}ar(BW) = [4S_1 p(1-p) + S_2 p(1-p)(1-4p(1-p))]$$

Notations :

$$S_0 = \sum_{i \neq j} w_{ij}, S_1 = \frac{1}{2} \sum_{i \neq j} (w_{ij} + w_{ji})^2, S_2 = \sum_{i \neq j} (w_{i+} + w_{+i})^2$$

$$w_{i+} = \sum_j w_{ij}, w_{+j} = \sum_j w_{ji}$$

Test "join counts" pour variable dichotomique

Le modèle "non free sampling" suppose qu'il y a $n_B = \sum_i X_i$ valeurs 1 et $n - n_B$ valeurs 0, et que l'on fait un tirage sans remise
Si $n^{(b)} = n(n-1)\cdots(n-b+1)$, on a

$$\mathbb{E}(BB) = \frac{S_0}{2}\frac{n_B^{(2)}}{n^{(2)}}$$

$$\begin{aligned} 4\mathbb{V}ar(BB) &= [S_1(\frac{n_B^{(2)}}{n^{(2)}} - 2\frac{n_B^{(3)}}{n^{(3)}} + \frac{n_B^{(4)}}{n^{(4)}}) \\ &+ S_2(\frac{n_B^{(3)}}{n^{(3)}} - \frac{n_B^{(4)}}{n^{(4)}}) + \frac{S_0^2 n_B^{(4)}}{n^{(4)}} - (\frac{S_0 n_B^{(2)}}{n^{(2)}})^2] \end{aligned}$$

$$4as\mathbb{V}ar(BB) = p^2(1-p)[S_1(1-p) + S_2 p - 4\frac{S_0^2 p}{n}]$$

Test "join counts" pour variable dichotomique

```
> joincount.test(as.factor(HICRIME),nb2listw(col.gal.nb))

        Join count test under nonfree sampling

data:  as.factor(HICRIME)
weights: nb2listw(col.gal.nb)

Std. deviate for FALSE = 4.6176, p-value = 1.941e-06
alternative hypothesis: greater
sample estimates:
Same colour statistic           Expectation              Variance
            9.4833333             6.2500000             0.4903158


        Join count test under nonfree sampling

data:  as.factor(HICRIME)
weights: nb2listw(col.gal.nb)

Std. deviate for TRUE = 4.9963, p-value = 2.921e-07
alternative hypothesis: greater
sample estimates:
Same colour statistic           Expectation              Variance
             9.206349              5.750000             0.478553
```

Test "join counts" pour variable dichotomique

```
joincount.mc(HICRIME,nb2listw(col.gal.nb),nsim=100)

        Monte-Carlo simulation of join-count statistic

data:  HICRIME
weights: nb2listw(col.gal.nb)
number of simulations + 1: 101

Join-count statistic for faible = 9.4833, rank of observed statistic = 101, p-value = 0.009901
alternative hypothesis: greater
sample estimates:
    mean of simulation variance of simulation
             6.3177480              0.5678434


        Monte-Carlo simulation of join-count statistic

data:  HICRIME
weights: nb2listw(col.gal.nb)
number of simulations + 1: 101

Join-count statistic for fort = 9.2063, rank of observed statistic = 101, p-value = 0.009901
alternative hypothesis: greater
sample estimates:
    mean of simulation variance of simulation
             5.7360655              0.5681022
```

Pratique des tests d'autocorrélation spatiale

Choix entre "free sampling" et "non free sampling" :
- guidé par le contexte
- si X suit une loi F inconnue de variance finie, on a toujours la même espérance et le moment d'ordre deux vérifie $\mathbb{E}(I^2) = \mathbb{E}(\mathbb{E}_R(I^2))$.

Choix entre I et G :
- l'indice de Geary est plus sensible aux points aberrants
- l'approximation gaussienne est meilleure pour I que pour G

Test d'autocorrélation des résidus d'un modèle linéaire ordinaire

L'indice de Moran généralisé s'écrit comme l'indice de Moran appliqué aux résidus du modèle WLS : ceux-ci n'étant pas des observations mais des estimations, il faut ajuster les calculs de moments dans le contexte "free sampling".
Dans le cas $D = I_n$, on montre que sous l'hypothèse d'absence d'autocorrélation spatiale avec une matrice de voisinage W

$$\mathbb{E}(I) = -\frac{trA}{n-k},$$

où k est le nombre de colonnes de X et $A = (X'X)^{-1}X'WX$.

Test d'autocorrélation des résidus d'un modèle linéaire ordinaire

```
> lmmod=lm(HOVAL~CRIME+INC,data=columbus)
> lm.morantest(lmmod,nb2listw(col.gal.nb))
        Global Moran's I for regression residuals

data:
model: lm(formula = HOVAL ~ CRIME + INC, data = columbus)
weights: nb2listw(col.gal.nb)

Moran I statistic standard deviate = 2.1947, p-value = 0.01409
alternative hypothesis: greater
sample estimates:
Observed Moran's I       Expectation          Variance
      0.167370309       -0.034246629        0.008439035
```

Hypothèse CSR

On dit qu'un processus ponctuel vérifie l' hypothèse d'homogénéité spatiale (hypothèse CSR pour "complete spatial randomness") si c'est un processus de Poisson homogène.
Cette hypothèse implique donc à la fois l'homogénéité de la répartition des points mais aussi l'indépendance entre les observations dans des zones disjointes.
Tester l'hypothèse CSR est la première étape dans la modélisation d'un processus ponctuel dans le sens où si cela est le cas, le processus sera entièrement caractérisé par le réel λ de la définition.
Si cela n'est pas le cas, c'est alors que le travail de modélisation peut commencer.
Il existe de nombreux tests de CSR mais nous allons seulement développer deux approches.

Test basé sur les quadrats

Diviser la fenêtre d'observation en m quadrats, c'est à dire en cellules rectangulaires ou carrées d'égale surface
Dénombrer les points du processus dans chaque cellule, notés $n_k, k = 1, \ldots, m$.
Avec $\bar{n} = \frac{n}{m}$, on définit

$$I = \sum_{k=1}^{m} \frac{(n_k - \bar{n})^2}{(m-1)\bar{n}}.$$

I peut d'abord être interprété comme le rapport entre la variance empirique des effectifs n_k et leur moyenne (coefficient de variation).

Test basé sur les quadrats

Sous l'hypothèse CSR, les effectifs sont équidistribuées (même surface), de loi de Poisson et comme la moyenne d'une loi de Poisson est égale à sa variance, I n'est autre que le ratio de deux estimateurs de la variance.
Conditionnellement au nombre total de points, $(m-1)I$ n'est autre que le χ^2 de Pearson d'ajustement de la série des effectifs des quadrats.
Sous l'hypothèse CSR, la loi de $(m-1)I$ peut être approximée asymptotiquement par une loi de χ^2 à $m-1$ degrés de liberté.

Test basé sur les quadrats

Interprétation :

- lorsque I est significativement grand et que l'homogénéité est respectée, il denote une tendance à l'**aggrégation**, c'est à dire une dépendance entre les points de type attraction.
- Inversement, lorsque I est significativement petit et que l'homogénéité est respectée, il traduit une tendance à la régularité, c'est à dire une dépendance entre les points de type répulsion.

