Outils et modèles de statistique spatiale

Christine Thomas-Agnan

Toulouse School of Economics
GREMAQ: Groupe de Recherche en Economie Mathématique et Quantitative
France

PEPI INRA

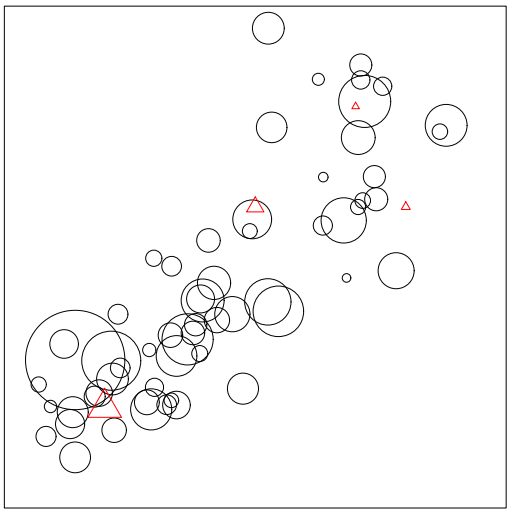
Jeu de données 1 : pompiers de Toulouse

SDIS 31, "Service Départemental d'Incendie et de Secours de la Haute-Garonne"
Positions et caractéristiques d'environ $20,000$ sinistres dans une zone autour de la ville de Toulouse en 2004
sinistre = tout évenement donnant lieu à un appel à une caserne : feux, accidents, toute sorte d' incidents.
localisation du sinistre, charge de travail associée (durée en minutes entre l'arrivée des premiers pompiers sur le site et départ du dernier véhicule multiplié par le nombre de pompiers sur le site). Localisation des $J = 6$ casernes existantes $(s_j, j = 1, \cdots, J)$ et leur nombre respectif de pompiers z_j.
Le nombre médian de pompiers par caserne est d'environ 64 et la charge médiane d'environ 174 minutes dans cette base.

Jeu de données 1 : pompiers de Toulouse

Autres caractéristiques des sinistres : nombre de véhicule mobilisés, type de sinistre (feu, accident, autre).
Objectif : Le SDIS 31 voudrait implanter un (ou plusieurs) caserne (s) pour réduire le temps d'arrivée sur le site d'un sinistre et réduire la charge totale des casernes existantes.
On suppose que le nombre de pompiers dans la nouvelle caserne (numérotée $J+1$) est fixé (égal à 60 dans l' application).

Jeu de données 1 : pompiers de Toulouse



Jeu de données 2 : les collèges de Midi-Pyrénées

226 collèges de Midi-Pyrénées (2003-2004)
les collèges sont localisés au centroïde de la commune
parmi les variables disponibles : the number of students per class, the cost per student and the occupancy rate which is the number of students in the school divided by the number of students the school has been designed for.

Jeu de données 2 : les collèges de Midi-Pyrénées

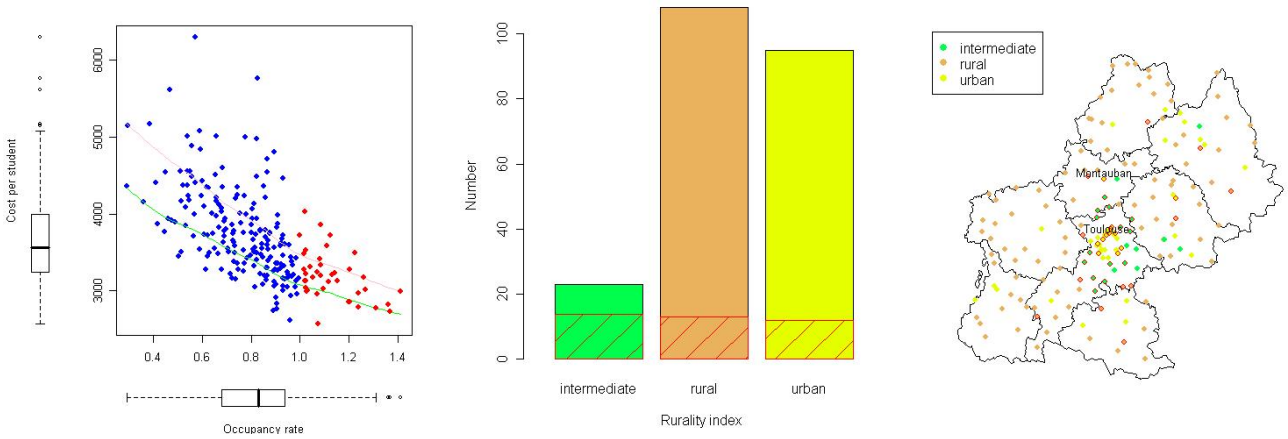


FIGURE: Scatterplot of cost per student versus occupancy rate and barplot of rurality index : selection of schools with occupancy rate greater than 1.

Jeu de données 2 : les collèges de Midi-Pyrénées

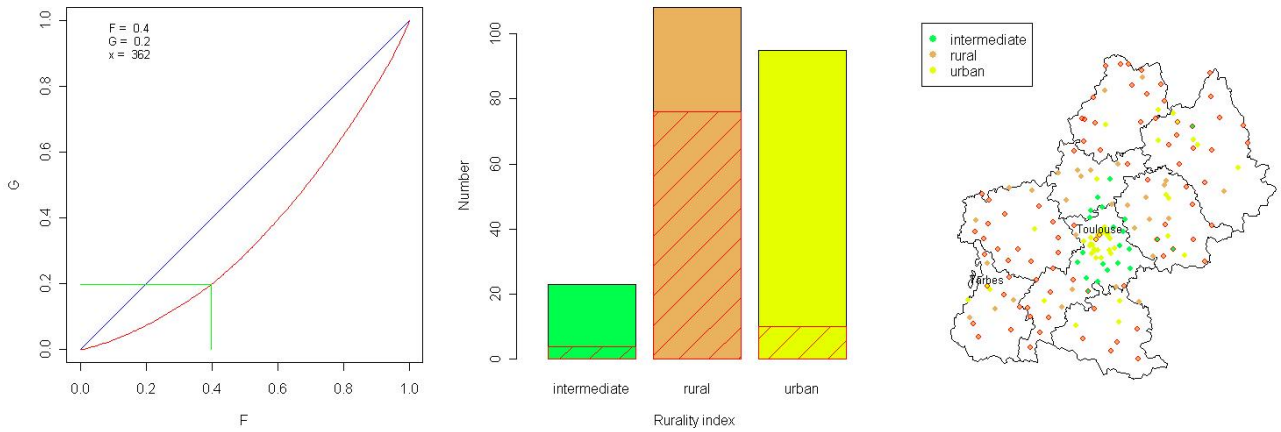


FIGURE: Lorenz curve and Gini index for the number of students : selection of the first 40 % of schools sorted by increasing number of students.

Jeu de données 3 : collèges de Midi-Pyrénées version 2

deuxième version du même jeu de données aggrégées au niveau "pseudo-canton"
A "canton" is a french administrative subdivision which usually is an aggregate of several communes. However, large "communes" may be divided into several cantons and in that case, a pseudo-canton corresponds simply to the commune. In the other cases, pseudo-cantons correspond to cantons.
155 pseudo-cantons with at least one public school.
variables : the mean number of students per class, the mean cost per student and the mean occupancy rate together with the number of schools in the pseudo-cantons and a rurality index. The rurality index takes the value 1 if the ratio of the number of rural communes in the pseudo-canton to the number of communes is larger than 1/2, and 0 otherwise.