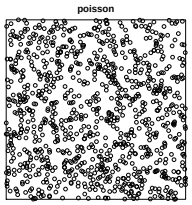
Test basé sur les quadrats

```
> poisson=rpoispp(10,win=window)
> quadrat.test(poisson)

        Chi-squared test of CSR using quadrat counts

data:  poisson
X-squared = 16.9478, df = 24, p-value = 0.8509

Quadrats: 5 by 5 grid of tiles
```



Test non significatif $\Rightarrow$ non rejet de CSR

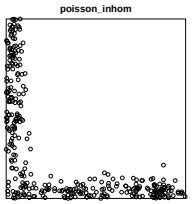
Test basé sur les quadrats

```
> poisson_inhom=rpoispp(function(x,y){100*exp(-3*x)+100*exp(-3
20,win=window)
> quadrat.test(poisson_inhom)

        Chi-squared test of CSR using quadrat counts

data:  poisson_inhom
X-squared = 652.0952, df = 24, p-value < 2.2e-16

Quadrats: 5 by 5 grid of tiles
```

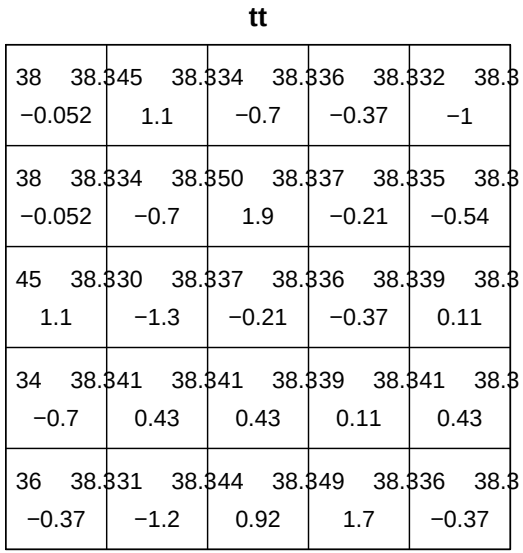


Test significatif $\Rightarrow$ rejet de CSR

Test basé sur les quadrats

```
> tt=quadrat.test(poisson)
> plot(tt)
```

A gauche les effectifs observés, à droite les effectifs estimés sous CSR, au centre les résidus de Pearson.



Diagnostic basé sur des simulations

Une autre approche pour évaluer l'hypothèse CSR consiste à simuler M réalisations d'un processus de Poisson homogène et de calculer des caractéristiques du processus (fonctions F,G, K ou L, voir + loin) pour chaque simulation.
On trace ensuite les enveloppes de ces courbes sur l'ensemble des simulations et on évalue si la caractéristique observée sur l'échantillon entre ou non dans ces enveloppes. Nous reviendrons sur cette méthode après avoir défini ces caractéristiques.

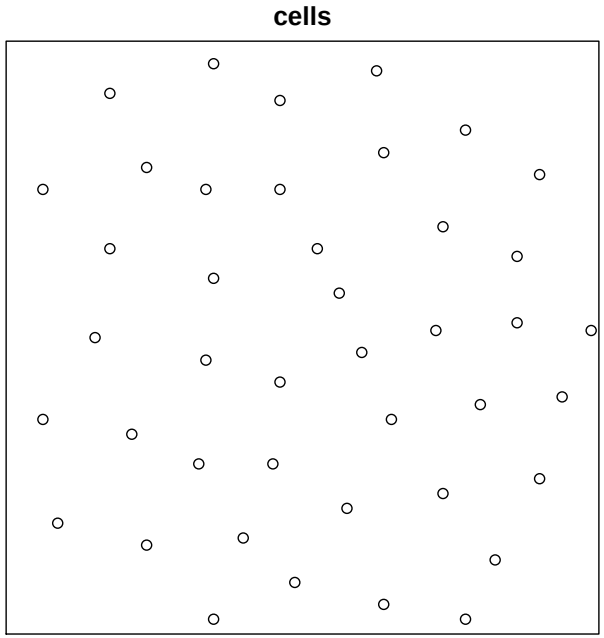
Test de CSR basé sur la fonction F

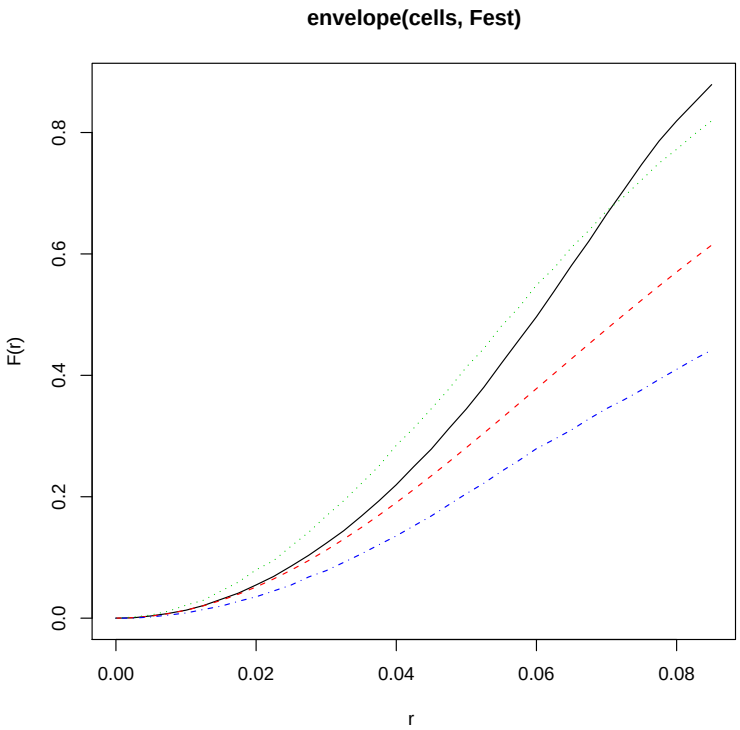
La méthode suivante permet d' évaluer qualitativement l'hypothèse CSR par des simulations. On simule M réalisations d'un processus de Poisson homogène dans E et on calcule la fonction $\hat{F}_k(r)$ pour chaque simulation k.
On détermine ensuite l'enveloppe supérieure F_U et inférieure F_L par

$$F_U(r) = \max_{k=1}^{M} \hat{F}_k(r), F_L(r) = \min_{k=1}^{M} \hat{F}_k(r).$$

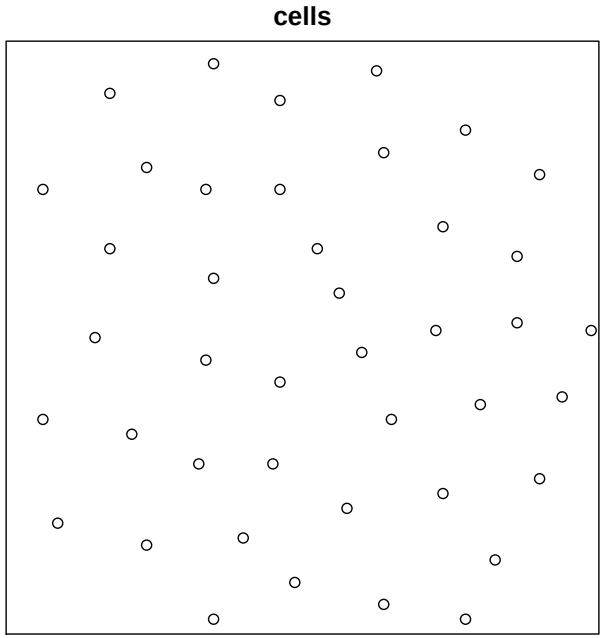
Si la fonction $\hat{F}(r)$ de notre réalisation se trouve dans l'enveloppe, on en déduit que le modèle de Poisson homogène est compatible avec les données. Pour le jeu de données cells (positions de cellules), on voit que la fonction $\hat{F}$ en noir sur la figure suivante sort de l'enveloppe (en pointillés).

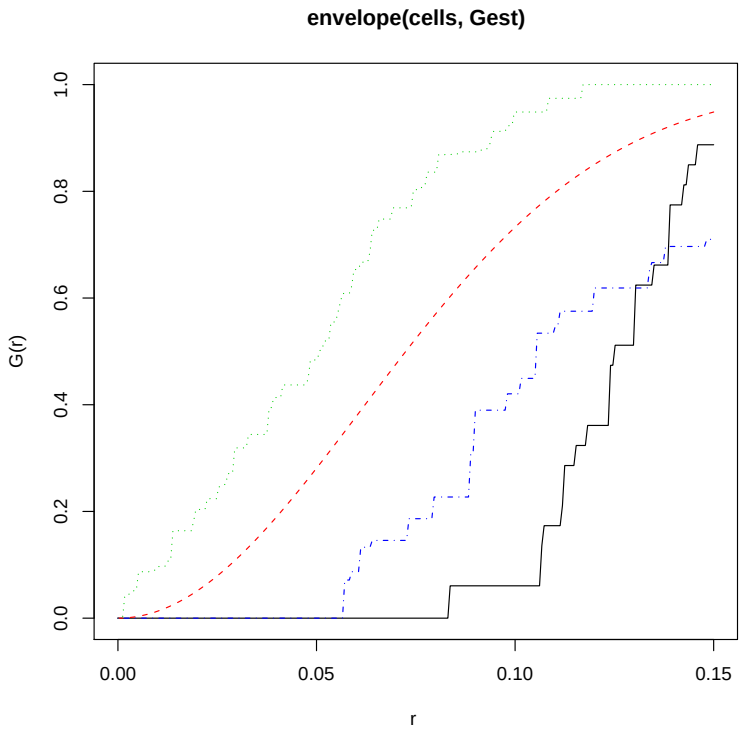
Test de CSR basé sur F





Test de CSR basé sur G





Analyse statistique des données spatiales VI

Christine Thomas-Agnan

Toulouse School of Economics

29 octobre 2012

Contexte

Variable dépendante : vecteur aléatoire Y (quantitatif, univarié) observé sur un nombre fini de zones représentées par leur centroide s_i.
Variable indépendante : vecteur aléatoire X (quantitatif , multivarié de dimension p), observé sur les mêmes zones.
En général on suppose de plus que X et Y sont gaussiens.
Modèle : $Y = \mu + \epsilon$ avec $\mu = \mathbb{E}(Y \mid X)$ (d'où $\mathbb{E}(\epsilon) = 0$ et $X \perp Y$), $\mathbb{V}ar(Y) = V$.
En général, on dispose d'une seule réalisation, c'est à dire de l'observation du couple (X, Y) en n sites.
Sans autre restriction sur ce modèle, on a n observations pour estimer $n + \frac{n(n+1)}{2}$ paramètres $\mapsto$ nécéssité de réduire le nombre de paramètres.

Modélisation de la tendance

On exprime la tendance comme une fonction

- des coordonnées géographiques
- de régresseurs + régresseurs spatialement décalés
- une combinaison des deux

Modèle non spatial WLS : V diagonale

$Y = X\beta + \epsilon$ avec $\mathbb{E}(\epsilon) = 0, \mathbb{V}ar(\epsilon) = \sigma^2 D$, où D est une matrice diagonale, $D = I_n$ correspondant au modèle OLS.
Présence de D : l'hétéroscédasticité est fréquente dans les variables spatiales.
exemple : T_i (resp : τ_i) est le taux de chomage observé (resp : théorique) dans la zone i et P_i est la population de la zone. Alors $var(T_i) = \frac{\tau_i(1-\tau_i)}{P_i}$ donc même si le taux de chomage est constant, il faut prendre des poids sur la diagonale de D proportionnels à $\frac{1}{P_i}$.

Estimateurs du maximum de vraisemblance dans le modèle WLS

$$
\begin{aligned}
\hat{\beta} &= (X'D^{-1}X)^{-1}X'D^{-1}Y \\
\mathbb{V}ar(\hat{\beta}) &= \sigma^2(X'D^{-1}X)^{-1} \\
\mathbb{V}ar(\hat{\epsilon}) &= \sigma^2 PDP', P = I_n - (X'D^{-1}X)^{-1}X'D^{-1} \\
\hat{\sigma}^2 &= \frac{(Y - X\hat{\beta})'D^{-1}(Y - X\hat{\beta})^2}{n-p}
\end{aligned}
$$

Test d'autocorrélation spatiale des résidus du modèle WLS

L'indice de Moran généralisé s'écrit comme l'indice de Moran appliqué aux résidus du modèle WLS : ceux-ci n'étant pas des observations mais des estimations, il faut ajuster les calculs de moments dans le contexte "free sampling".
Dans le cas $D = I_n$, on montre que sous l'hypothèse d'absence d'autocorrélation spatiale

$$\mathbb{E}(I) = -\frac{trA}{n-k},$$

où k est le nombre de colonnes de X et $A = (X'X)^{-1}X'WX$.

Test d'autocorrélation spatiale des résidus du modèle WLS

Si $k = 1$ (aucun régresseur), on retrouve la formule $\mathbb{E}(I) = -\frac{1}{n-1}$.
Si $k = 2$ (un seul régresseur), on obtient $\mathbb{E}(I) = -\frac{1+I_X}{n-2}$, où I_X est l'indice de Moran pour la variable X.

$$\mathbb{V}ar(I) = \frac{1}{(n-k)(n-k+2)}[S_1 + 2trA^2 - trB - \frac{2(trA)^2}{n-k}]$$

Un catalogue de modèles

Les modèles spatiaux consistent à introduire une variable spatialement décalée dans un modèle OLS ou WLS pour introduire de l'autocorrélation spatiale.

- modèle régressif croisé
- modèle LAG : spatial autorégressif
- modèle SDM : "spatial Durbin"
- modèle SEM : à erreurs spatialement corrélées
- modèle SAC : combine LAG et SEM
- modèle SARMA
- modèle CAR : conditionnel autorégressif

Etude de cas : Columbus

Nous utiliserons pour illustrer les notions un jeu de données économiques de Luc Anselin sur la ville de Columbus (Ohio, US) en 1980. Ce jeu de trouve dans le package spdep au format .Rdata et dans le package maptools au format .shp. La ville de Columbus est découpée en 49 quartiers pour lesquels on dispose de 18 attributs parmi lesquels nous avons choisi

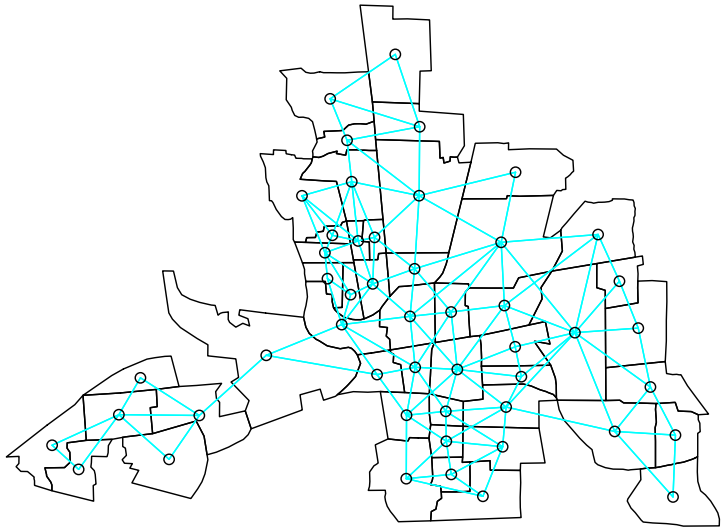
- HOVAL valeur immobilière en $ 1000
- INC revenu moyen des ménages en $ 1000
- CRIME nombre de cambriolages et vols de voitures pour 1000 habitants

On va chercher à expliquer la criminalité dans les quartiers par la valeur immobilière et le revenu des ménages.