Jeu de données 3 : collèges de Midi-Pyrénées version 2

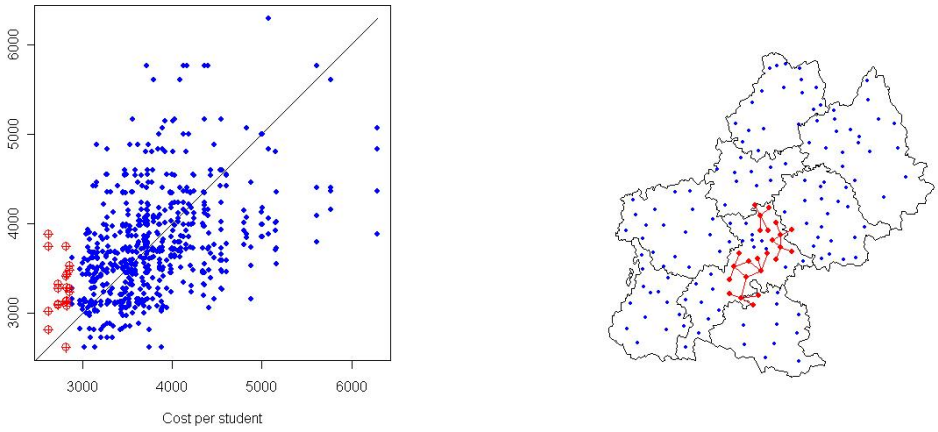


FIGURE: Neighbor plot for cost per student : selection of small costs.

Jeu de données 4 : le prix des maisons à Columbus (Ohio)

Nous utiliserons 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

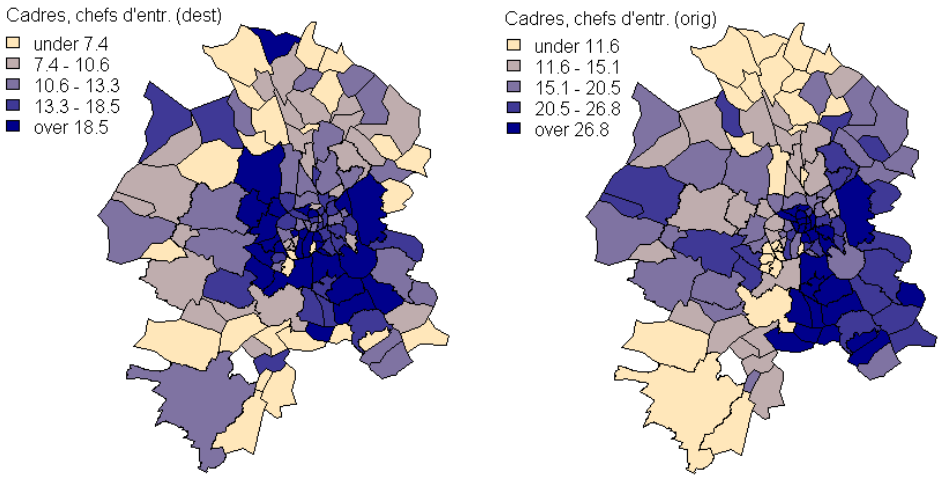
- HOVAL valeur immobilière en $ 1000
- INC revenu moyen des ménages en $ 1000
- CRIME nombre de cambriolages et vols de voitures pour 1000 habitants

Jeu de données 5 : les navettes domicile-travail

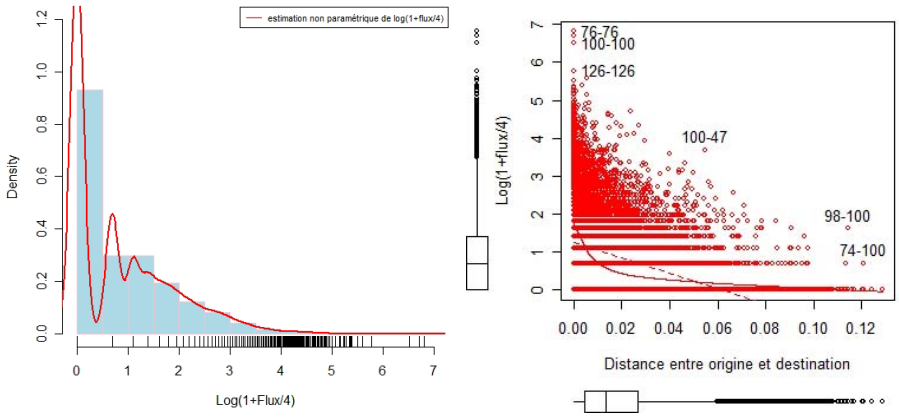
Les données de flux sont issues du recensement de la population de 1999 : sondage au quart, fichier d'emploi au lieu de travail
" îloté ", avec localisation infra-communale des emplois à l'îlot.
Champ : sont exclus les actifs travaillant dans le secteur d'activité de la défense, et les actifs ayant déclaré travailler en des lieux variables ou chez des particuliers (lieu de travail méconnu, ou ne se prêtant pas à une localisation unique et précise).
Variables explicatives :

- caractéristiques de la population active (catégorie socioprofessionnelle, secteur d'activité, tranche d'âge) issues du même fichier que les données de flux, et portant sur le même champ.
- caractéristiques des logements : exploitation exhaustive du recensement de la population de 1999.
- surface des zones, latitude et longitude des centroïdes (utilisées pour calculer la distance entre deux zones) : issues des fonds de cartes de l'IGN.

Jeu de données 4 : les navettes domicile-travail



Jeu de données 4 : Marginale des flux et lien avec distance



Quatre types de données spatiales

On distingue quatre grands types de données géoréférencées :

- les données de type **géostatistique** :
  L'outil de modélisation des données géoréférencées est le champ aléatoire. Lorsqu'une caractéristique $X(s, \omega)$ d'une unité statistique est mesurée en la position s pour la réalisation ω, on notera X_s la variable aléatoire associée, où l'indice s varie dans une partie $\mathcal{D}$ de $\mathbb{R}^d$ (contenant un rectangle de volume strictement positif).
  Les données sont en général des observations du champ en des points discrets et **déterministes** s_i de $\mathcal{D}$. Exemple en météorologique : le champ "vitesse du vent" peut être défini en tout point d'une zone géographique mais est mesuré en un nombre fini de stations météo.

Quatre types de données spatiales

- les données de type **treillis** ou **latticiel** : X_s n'a de sens que sur une collection dénombrable de zones inclues dans $\mathcal{D} \in \mathbb{R}^d$, souvent représentées par leur centroïde. C'est le cas pour la plupart des données de type économique. Exemple : le champ "revenu par foyer fiscal" n'a de sens qu'en moyenne sur une zone telle qu'une commune ou un canton.
- les données de type **processus ponctuel** ou "semis de points". Ce qui les distingue des deux autres cas est le fait que la position des observations est à présent aléatoire. Si seule la position est observée, il s'agit d'un simple processus ponctuel. Si de plus, une variable (que l'on appelle la marque) est observée en ces positions aléatoires, il s'agit d'un processus ponctuel marqué.
- les données bilocalisées (flux)

Les logiciels d'analyse spatiale

- Les SIG ou Geographic Information System : ARCINFO, MAPINFO, ARCVIEW (version légère de ARCINFO), SAS/GIS, CARTE ET BASE, ASTEROP, GRASS
- Les boite 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)

Les logiciels d'analyse spatiale

- Matlab : boite à outils spatialeconometrics.com (Le Sage) pour les modèles d'économétrie spatiale
- les packages de R :GeoXp (Toulouse, exploratoire interactif), spdep (Econométrie spatiale) , geoR (géostatistique) , spatstat (semis de points), rgdal, sp, maptools (lecture, gestion, cartographie de données spatiales)

Mes articles en géostatistique

- Cressie N., Perrin O. and Thomas-Agnan C., 2005. Likelihood based estimation for Gaussian MRFs. Statistical Methodology 2, pp. 1-16
- Cressie N., Perrin O. and Thomas-Agnan C., 2005. Doctors's prescribing patterns in the Midi-Pyrénées region of France : point process aggregation. In : "Case studies in spatial point process models", Lecture Notes in Statistics 185, Springer Verlag.
- Elogne S., Perrin O., Thomas-Agnan C., 2008. Non parametric estimation of smooth stationary covariance functions by interpolation methods, Statistical Inference for Stochastic Processes, 12(2), 177-205.
- Laurent T., Ruiz-Gazen A. and Thomas-Agnan C., GeoXp : an R package for exploratory spatial data analysis, to appear in Journal of Statistical Software.

Mes articles en processus ponctuels

- Cucala, L. and Thomas-Agnan, C., 2006. Spacings-based tests for spatial randomness and coordinate-invariant procedures. Annales de l'. I.S.U.P., 50, no 1-2, 31-45.
- Bonneu F. and Thomas-Agnan C., 2009, Spatial point process models for location-allocation problems, Computational Statistics and Data Analysis, 53 (8), 3070-3081.
- Bonneu F. Coelho S., Magrini M.B., and Thomas-Agnan C. (2011) The decision to create new classes or to close existing ones : A study of distance based choices of agricultural training establishments, to appear in Environment and Planning B, Planning and Design.
- Bonneu F. and Thomas-Agnan C., Concentration measures for micro-geographic data based on second order characteristics of spatial point processes, WP 2011.

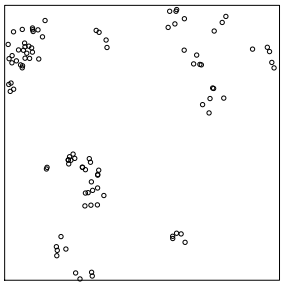
Mes articles en économétrie spatiale

- Laurent T., Goulard M. and Thomas-Agnan C., About predictions in spatial SAR models : optimal and almost optimal strategies, WP 2011.
- J.P. Lesne, Ruiz-Gazen A., Tranger H. and Thomas-Agnan C., Predicting population annual growth rate with spatial models, WP2011.
- LeSage J. and Thomas-Agnan C., Impacts in spatial interaction models, WP2011.
- Laurent T., LeSage J., Rousseau C. and Thomas-Agnan C., Modelling home to work commuting data in the region of Toulouse with spatial interaction models, WP2011.

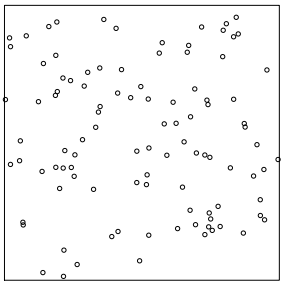
Phénomènes à modéliser

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 CSR, détecter régularité ou agrégation, ajuster un modèle, détecter agrégats.

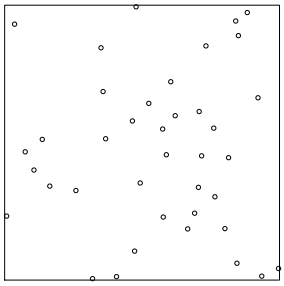
Exemples



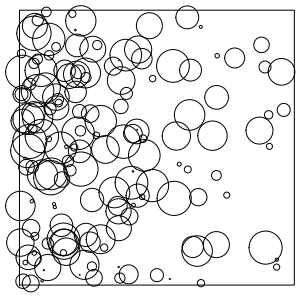
(a) Agrégé



(b) Poisson Homogène



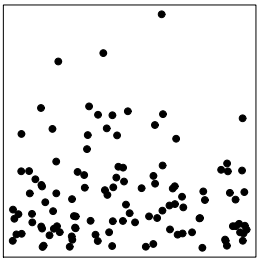
(c) Régulier



(d) Poisson inhomogène marqué

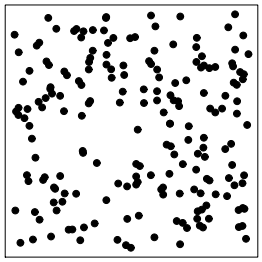
Stationarité - Isotropie

Homogénéité = stationnarité = propriétés invariantes par translation,
Isotropie = propriétés invariantes par rotation.



(e) Non stationnaire

(f) Anisotrope



(g) Stationnaire et isotrope

Premier ordre : intensité

L'intensité est l'analogue pour le processus ponctuel de l'espérance pour une variable aléatoire.
Mesure d'intensité

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

$\Lambda(B)$ représente le nombre moyen de points du processus dans B.
Si le processus est stationnaire, cette mesure est proportionnelle à l'aire du domaine B et le facteur de proportionalité, λ, appellé intensité, représente le nombre moyen de points du processus par unité de surface.
Estimateur de Λ
$\hat{\Lambda}(B) = \sum_{\xi \in X} \mathbf{1}(\xi \in B)$ désigne le nombre de points du processus dans B.

Premier ordre : intensité

Fonction intensité

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

Cette fonction λ porte le nom de fonction d'intensité du processus ponctuel.
$\lambda(u)du$ est la probabilité d'occurence d'un point dans la boule infinitésimale de centre u et de volume $|du|$
Si le processus est stationnaire, la fonction d' intensité est constante.
Lorsque l'intensité est constante, on dit que le processus est homogène.

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

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

où W est la fenêtre d'observation, $\mid W \mid$ son aire et N 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.

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

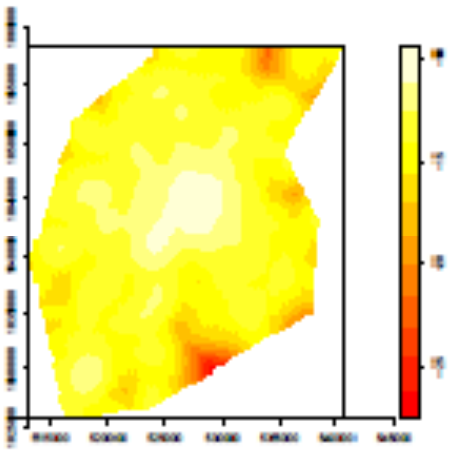
L' estimateur de Diggle est, de même qu'un estimateur non paramétrique de densité, peu sensible au choix du noyau K.
Le choix de la largeur de bande ou fenêtre h permettant de minimiser l'erreur quadratique moyenne intégrée

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

se fait selon les mêmes méthodes que dans le cadre de l'estimation de densité [Scott, 1992].

Estimation de l'intensité des sinistres

Logarithme de l'intensité des sinistres en juin



Estimation de l'intensité avec covariables

On peut spécifier un modèle paramétrique pour l'intensité usuellement de la forme

$$\lambda(s) = \exp(\sum_k \beta_k X_k)$$

où X_k sont des coavariables connues sur le domaine.
On ajuste les coefficients β_k par pseudo-maximum de vraisemblance.
Exemple des pompiers : on utilise la population.

Processus de Poisson

Homogène : Modèle de processus le plus élémentaire qui permet de générer des semis de points avec une répartition spatiale totalement aléatoire.

Inhomogène : Modèle le plus simple de processus inhomogène sans interaction entre les points.

Définition :
1. $\Lambda(B)$ suit une loi de Poisson de moyenne $\int_B \lambda(u)du$,
2. Sachant $\Lambda(B) = n$, les n points de B sont indépendants et issus de la loi de densité $\frac{\lambda(x)}{\int_B \lambda(u)du}$.
3. Si B et B' sont disjoints, $\Lambda(B)$ et $\Lambda(B')$ sont indépendants.

Processus de Poisson

Relation densité - intensité : si le processus est de Poisson inhomogène d'intensité λ, conditionnellement à $N = n$, les n localisations sont alors i.i.d. et leur densité marginale est liée à l'intensité par

$$\mathbb{E}(N) f(s) = \lambda(s) \mid \Omega \mid$$

Les probabilités fini-dimensionnelles de ce processus sont données par

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

Caractéristiques du 2nd ordre

Du fait de la propriété (3), le processus de Poisson implique une **absence d'interaction** entre les localisations des 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**) et ceux pour lesquels les évènements ont tendance à se repousser (**régularité**).