Etude de cas : Columbus

La structure de voisinage est une matrice de contiguité de type "queen" notée W

```
plot(columbus)
plot(col.gal.nb,coord,add=TRUE)
```



Etude de cas : Columbus

Ajustement d'un modèle OLS

```
mod=lm(CRIME ~ INC + HOVAL, data = columbus)
Call:
lm(formula = CRIME ~ INC + HOVAL, data = columbus)
Residuals:
    Min      1Q  Median      3Q     Max
-34.418  -6.388  -1.580   9.052  28.649
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  68.6190     4.7355  14.490  < 2e-16
INC          -1.5973     0.3341  -4.780 1.83e-05
HOVAL        -0.2739     0.1032  -2.654   0.0109
Residual standard error: 11.43 on 46 degrees of freedom
Multiple R-squared: 0.5524,     Adjusted R-squared: 0.5329
F-statistic: 28.39 on 2 and 46 DF,  p-value: 9.34e-09
```

Etude de cas : Columbus

Test de Moran des résidus de ce modèle (test gaussien)

```
lm.morantest(mod,nb2listw(col.gal.nb))
    Global Moran's I for regression residuals

data:
model: lm(formula = CRIME ~ INC + HOVAL, data = columbus)
weights: col.listw

Moran I statistic standard deviate = 2.681, p-value = 0.00367
alternative hypothesis: greater
sample estimates:
Observed Moran's I        Expectation           Variance
       0.212374153       -0.033268284        0.008394853
```

Le modèle régressif croisé

Une première façon simple d'introduire de l'interaction entre unités spatiales est d'introduire une variable spatialement décalée parmi les explicatives :

$$Y = X\beta + WZ\delta + \epsilon,$$

avec $\mathbb{E}(\epsilon) = 0, \mathbb{V}ar(\epsilon) = \sigma^2 D$, où D est une matrice diagonale de pondération.
L'observation Y pour une unité spatiale donnée est donc ainsi expliquée par la valeur de X pour cette unité et par la moyenne des valeurs de Z pour les unités voisines. Par exemple, la production d'une région peut être expliquée par la disponibilité du travail et par le montant du capital public dans les zones voisines.

Le modèle régressif croisé

μ et V :

$$\mu = X\beta + WZ\delta$$

et

$$V = \sigma^2 D$$

L'ajustement de ce modèle peut se faire par MCO. Attention : si W est normalisée, il ne faut pas que la constante apparaisse à la fois dans X et dans Z sous peine de non identifiabilité.

Le modèle régressif croisé : application à Columbus

```
lm(formula = CRIME ~ INC + HOVAL + lag_INC + lag_HOVAL, data = columbus)

Residuals:
     Min       1Q   Median       3Q      Max
-36.2447  -7.6130   0.1881   7.8635  25.9821

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  74.0290     6.7218  11.013 3.13e-14
INC          -1.1081     0.3750  -2.955  0.00501
HOVAL        -0.2949     0.1014  -2.910  0.00565
lag_INC      -1.3834     0.5592  -2.474  0.01729
lag_HOVAL     0.2262     0.2026   1.116  0.27041

Residual standard error: 10.94 on 44 degrees of freedom
Multiple R-squared: 0.6085,     Adjusted R-squared: 0.5729
F-statistic: 17.09 on 4 and 44 DF,  p-value: 1.581e-08
```

Modèle spatial simultané autorégressif LAG

Le modèle LAG propose de prendre en compte dans la moyenne de Y sur une zone, outre les variables explicatives X, la moyenne de Y sur les zones voisines

$$Y = \rho WY + X\beta + \epsilon$$

WY est la variable endogène décalée et $(I - \rho W)Y$ la variable endogène filtrée.
Notons que si la matrice $(I - \rho W)$ est non singulière, ce modèle admet l'écriture équivalente suivante dite forme réduite ou DGP

$$Y = (I - \rho W)^{-1} X\beta + (I - \rho W)^{-1}\epsilon.$$

Modèle spatial simultané autorégressif LAG

μ et V :

$$\mu = (I - \rho W)^{-1} X\beta$$

$$\mathbb{V}ar(Y) = \sigma^2\{(I - \rho W')(I - \rho W)\}^{-1}.$$

Notons que cette variance implique une hétéroscédasticité même dans le cas où les erreurs sont homoscédastiques.

Modèle SDM Spatial Durbin

Une combinaison du modèle régressif croisé et du modèle LAG donne le modèle dit "Spatial Durbin"

$$Y = \rho WY + X\beta + WZ\delta + \epsilon$$

Forme réduite :

$$Y = (I - \rho W)^{-1}(X\beta + WZ\delta) + (I - \rho W)^{-1}\epsilon.$$

μ et V :

$$\mu = (I - \rho W)^{-1}(X\beta + WZ\delta)$$

et

$$\mathbb{V}ar(Y) = \sigma^2\{(I - \rho W')(I - \rho W)\}^{-1}.$$

Modèle à erreurs spatialement corrélées : SEM

Voyons à présent un autre modèle dans lequel l'autocorrélation spatiale intervient par l'intermédiaire d'un modèle LAG sur les erreurs.

$$Y = X\beta + \epsilon$$

$$\epsilon = \lambda W\epsilon + U,$$

où U est un bruit blanc Le paramètre λ mesure l'intensité de l'autocorrélation spatiale entre les résidus.

Modèle à erreurs spatialement corrélées : SEM

On a l'écriture équivalente

$$(I - \lambda W)Y = (I - \lambda W)X\beta + U.$$

Notons que si la matrice $(I - \lambda W)$ est non singulière, ce modèle admet la forme réduite suivante

$$Y = X\beta + (I - \lambda W)^{-1}U$$

μ et V :

$$\mu = X\beta$$

$$\mathbb{V}ar(Y) = \sigma^2\{(I - \lambda W')(I - \lambda W)\}^{-1}.$$

Notons que cette variance implique une hétéroscédasticité (les éléments de la diagonale ne sont pas constants) même dans le cas où les erreurs U sont homoscédastiques.