Caractéristiques du 2nd ordre : mesure d'ordre 2

Mesure d'ordre 2 :

$$\Lambda^{(2)}(B_1 \times B_2) = \mathbb{E}\left[\sum_{\xi,\eta\neq\xi\in X} \mathbf{1}(\xi \in B_1)\mathbf{1}(\eta \in B_2)\right] = \int_{B_1}\int_{B_2} \rho^{(2)}(u,v)dvdu.$$

$\rho^{(2)}(u,v)dudv$ s'interpréte comme la probabilité d'occurence jointe d'un point dans la boule infinit. de centre u et de vol. $|du|$ et d'un point dans la boule infinit. de centre v et de vol. $|dv|$.
Pour un processus stationnaire, la fonction $\rho^{(2)}(x,y)$ ne dépends que de $x-y$.
Si de plus le processus est isotrope, elle ne dépends que de $\| x-y \|$.

Caractéristiques du 2nd ordre : corrélation des paires

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

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

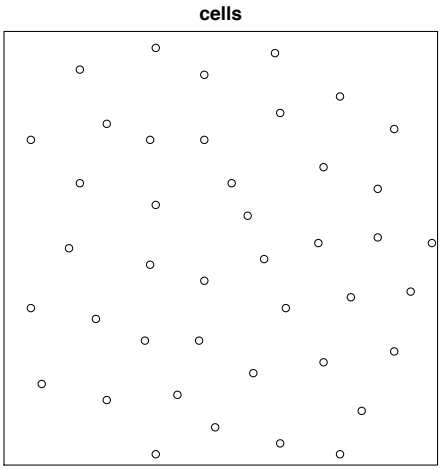
Fonction g et interaction :
Pour un processus de Poisson, on a $g(x,y) = 1$.
Si $g(x,y) > 1$, cela indique que pour ce PP, il est plus probable d'observer un couple de points en x et y que pour un processus de Poisson ayant la même intensité.

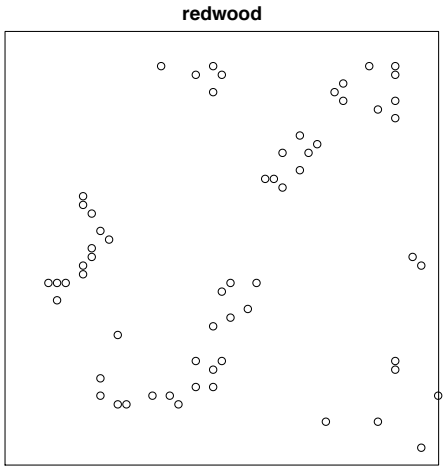
Cadre Stationnaire, isotrope
$g(r) = \rho^{(2)}(r)/\lambda^2$.

- $g(r) > 1$ indique une tendance à l'aggrégation pour des points à distance r,
- inversement, $g(r) < 1$ indique une tendance à la répulsion pour des points à distance r.

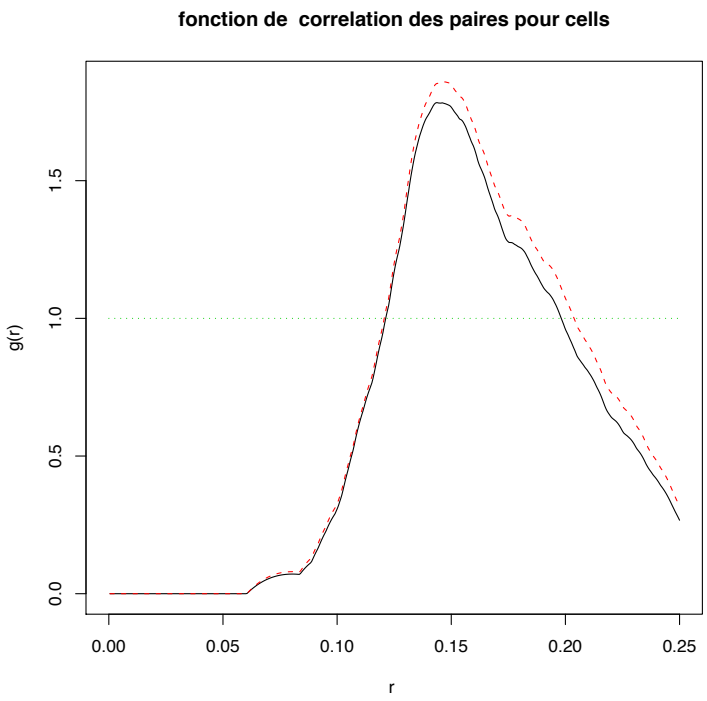
Estimation de g : exemple

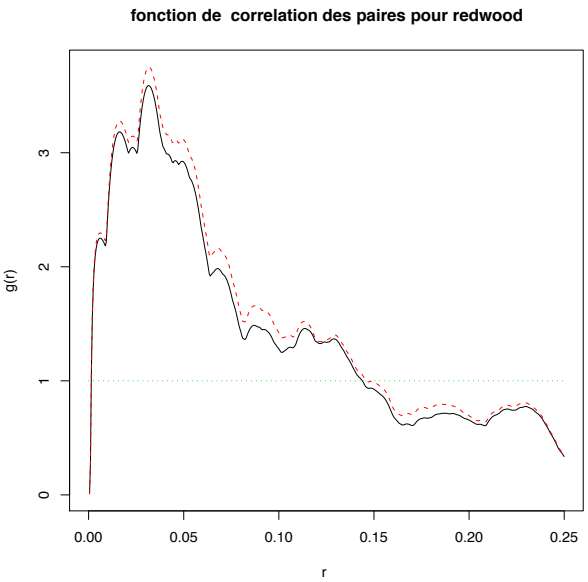
Deux forêts





Estimation de g : exemple





Caractéristiques du 2nd ordre : 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.
Fonction K de Ripley :
$\lambda K(r) = \mathbb{E}[\text{nb voisins à dist. } \leq r] = 2\pi \int_0^r ug(u)du$.
$K(r)$ peut aussi s'interpréter comme le nombre moyen de points du processus dans une boule centrée en un des points du processus, horsmis le centre lui-même.
Pour un processus de Poisson homogène, $K(r) = \pi r^2$.
Version linéarisée : 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 $L(r) = \sqrt{\frac{K(r)}{\pi}} - r$.

Caractéristiques du 2nd ordre : fonction K de Ripley pour cas inhomogène

Baddeley *et al.* (2000) ont étendu la définition de K et g à une sous-classe de processus ponctuels inhomogènes (second-order intensity-reweighted stationary).

Estimations :

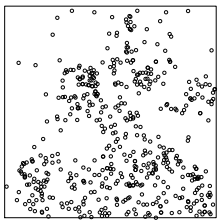
$$\hat{K}(r) = \frac{1}{|W|} \sum_{i=1}^{n} \sum_{j \neq i} \hat{w}_{x_i,x_j,r} \frac{\mathbb{1}(\|x_i - x_j\| \leq r)}{\hat{\lambda}(x_i)\hat{\lambda}(x_j)}, \quad r \geq 0$$

$$\hat{g}(r) = \frac{1}{2\pi r} \sum_{i=1}^{n} \sum_{j \neq i} \hat{w}_{x_i,x_j,r} \frac{h^{-1} K\left(\frac{r - \|x_i - x_j\|}{h}\right)}{\hat{\lambda}(x_i)\hat{\lambda}(x_j)}, \quad r \geq 0$$

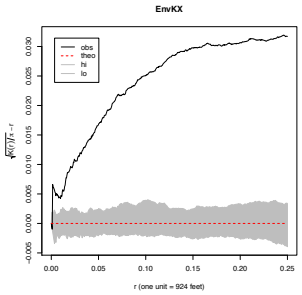
où $\hat{w}_{x_i,x_j,r}$ est une correction d'effet de bord.

Test de l'hypothèse CSR avec la fonction L

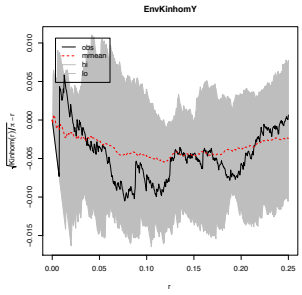
Hypothèse CSR : le processus observé est-il compatible avec un modèle de Poisson homogène ? Il existe de multiples tests (par exemple : quadrats) mais ils nous amènent parfois à rejeter l'hypothèse sans savoir si le rejet vient d'une inhomogenéité pure (sans interaction) ou d'une présence d'interaction. Souvent, nous disposons d'une seule réalisation et seule des connaissances a priori peuvent nous permettre de trancher.
Un test basé sur l'estimation de la fonction L_{inhom} permet de trancher.



(h) Lansing



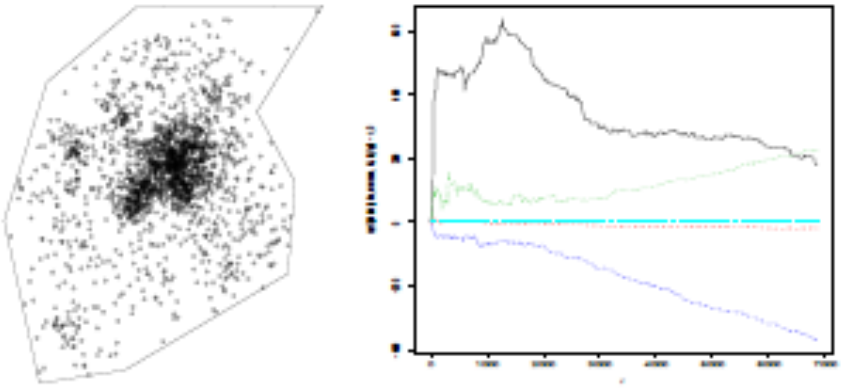
(i) Enveloppe $L(r)$ sous CSR



(j) Enveloppe $L(r)$ sous Poisson inhomogène

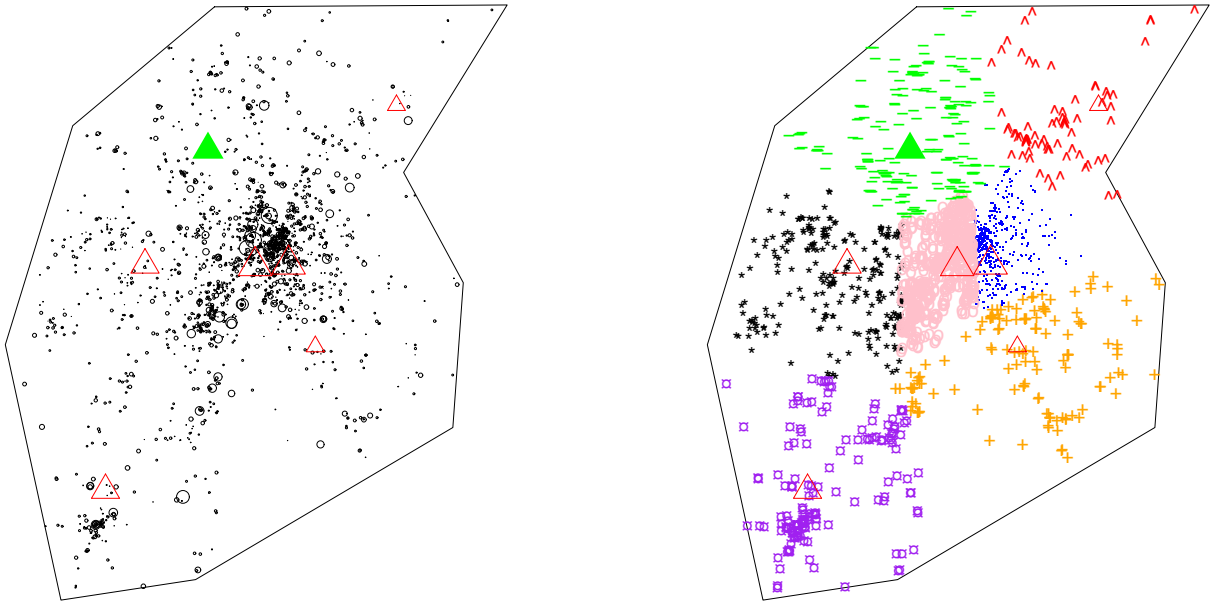
Résultats analyse processus des pompiers

Gauche : simulation d'un Poisson inhomogène avec intensité estimée, droite : enveloppe de la fonction Linhom



Résultats analyse positionnement d'une caserne

Left : Final positions (existing fire station in red triangle, optimal position in green solid triangle), right : final allocations for the genetic method



Exemple de code dans Spatstat

```
load("C:\\CHEMIN\\Pompiers_janvier+region.Rdata")
attach(sinistres_janvier)
PP=ppp(X,Y,window=Region)
s=summary(PP)
str(s)
plot(PP,main="Sinistres Region de Toulouse")
h=5000
Z=density.ppp(PP,h, edge=TRUE)
plot(Z)
ppm(PP, trend, interaction, ...)
ppm(PP, ~ sqrt(x^2 + y^2), Poisson())
```

Cadre général

Variable dépendante : Y (quantitative, univariée) observée sur un nombre fini de zones representées par leur centroides s_i.
Variable indépendente : X (quantitative , multivariaée de dimension p), observée sur les mêmes zones.
En général on suppose que X et Y sont gaussiens.
Modèle de base : $Y = \mu + \epsilon$ avec $\mu = \mathbb{E}(Y \mid X)$ ($\mathbb{E}(\epsilon) = 0$ et $X \perp Y$), $\mathbb{V}ar(Y \mid X) = V$.
Une seule réalisation, i.e. une seule observation du couple (X, Y) pour les n sites.
Sans autre restriction sur le modèle, on a n observations pour estimer $n + \frac{n(n+1)}{2}$ paramètres $\mapsto$ besoin de réduire le nombre de paramètres.

Cadre général

Le terme déterministe $\mathbb{E}(Y_s)$ s'appelle la **tendance** et modélise les variations à grande échelle du phénomène décrit par ce champ.

Le terme aléatoire $(Y_s - \mathbb{E}(Y_s))$ s'appelle la **fluctuation** et modélise les variations du champ à petite échelle. Notons que la fluctuation a une moyenne nulle.

Dans la pratique, cette décomposition en deux termes pour un phénomène observé n'est pas unique et c'est le choix du modélisateur d'affecter certains aspects à la partie aléatoire ou à la partie déterministe.

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.

Modélisation de la tendance

On peut exprimer la variation à grande échelle par une fonction de

- les coordonnées géographiques
- d'autres facteurs explicatifs + leur version spatialement décalée
- une combinaison des deux

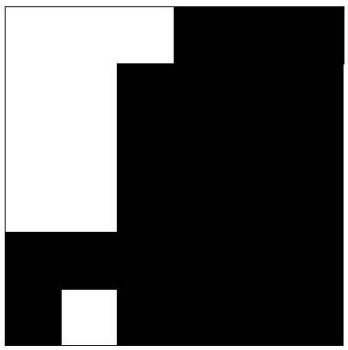
Intuition de l'autocorrélation

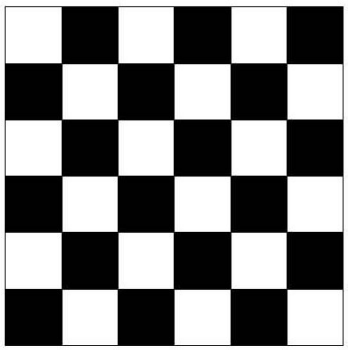
Si la tendance est spécifique au moment d'ordre un d'un champ, l'autocorrélation concerne le moment d'ordre deux que l'on supposera exister.
Pour les données spatiales, une corrélation peut se produire entre X_s et X_t du fait de leur proximité géographique.