Modèle général : SAC

Ce modèle combine les modèles LAG et SEM de la façon suivante

$$Y = \rho W_1 Y + X\beta + \epsilon$$

$$\epsilon = \lambda W_2 \epsilon + U,$$

où U est un bruit blanc
Forme réduite :

$$Y = (I - \rho W_1)^{-1} X\beta + (I - \rho W_1)^{-1}(I - \lambda W_2)^{-1} U$$

μ et V :

$$\mu = (I - \rho W_1)^{-1} X\beta$$

et

$$V = [(I - \rho W_1')(I - \lambda W_2')(I - \lambda W_2)(I - \rho W_1)]^{-1}$$

Modèle SARMA

Le modèle MA à un paramètre s'écrit :
$Y_i = \mu + \rho \sum_{i=1}^{n} w_{ij}\epsilon_j + \epsilon_i$ où ϵ est un bruit blanc $\mathbb{E}(\epsilon) = 0, \mathbb{V}ar(\epsilon) = \sigma^2 D$ (D matrice diagonale).
alors $V = \sigma^2(I_n + \rho W)D(I_n + \rho W)'$.
On peut utiliser ce modèle combiné avec un modèle LAG

$$Y = \rho W_1 Y + X\beta + \epsilon$$

$$\epsilon = (I - \lambda W_2)u,$$

où U est un bruit blanc

Modèle conditionnel autorégressif CAR

Ce modèle est défini par une contrainte de type markovien sur la loi conditionnelle de Y_i sachant la valeur de Y pour les autres sites

$$Y_i \mid Y_1, \cdots, Y_{i-1}, Y_{i+1}, \cdots, Y_n \sim \mathcal{N}(\mu_i + \sum_{j=1}^{n} c_{ij}(Y_j - \mu_j), \tau_i^2),$$

où

- $C = (c_{ij})$ et $D = diag(\tau_1^2, \cdots, \tau_n^2)$ doivent satisfaire les deux conditions $D^{-1}C$ symétrique et $D^{-1}(I - C)$ définie positive.
- μ s'exprime par une combinaison linéaire d'explicatives $\mu = X\beta$

De façon équivalente dans le cas gaussien $Y \sim \mathcal{N}(X\beta, \tau^2(I - C)^{-1}D)$
Pour le modèle CAR à un paramètre $C = \rho W$ avec W matrice de voisinage, la variance s'écrit alors $V = \tau^2(I_n - \rho W)^{-1}D$.

Lien CAR-LAG

En faisant une hypothèse gaussienne, on peut écrire le modèle LAG

$$Y \sim \mathcal{N}((I-\rho W)^{-1}X\beta, \sigma^2\{(I-\rho W')(I-\rho W)\}^{-1})$$

et le modèle CAR

$$Y \sim \mathcal{N}(X\beta, \tau^2(I-C)^{-1})$$

d'où la même structure de covariance en posant
$C = \rho(W+W') - \rho^2 WW'$ et $\sigma = \tau$ mais des moyennes modélisées de façon différente.

Modèle LAG : contrainte sur les coefficients

Il y a dans ce modèle des contraintes sur le paramètre ρ qui sont dues à la nécéssité d'imposer la non singularité de $I - \rho W$. Soient ω_{min} et ω_{max} la plus petite et la plus grande valeurs propres de la matrice de voisinage W.
Si W est symétrique,

$$\frac{1}{\omega_{min}} < \rho < \frac{1}{\omega_{max}},$$

est une condition suffisante de non singularité.
Si W normalisée, alors $\omega_{max} = 1$ et $\rho \in [0, 1[$ est une condition suffisante de non singularité de $I - \rho W$.

Columbus : conditions sur paramètre ρ

La matrice W n'est pas symétrique mais est normalisée. Ses valeurs propres sont

```
eigen(Wmat, symmetric = FALSE,only.values = TRUE)$values
 [1]  1.000000e+00  9.687970e-01  9.388159e-01  8.748731e-01  8.476441e-01
 [6]  7.655969e-01  6.907270e-01 -6.519546e-01 -6.009133e-01  5.873411e-01
[11] -5.637492e-01  5.508182e-01  5.361444e-01 -5.042972e-01 -5.000000e-01
[16] -4.955955e-01 -4.823929e-01 -4.750630e-01 -4.452039e-01  4.418332e-01
[21] -4.222511e-01 -4.122630e-01 -3.889661e-01 -3.826030e-01 -3.655755e-01
[26] -3.544676e-01  3.372218e-01  3.237003e-01 -3.179893e-01 -3.094258e-01
[31]  2.852730e-01 -2.721972e-01 -2.556928e-01 -2.500000e-01 -2.289888e-01
[36] -2.066596e-01  1.975947e-01 -1.935817e-01 -1.820426e-01  1.704262e-01
[41] -1.468052e-01  1.245939e-01 -1.089779e-01 -8.386006e-02 -5.486559e-02
[46] -3.749353e-02  3.428778e-02  1.818743e-02  8.322744e-17
```

La condition sur le paramètre ρ est donc $-0.652 < \rho < 1$

EMV dans le modèle LAG

On montre aisément que les estimateurs MCO sont biaisés dans ce modèle et c'est pourquoi on doit recourir au maximum de vraisemblance.
Sous l'hypothèse de normalité des erreurs $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$, avec la notation $A(\rho) = (I - \rho W)$, la vraisemblance $L = L(y \mid \rho, \sigma^2)$ dans ce modèle s'écrit

$$
\begin{aligned}
L = \quad & f_Y(y) = f_\epsilon(\epsilon) \mid \det(\frac{\partial \epsilon}{\partial Y}) \mid = f_\epsilon(\epsilon) \mid \det(A(\rho)) \mid \\
= \quad & \frac{1}{(\sigma\sqrt{2\pi})^n} \exp(-\frac{\parallel \epsilon \parallel^2}{2\sigma^2}) \mid \det(A(\rho)) \mid \\
= \quad & \frac{1}{(2\pi)^{n/2}\sigma^n} \mid \det(A(\rho)) \mid ... \\
& .. \exp\{-\frac{1}{2\sigma^2}(y - A(\rho)^{-1}X\beta)'A(\rho)'A(\rho)(y - A(\rho)^{-1}X\beta)\},
\end{aligned}
$$

Calcul de LL dans le modèle LAG

D'où la log-vraisemblance $LL = \log L(y \mid \rho, \sigma^2)$

$$\begin{aligned} LL &= -\frac{n}{2}\log(2\pi) - n\log(\sigma) + \log(det((I - \rho W)) \\ &\quad -\frac{1}{2\sigma^2}(y - A(\rho)^{-1}X\beta)'A(\rho)'A(\rho)(y - A(\rho)^{-1}X\beta). \end{aligned}$$

avec $A(\rho) = (I - \rho W)$

EMV dans le modèle LAG

Si l'on dérive par rapport à σ, β et ρ, on peut obtenir l' expression explicite suivante de $\hat{\sigma}$ et $\hat{\beta}$ en fonction de $\hat{\rho}$

$$\hat{\sigma}^2(\rho) = \frac{1}{n}(y - A(\rho)^{-1}X\hat{\beta}(\rho))'A(\rho)'A(\rho)(y - A(\rho)^{-1}X\hat{\beta}(\rho)),$$

et

$$\hat{\beta}(\rho) = (X'X)^{-1}X'A(\rho)Y.$$

avec $A(\rho) = (I - \rho W)$

EMV dans le modèle LAG

Lorsqu'on reporte ces expressions dans le log-vraisemblance, on obtient ce qui s'appelle la log-vraisemblance concentrée qu'il reste à minimiser par rapport à ρ et qui vaut à constante près

$$\begin{aligned} \log L(y \mid \rho) &= \log(detA(\rho)) \\ &- \frac{n}{2} \log(y - A(\rho)^{-1} X\beta)' A(\rho)' A(\rho)(y - A(\rho)^{-1} X\beta)/n. \end{aligned}$$

avec $A(\rho) = (I - \rho W)$
Cette vraisemblance concentrée doit être optimisée numériquement et le problème principal est celui de l'évaluation du terme en log déterminant qui peut être couteux lorsque le nombre de sites devient grand : il faut alors recourir à des approximations de ce terme (il en existe plusieurs).