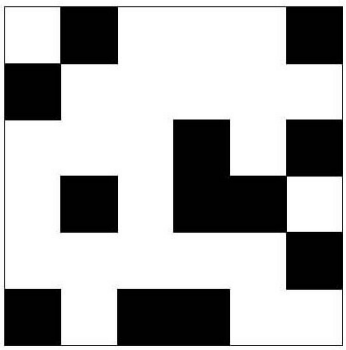
- autocorrélation spatiale **positive** : regroupement géographique de valeurs similaires de la variable.
- autocorrélation spatiale **négative** : regroupement géographique de valeurs dissemblables de la variable.
- **absence** d'autocorrélation : pas de relation entre la proximité géographique et le degré de ressemblance des valeurs de la variable.

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.



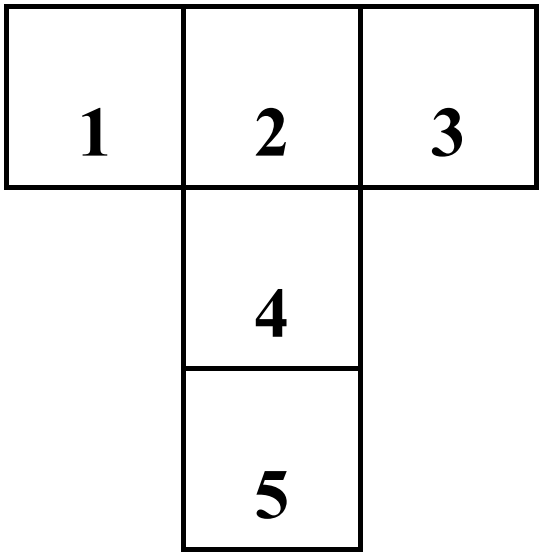




Matrices de poids W

- Version spatiale de l'opérateur retard en séries temporelles.
- Pour n sites géographiques, W est de taille $n \times n$ et w_{ij} mesure intensité dépendance de la zone i par rapport à la zone j.
- Par convention $w_{ii} = 0$.
- W n'est pas nécéssairement symétrique. Si W quelconque, $(W + W')/2$ est symétrique.
- On dit qu'une matrice de poids est normalisée si $\sum_{j=1}^{n} w_{ij} = 1$. Cette contrainte permet de rendre les paramètres spatiaux comparables entre divers modèles. 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.

Petit exemple de matrice de contiguité



Matrice de contiguité

$$W = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

Petit exemple de matrice de contiguité : normalisation

W normalisée

$$W = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1/3 & 0 & 1/3 & 1/3 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 0 & 1/2 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

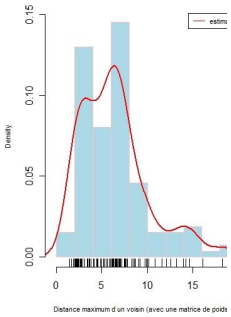
Autres exemples

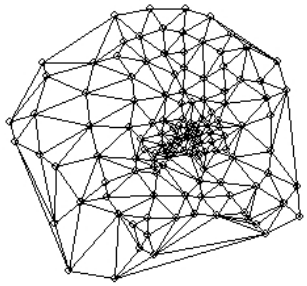
- Matrice basée sur des distances : par exemple $w_{ij} = \mathbf{1}_{(d(s_i,s_j) \leq seuil)}$
- Matrice basée sur les k plus proches voisins : étant donné un entier k, pour un site i, les indices j tels que $w_{ij} = 1$ sont ceux de son plus proche voisin, de son deuxième plus proche voisin, etc... jusqu'à son k-ième plus proche voisin.
- Matrice basée sur la 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 joignant est une arête de la triangulation.

Exemple : Analyse d'une matrice de Delaunay

```
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
```





Variable spatialement décalée

Si $\mathbf{z}$ est une variable et W une matrice de poids, la variable spatialement décalée associée à $\mathbf{z}$ est $W\mathbf{z}$.

Si W est binaire, le terme i de $W\mathbf{z}$ est la somme des valeurs de $\mathbf{z}$ associées aux voisins du site i.

si W est normalisée, le terme i de $W\mathbf{z}$ est la moyenne (pondérée par la proximité) des valeurs de $\mathbf{z}$ sur les voisins du site i.

Indice de Moran

$$I = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(Z_i - \bar{Z})(Z_j - \bar{Z})}{\sum_{i=1}^{n}(Z_i - \bar{Z})^2}$$

Si $\mathbf{z} = Z - \bar{Z}$, les valeurs de $\mathbf{z}$ 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

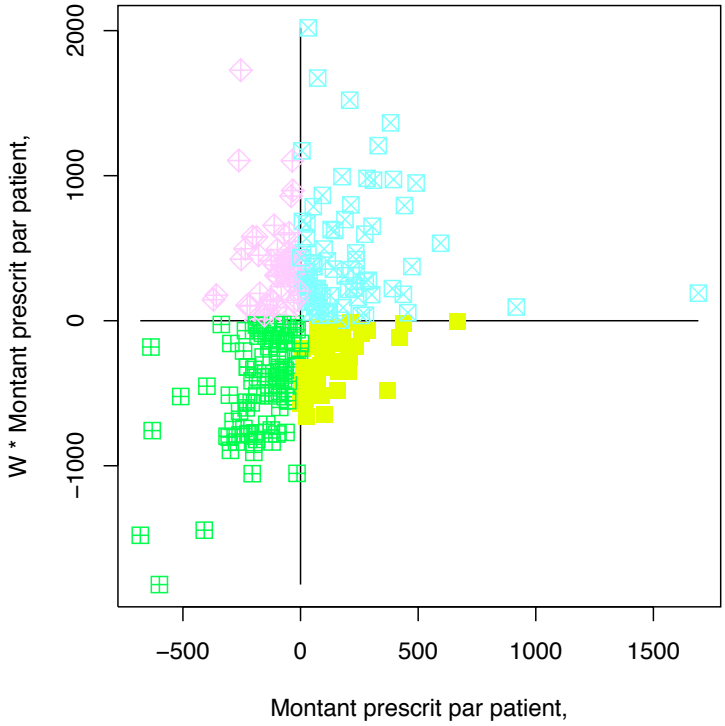
Diagramme de Moran (Anselin, 1993)

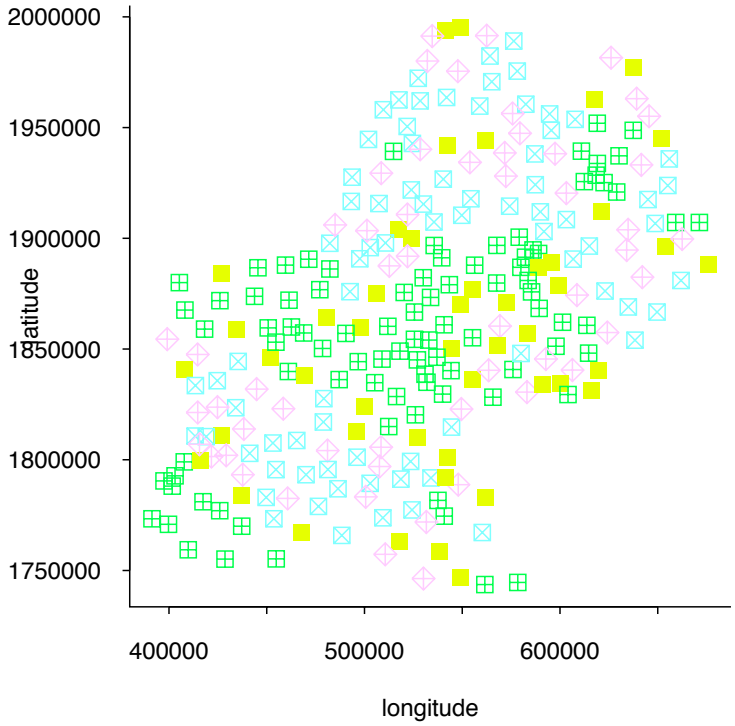
Le **"Moran scatterplot"** est un nuage de points de WX contre X, où X est centrée et W normalisée.
Les deux propriétés X centrée et W normalisée impliquent que la moyenne empirique de WX est égale à $\bar{X}$ et donc à 0.
On peut superposer au nuage la droite de régression qui passe donc par l'origine.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

Diagramme de Moran : exemple





Statistiques "join counts" pour variable dichotomique

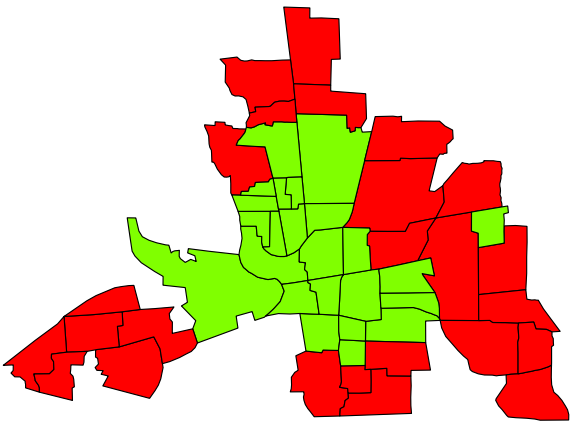
Si X_i a deux modalités 0 et 1 avec : $P(X_i = 1) = p$, on introduit les statistiques suivantes appellées "join counts"

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

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

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
```



Test de Moran pour variable 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 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.

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 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.

Tester l' autocorrelation des résidus d'un modèle WLS

L'indice de Moran généralisé est donné par la même formule que l'ordinaire mais il s'applique aux résidus d'un modèle WLS. Comme ces résidus sont corrélés, cela doit être pris en compte pour le calcul des moments et de la distribution asymptotique. Dans le modèle "free sampling" ($X_1, \cdots, X_n$ i.i.d. $\mathcal{N}(0, \sigma^2)$), avec $D = I_n$, on prouve que sous l'hypothèse d'absence d'autocorrélation,

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

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

Tester l' autocorrelation des résidus d'un modèle WLS

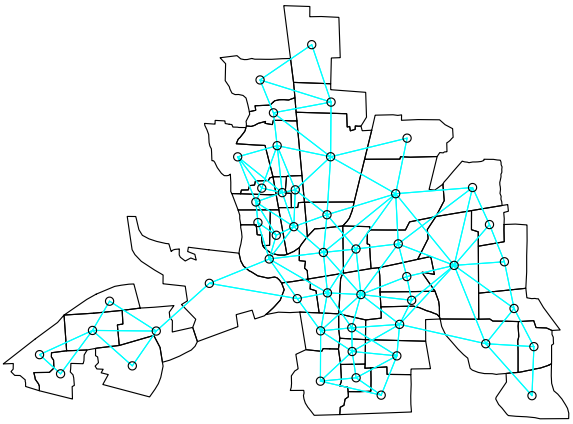
Si $k = 1$ (aucune explicative), on trouve la formule classique $\mathbb{E}(I) = -\frac{1}{n-1}$.
Si $k = 2$ (une seule explicative), on trouve $\mathbb{E}(I) = -\frac{1+I_X}{n-2}$, où I_X est l'indice de Moran pour la variable X.

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

où $B = (X'X)^{-1}X'(W + W')^2X$

Etude de cas : Columbus

On va chercher à expliquer la criminalité dans les quartiers par la valeur immobilière et le revenu des ménages.
La structure de voisinage est une matrice de contiguité de type "queen" notée W



Etude de cas : Columbus

Ajustement d'un modèle OLS

```
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
```

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

```
    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
```

La famille SAR

Etant données une matrice de poids W et une variable Z, la variable spatialement décalée WZ présente automatiquement de l' autocorrélation spatiale.
La famille des modèles simultanés autorégressifs SAR est obtenue en introduisant de l' autocorrélation spatiale par l'usage d'une variable spatialement décalée d'une façon ou d'une autre dans un modèle OLS ou WLS. Une autre famille est celle des modèles conditionnels autorégressifs CAR.

La famille SAR

- introduire WX dans le modèle WLS conduit au modèle SLX : spatially lagged-X model
- introduire WY dans le modèle WLS conduit au modèle LAG : lag model
- introduire WX dans le modèle LAG conduit au modèle SDM : "Spatial Durbin"
- utiliser un modèle LAG pour le terme d'erreur conduit au modèle SEM : "Spatial Error model"

La famille SAR

- combiner les modèles LAG et SEM models conduit au modèle SAC
- introduire $W\epsilon$ dans le modèle WLS model conduit au modèle MA
- combiner les modèles LAG et MA models conduit au modèle SARMA

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. Le paramètre ρ est lié à l'intensité de l'autocorrélation dans Y 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 : tendance et autocorrélation

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.

Matrice de précision $Q = \Sigma^{-1} = \frac{1}{\sigma^2}(I - \rho W')(I - \rho W)\Phi^{-1}$

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
```

Modèle conditionnel autorégressif CAR

Ce modèle est défini par des contraintes de type markovien sur la loi conditionnelle de Y_i sachant Y sur 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 des cobditions $D^{-1}C$ symmétrique et $D^{-1}(I - C)$ définie positive.
- μ est une combinaison linéaire des variables 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 model à un paramètre de variance $C = \rho W$ avec W une matrice de voisinage.
La variance est alors donnée par $V = \tau^2(I_n - \rho W)^{-1}D$.

Comparaison CAR-LAG

Sous l'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 CAR

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

Si l'on pose $C = \rho(W+W') - \rho^2 WW'$ et $\sigma = \tau$, on voit qu'on a la même structure de covariance mais la moyenne est modélisée différemment.
Note : pas de problème d' identification pour le modèle LAG
Différences : les effets de débordement existent dans le modèle LAG mais pas dans CAR, par contre les estimateurs OLS sont convergents pour CAR (pas pour LAG)

Navettes domicile-travail

Approche statistique log-linéaire

$$\log(Y_{od}) = \log(\alpha) + \beta_o \log(Y_o) + \beta_d \log(Y_d) + \gamma \log(D_{od}) + \epsilon$$

Usuellement traité de façon homoscédastique, et en mettant de coté les flux diagonaux et les flux nuls.
Estimation : MCO

$$\min_{\alpha,\beta_o,\beta_d,\gamma} \sum_o \sum_d (\log(Y_{od}) - \log(\alpha) - \beta_o \log(Y_o) - \beta_d \log(Y_d) - \gamma \log(D_{od}))^2$$

Approche statistique Poissonnienne : Le flux Y_{od} suit une loi de Poisson de paramètre λ_{od} avec

$$\lambda_{od} = \exp(\log(\alpha) + \beta_o \log(Y_o) + \beta_d \log(Y_d) + \gamma \log(D_{od}))$$

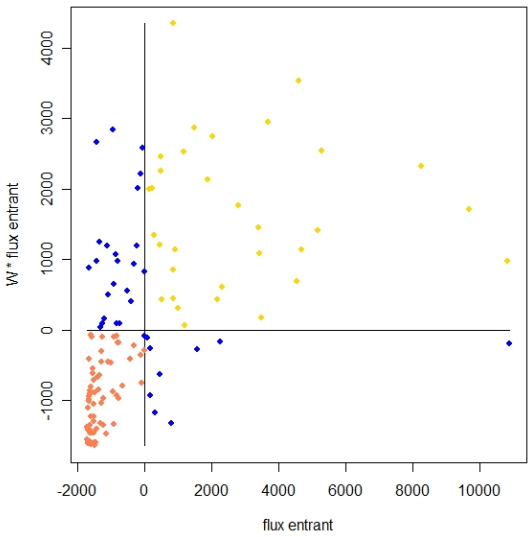
Modèle gravitaire - approche log-linéaire