Columbus : EMV du modèle LAG

```
Call:lagsarlm(formula = CRIME ~ INC + HOVAL, data = columbus, listw = listw)

Residuals:
        Min          1Q      Median          3Q         Max
-37.4497095  -5.4565566   0.0016389   6.7159553  24.7107975

Type: lag
Coefficients: (asymptotic standard errors)
             Estimate Std. Error z value  Pr(>|z|)
(Intercept) 46.851429   7.314754  6.4051 1.503e-10
INC         -1.073533   0.310872 -3.4533 0.0005538
HOVAL       -0.269997   0.090128 -2.9957 0.0027381

Rho: 0.40389 LR test value: 8.4179 p-value: 0.0037154
Asymptotic standard error: 0.12071 z-value: 3.3459 p-value: 0.00082027
Wald statistic: 11.195 p-value: 0.00082027

Log likelihood: -183.1683 for lag model
ML residual variance (sigma squared): 99.164, (sigma: 9.9581)
Number of observations: 49
Number of parameters estimated: 5
AIC: 376.34, (AIC for lm: 382.75)
LM test for residual autocorrelation
test value: 0.19184 p-value: 0.66139
```

Interprétation des coefficients dans le modèle LAG

Dans un modèle OLS linéaire ordinaire $Y = X\beta + \epsilon$, les dérivées des coordonnées de Y par rapport à celles de X sont données par $\frac{\partial y_i}{\partial x_{ik}} = \beta_k$, pour tout i et k et $\frac{\partial y_i}{\partial x_{jk}} = 0$, pour tout k et $j \neq i$.
β_k s' interprète classiquement comme l'accroissement de $\mathbb{E}(Y)$ quand la k-ème variable explicative augmente d'une unité toutes choses égales par ailleurs. Plus précisément, l'augmentation d'une unité de x_{ik}

- n'a aucun effet sur Y_j pour $j \neq i$
- a le même effet sur Y_i que l'augmentation d'une unité de $x_{i'k}$ sur $Y_{i'}$

Interprétation des coefficients dans le modèle LAG

L'écriture de LAG par composante est $y_i = \sum_{t=1}^{p} S_t(W)_{it} x_t + \tilde{\epsilon}_i$, où p est le nombre de variables explicatives, x_t est la t-ème colonne de la matrice X et $\tilde{\epsilon} = (I - \rho W)^{-1}\epsilon$.
Alors, les dérivées partielles de $\mathbb{E}(y_i)$ par rapport à x_{jt} sont

$$\frac{\partial \mathbb{E}(y_i)}{\partial x_{jt}} = S_t(W)_{ij}.$$

On remarque d'abord que la dérivée croisée de la i-ème composante $\mathbb{E}(y_i)$ par rapport à x_{jt} pour $j \neq i$ n'est plus nécéssairement nulle mais égale à $S_t(W)_{ij}$.
On en déduit qu'un changement sur l'une des variables explicatives pour l'individu i va affecter non seulement y_i mais aussi tous les y_j : un changement de la variable explicative dans une unité spatiale peut se répercuter sur les Y de toutes les autres unités.

Interprétation des coefficients dans le modèle LAG

De plus, l'effet sur $\mathbb{E}(y_i)$ de l'accroissement d'une unité de la i-ème composante de la t-ème variable explicative x_{it} n'est plus nécéssairement constant sur les i car égal à $S_t(W)_{ii}$. On définit alors trois mesures résumant ces effets pour chaque variable explicative t :
L'**impact direct moyen** $ADI = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \mathbb{E}(y_i)}{\partial x_{it}}$ mesure la moyenne de l'effet de l'accroissement d'une unité de la variable t pour l'individu i sur $\mathbb{E}(Y_i)$ pour ce même individu.
L'**impact moyen total** $ATI = \frac{1}{n}\sum_{i,j} \frac{\partial \mathbb{E}(y_i)}{\partial x_{jt}}$, mesure l'effet moyen sur $\mathbb{E}(Y)$ de l'accroissement de x_t d'une unité pour tous les individus. C'est la moyenne sur les individus i de l'impact total de cet accroissement sur $\mathbb{E}(Y_i)$ qui est mesuré par $\sum_j \frac{\partial \mathbb{E}(y_i)}{\partial x_{jt}}$.

Interprétation des coefficients dans le modèle LAG

L'**impact indirect moyen** ou "spillover" $AII = \frac{1}{n}\sum_{i\neq j}\frac{\partial \mathbb{E}(y_i)}{\partial x_{jt}}$ mesure la moyenne de l'effet indirect sur chaque composante de $\mathbb{E}(Y)$. L'effet indirect sur $\mathbb{E}(Y_i)$ est mesuré par l'effet de l'accroissement d'une unité de x_{jt} pour tous les individus $j \neq i$.

L'impact moyen total est la somme de l'impact direct moyen et de l'impact indirect moyen : $ATI = ADI + AII$
En raison de l'effet non linéaire de ρ, ces mesures d'impact sont des fonctions non linéaires des paramètres : on recourt à des méthodes de Monte Carlo pour tester leur significativité.

Columbus : calcul des effets

```
$direct.eff
       INC      HOVAL
-1.1225155 -0.2823163

$indirect.eff
       INC      HOVAL
-0.6783818 -0.1706152

$total.eff
       INC      HOVAL
-1.8008973 -0.4529315
```

Comparer aux coefficients

```
Coefficients:
             Estimate
INC         -1.073533
HOVAL       -0.269997
```

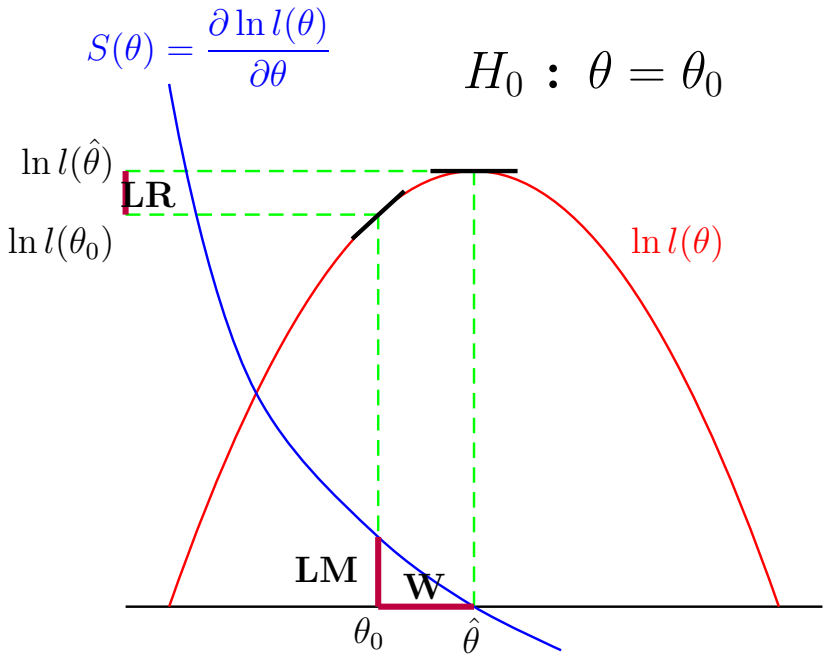
Les trois tests sur les coefficients

Il existe trois tests classiques pour tester $H_0 : \theta = \theta_0$ contre l'alternative $H_0 : \theta \neq \theta_0$, où θ peut-être soir l'un des paramètres β soit le paramètre ρ

- test de Wald : TW
- test du rapport de vraisemblance LR
- test du multiplicateur de Lagrange LM

Ces trois tests sont asymptotiquement équivalents mais à distance finie on a $TW \geq LR \geq LM$. Le test de Wald requiert l'estimation des paramètres sous l'hypothèse alternative, le test du multiplicateur de Lagrange requiert l'estimation des paramètres sous l'hypothèse nulle et le test du rapport de vraisemblance requiert les deux estimations.

Les trois tests sur les coefficients : graphique



Le test du rapport de vraisemblance

$\hat{\theta}$ estimateur du maximum de vraisemblance de θ sous H_1
Sous H_0 :

$$LR_\theta = -2(LL(\theta_0) - LL(\hat{\theta})) \rightarrow \chi^2(1)$$

Le test du score ou LM

Fonction Score : $S(\theta_0) = \dfrac{\partial LL(\theta_0)}{\partial \theta}$

Sous H_0

$$LM_\theta = \frac{S^2(\theta_0)}{nI(\theta_0)} \rightarrow \chi^2(1)$$

où $I(\theta)$ est ma matrice d'information de Fisher basée sur une observation.

Le test de Wald

Le test de Wald est basé sur $(\hat{\theta} - \theta_0)$
Sous H_0

$$TW_\theta = \frac{(\hat{\theta} - \theta_0)^2}{\hat{V}(\hat{\theta})} \rightarrow \chi^2(1)$$

Le test du coefficient ρ : LM-lag

$H_0 : \rho = 0$

c'est un test du modèle OLS sous H_0 contre le modèle alternatif LAG
on fait un test de type LM avec la statistique

$$LM_{LAG} = \frac{[\hat{\varepsilon}' Wy/\hat{\sigma}^2]^2}{T_{sar}}$$

où $\hat{\varepsilon}$ résidus du modèle OLS, $\hat{\sigma}^2$ variance résiduelle estimée par OLS et $T_{sar} = [(WX\hat{\beta})' P(WX\hat{\beta})]/\hat{\sigma}^2 + trace((W + W')W)$

Sous H_0

$$LM_{LAG} \to \chi^2(1)$$

Il existe une version robuste de ce test RLM_{LAG}.

Columbus : le test de OLS contre LAG

```
lm.LMtests(ols_mod, col.listw, test="LMlag")

        Lagrange multiplier diagnostics for spatial dependence

data:
model: lm(formula = CRIME ~ INC + HOVAL, data = columbus)
weights: col.listw

LMlag = 7.8557, df = 1, p-value = 0.005066
```

Le test rejette le modèle non spatial.

Prédiction dans le modèle LAG

Pour le calcul de $\hat{Y}_i$ pour les unités spatiales i de l'échantillon, on dispose de trois alternatives

1. $\hat{Y}^{TC} = (I - \hat{\rho}W)^{-1}X\hat{\beta}$
2. $\hat{Y}^{TS} = X\hat{\beta} + \hat{\rho}WY$
3. $\hat{Y}^{BP} = \hat{Y}^{TC} - \text{Diag}(Q)^{-1}(Q - \text{Diag}(Q))(Y - \hat{Y}^{TC})$, où $Q = \frac{1}{\sigma^2}(I - \rho W')(I - \rho W)$.

La meilleure prédiction (BLUP) est donnée par $\hat{Y}^{BP}$ et $\hat{Y}^{TS}$ a un efficacité relative assez bonne ; par contre $\hat{Y}^{TC}$ est mauvais.

Columbus : modèle spatial Durbin

```
    Call:lagsarlm(formula = CRIME ~ INC + HOVAL,
    data = columbus, listw = col.listw, type = "mixed")

Type: lag
Coefficients: (asymptotic standard errors)
             Estimate Std. Error z value  Pr(>|z|)
(Intercept) 45.592896  13.128680  3.4728 0.0005151
INC         -0.939088   0.338229 -2.7765 0.0054950
HOVAL       -0.299605   0.090843 -3.2980 0.0009736
lag_INC     -0.618375   0.577052 -1.0716 0.2838954
lag_HOVAL    0.266615   0.183971  1.4492 0.1472760

Rho: 0.38251 LR test value: 4.1648 p-value: 0.041272
Asymptotic standard error: 0.16237 z-value: 2.3557 p-value: 0.018488
Wald statistic: 5.5493 p-value: 0.018488

Log likelihood: -182.0161 for lag model
ML residual variance (sigma squared): 95.051, (sigma: 9.7494)
Number of observations: 49
Number of parameters estimated: 7
AIC: 378.03, (AIC for lm: 380.2)
LM test for residual autocorrelation
test value: 0.101 p-value: 0.75063
```

Modèle à erreurs spatialement corrélées : SEM

Rappel

$$Y = X\beta + \epsilon$$

$$\epsilon = \lambda W\epsilon + U,$$

Il y a dans ce modèle des contraintes sur le paramètre λ qui sont les mêmes que les contraintes sur ρ dans le modèle LAG.
Si l'on pose $A(\lambda) = I - \lambda W$, on a alors $Y = X\beta + A(\lambda)^{-1}\epsilon$ et $\epsilon = A(\lambda)(Y - X\beta)$.

Modèle à erreurs spatialement corrélées : SEM

Sous l'hypothèse de normalité des erreurs $U \sim \mathcal{N}(0, \sigma^2 I)$, la vraisemblance de Y s'écrit alors :

$$\begin{aligned} L = f_Y(y) &= f_\epsilon(\epsilon) \mid det(\frac{\partial \epsilon}{\partial Y}) \mid \\ &= f_\epsilon(\epsilon) \det(A(\lambda)) \\ &= \frac{1}{(\sigma\sqrt{2\pi})^n} \exp(-\frac{\| \epsilon \|^2}{\sigma^2}) \mid \det(A(\lambda)) \mid \end{aligned}$$

d'où la log-vraisemblance

$$\begin{aligned} LL = \ln(L) &= -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 + \ln(det(A(\lambda))) \\ &- \frac{1}{2\sigma^2}(Y - X\beta)'A(\lambda)'A(\lambda)(Y - X\beta) \end{aligned}$$

EMV dans le modèle SEM

A λ fixé, la maximisation de la log-vraisemblance se fait de façon analytique et on obtient :

$$\hat{\beta}(\lambda) = (X'A(\lambda)'A(\lambda)X)^{-1}X'(A(\lambda)'A(\lambda))Y$$

$$\hat{\sigma}^2(\lambda) = \frac{1}{n} \parallel Y - X\hat{\beta}(\lambda) \parallel^2_{A(\lambda)'A(\lambda)}$$

Lorsqu'on reporte ces expressions dans la log-vraisemblance, on obtient la log-vraisemblance "concentrée" qu'il reste à minimiser par rapport à λ et qui vaut à constante près

$$\log L(y \mid \lambda) = \log(det((I - \lambda W)) - \frac{n}{2} \log \parallel Y - X\hat{\beta}(\lambda) \parallel^2_{A(\lambda)'A(\lambda)} .$$

Cette vraisemblance concentrée doit être optimisée numériquement (pb du terme en log déterminant)

Columbus : EMV du modèle SEM

```
Call:
spautolm(formula = CRIME ~ INC + HOVAL, data = columbus, listw = col.listw)

Residuals:
      Min        1Q    Median        3Q       Max
-34.45950  -6.21730  -0.69775   7.65256  24.23631

Coefficients:
             Estimate Std. Error z value  Pr(>|z|)
(Intercept) 61.053619   5.314875 11.4873 < 2.2e-16
INC         -0.995473   0.337025 -2.9537 0.0031398
HOVAL       -0.307979   0.092584 -3.3265 0.0008794

Lambda: 0.52089 LR test value: 6.4441 p-value: 0.011132

Log likelihood: -184.1552
ML residual variance (sigma squared): 99.98, (sigma: 9.999)
Number of observations: 49
Number of parameters estimated: 5
AIC: 378.31
```