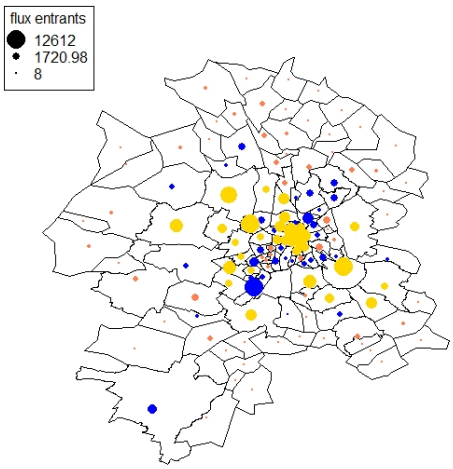
n sites, $N = n^2$ flux (chaque origine est aussi une destination)

$$\log(1 + Y_{od}) = \alpha + X_d\beta_d + X_o\beta_o + \gamma \log(g) + \epsilon,$$

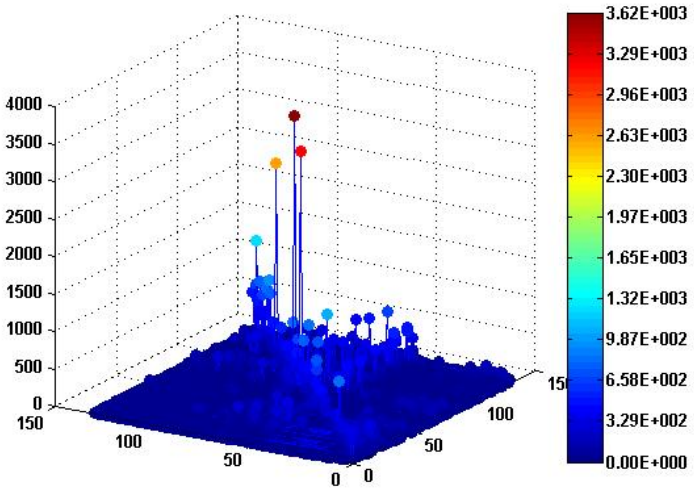
où
X_d est la matrice des caractéristiques des origines,
X_o la matrice des caractéristiques des destinations,
g est la version vectorisée de la matrice des distances et $\epsilon \sim \mathcal{N}(0, \sigma^2 I_N)$.

Autocorrélation dans les flux entrants ?

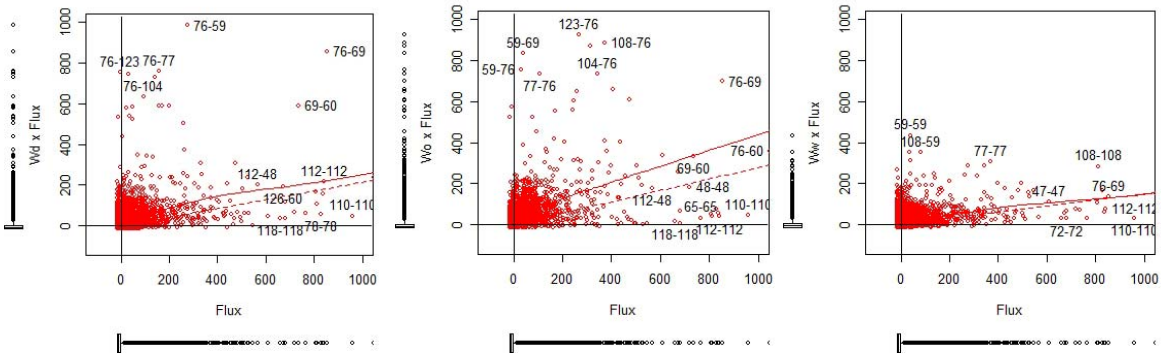




Plot3D des flux



Autocorrélation dans les flux ?



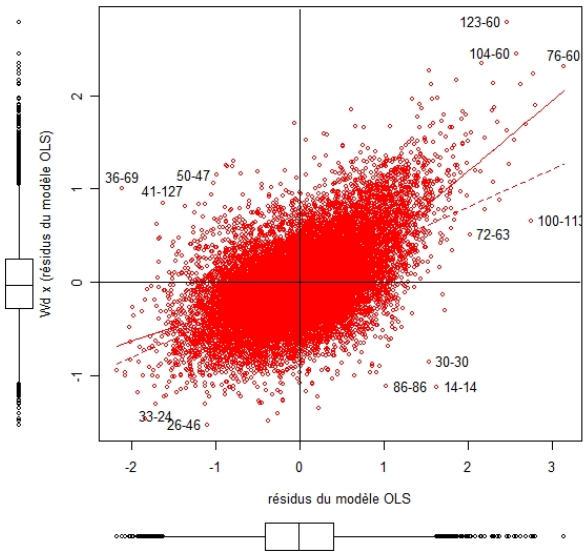
Voisinages pour données bilocalisées

Difficulté : les flux sont bilocalisés. Définir des voisinages ? Trois façons simples à partir d'une matrice W définie sur les sites

1. Wo : sont voisins les flux de même destination dont l'origine est voisine ($Wo = I_n \otimes W$)
2. Wd : sont voisins les flux de même origine dont la destination est voisine ($Wo = W \otimes I_n$)
3. Ww : sont voisins les flux dont soit origine et destination sont voisines ($Wo = W \otimes W$)

S'écrivent simplement avec des produits tensoriels de W et I

Diagramme de Moran des résidus



Adaptation du modèle LAG aux flux

D'après un article de LeSage et Pace (2007)

$$\log(1+Y_{od}) = \rho_d W_d Y + \rho_o W_o Y + \rho_w W_w Y + \alpha + X_d\beta_d + X_o\beta_o + \gamma \log(g) + \epsilon,$$

Si $\rho_d = \rho_o = \rho_w = 0$, on retrouve le modèle précédent.

Résultats

ols model

Ordinary Least-squares Estimates
Dependent variable = y
R-squared = 0.4424
Rbar-squared = 0.4423
sigma^2 = 0.5712
Durbin-watson = 1.6182
Nobs, Nvars = 17161, 5

Variable	Coefficient	t-statistic	t-probability
constant	-3.955709	-55.286630	0.000000
D_working_pop	0.311319	46.050898	0.000000
O_working_pop	0.344348	50.180599	0.000000
OD_employment_over_pop	-0.002570	-7.682573	0.000000
distance	-0.588300	-83.149924	0.000000

Wo only model

Spatial autoregressive Model Estimates
Dependent Variable = y
R-squared = 0.3556
Rbar-squared = 0.3554
sigma^2 = 0.3073
Nobs, Nvars = 17161, 5
log-likelihood = -9143.3339
# of iterations = 16
min and max rho = -1.0000, 1.0000
total time in secs = 1.8690
time for lndet = 0.7430
time for t-stats = 0.3290
Pace and Barry, 1999 MC lndet approximation used
order for MC appr = 50
iter for MC appr = 30

Variable	Coefficient	Asymptot t-stat	z-probability
constant	-2.798807	-46.915125	0.000000
D_working_pop	0.098795	16.125155	0.000000
O_working_pop	0.321554	63.334826	0.000000
OD_employment_over_pop	-0.000434	-1.756088	0.079073
distance	-0.253517	-39.245762	0.000000
rho	0.655960	87.995982	0.000000

Wd only model

Spatial autoregressive Model Estimates
Dependent Variable = y
R-squared = 0.4406
Rbar-squared = 0.4405
sigma^2 = 0.4802
Nobs, Nvars = 17161, 5
log-likelihood = -12459.013
# of iterations = 14
min and max rho = -1.0000, 1.0000
total time in secs = 1.1430
time for lndet = 0.8150
time for t-stats = 0.1380
Pace and Barry, 1999 MC lndet approximation used
order for MC appr = 50
iter for MC appr = 30

Variable	Coefficient	Asymptot t-stat	z-probability
constant	-2.623126	-36.108313	0.000000
D_working_pop	0.240034	37.412063	0.000000
O_working_pop	0.181859	25.494931	0.000000
OD_employment_over_pop	-0.002145	-6.977302	0.000000
distance	-0.375956	-49.470637	0.000000
rho	0.442967	53.905576	0.000000