Columbus : EMV du modèle SEM, syntaxe alternative

```
Call:
errorsarlm(formula = CRIME ~ INC + HOVAL, data = columbus, listw = col.listw)

Residuals:
      Min        1Q    Median        3Q       Max
-34.45950  -6.21730  -0.69775   7.65256  24.23631

Type: error
Coefficients: (asymptotic standard errors)
             Estimate Std. Error z value  Pr(>|z|)
(Intercept) 61.053618   5.314875 11.4873 < 2.2e-16
INC         -0.995473   0.337025 -2.9537 0.0031398
HOVAL       -0.307979   0.092584 -3.3265 0.0008794

Lambda: 0.52089 LR test value: 6.4441 p-value: 0.011132
Asymptotic standard error: 0.14129 z-value: 3.6868 p-value: 0.00022713
Wald statistic: 13.592 p-value: 0.00022713

Log likelihood: -184.1552 for error model
ML residual variance (sigma squared): 99.98, (sigma: 9.999)
Number of observations: 49
Number of parameters estimated: 5
AIC: 378.31, (AIC for lm: 382.75)
```

Notons que la syntaxe “spautolm” autorise des poids d’hétérsocédasticité contrairement à “errorsarlm”.

Le test LM du coefficient λ : LM-err

$H_0 : \lambda = 0$

c'est un test du modèle OLS sous H_0 contre le modèle alternatif SEM
on fait un test de type LM avec la statistique

$$LM_{ERR} = \frac{[\hat{\varepsilon}' W \hat{\varepsilon}/\hat{\sigma}^2]^2}{T_{sem}}$$

où $T_{sem} = tr[(W' + W)]$, $\hat{\varepsilon}$ résidus du modèle OLS, et $\hat{\sigma}^2$ estimateur OLS de σ^2.

Sous H_0

$$LM_{ERR} \rightarrow \chi^2(1)$$

Il existe une version robuste de ce test RLM_{ERR}.

Columbus : test de OLM contre SEM

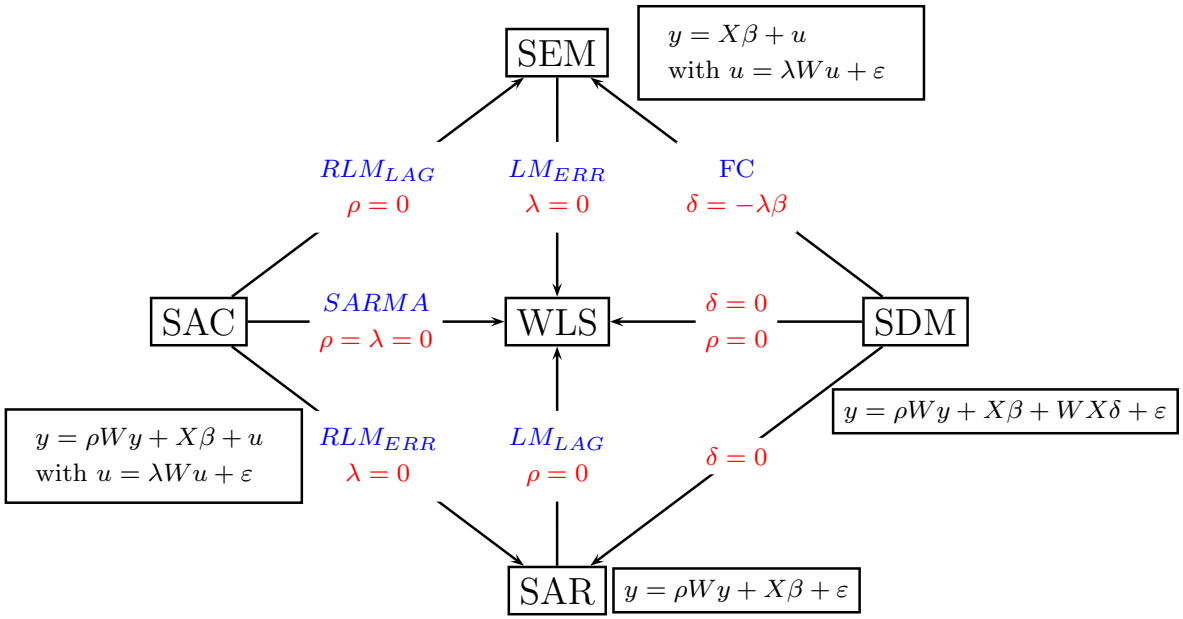
```
        Lagrange multiplier diagnostics for spatial dependence

data:
model: lm(formula = CRIME ~ INC + HOVAL, data = columbus)
weights: col.listw

LMerr = 4.6111, df = 1, p-value = 0.03177
```

Ce test est moins significatif que le LM_{LAG} donc on va préferer un modèle LAG.

Synthèse sur les tests



Stratégie par tests

1. Ajuster un modèle WLS puis un modèle mixte et choisir
2. faire les tests LM_{ERR} et LM_{LAG}
3. si aucun des deux n'est significatif, garder le modèle de l'étape 1
4. si un seul est significatif : si c'est LM_{ERR}, garder un modèle SEM, si c'est LM_{LAG}, garder un modèle LAG
5. si les deux sont significatifs, faire les tests RLM_{ERR} et RLM_{LAG}

Stratégie par tests

Suite

1. si seul RLM_{ERR} est significatif, choisir SEM
2. si seul RLM_{LAG} est significatif, choisir LAG
3. si les deux sont significatifs, choisir SAC
4. si aucun, choisir LAG (resp SEM) lorsque LM_{LAG} est plus significatif que LM_{ERR} (resp LM_{ERR} est plus significatif que LM_{LAG})

Stratégie par critères

La stratégie par critère consiste à minimiser le critère d'Akaiké ou le critère de Schwartz qui s'expriment en fonction de la log-vraisemblance et le nombre de paramètres k

- Akaiké : $AIC = -2\log(L) + 2k$
- Schwartz : $BIC = -2\log(L) + k\log(n)$

Columbus : choix de modèle par tests

```
lm.LMtests(ols_mod, col.listw,test="all")

LMerr = 4.6111, df = 1, p-value = 0.03177

LMlag = 7.8557, df = 1, p-value = 0.005066

RLMerr = 0.0335, df = 1, p-value = 0.8547

RLMlag = 3.2781, df = 1, p-value = 0.07021

SARMA = 7.8892, df = 2, p-value = 0.01936
```

l'algorithme choisit le modèle LAG.

Columbus : choix de modèle par critère

```
AIC(ols_mod,ols_croise,lagmodel,semmod,durbin)
           df      AIC
ols_mod     4 382.7545
ols_croise  6 380.1970
lagmodel    5 376.3366
semmod      5 378.3104
durbin      7 378.0322
```

le critère d'Akaiké choisit le modèle LAG.

Conclusion

- lorsqu'un test met en évidence de l'autocorrélation spatiale dans les résidus d'un modèle WLS, on peut commencer par introduire d'autres variables exogènes ou des exogènes spatialement décalées avant de se tourner vers un modèle spatial
- comment articuler le choix de variables et le choix de famille de modèle ?
- problème de niveau pour les tests consécutifs