Introduction à l'analyse spatiale

Marie-Pierre de Bellefon

Division Méthodes et Référentiels Géographiques

Séminaire de Méthodologie Statistique - 5 juillet 2016

Insee
Mesurer pour comprendre

1. Qu'est-ce qu'une donnée spatiale ?
2. Définir le voisinage d'un objet spatial
3. Mesurer la dépendance spatiale globale
4. Mesurer l'association spatiale locale

Sommaire

Donnée spatiale :

- Toute observation dont on connait non seulement la valeur, mais aussi la localisation.
  $\Rightarrow$ **Le support de l'observation contient des informations**

Analyse spatiale :

- identifie la **structure** spatiale des observations.
- quantifie la force des **interactions** spatiales.
- explique les **mécanismes** sous-jacents.

Trois types de données spatiales :

- Continues
- Ponctuelles
- Surfaciques

Références :

- N.L. Cressie *Statistics for spatial data* (1993)
- M. Hannoun *Un survol des méthodes élémentaires en statistique spatiale* (2000)
- M. Fischer, A. Getis *Handbook of applied Spatial Analysis* (2010)
- J.M. Floch, *Détection des disparités socio-économiques : l'apport de la statistique spatiale* (2012)
- L. Anselin *Spatial Econometrics : Methods and Models* (2013)
- A. Baddeley, E. Rubak, R. Turner *Spatial Point Patterns* (2015)

Données spatiales continues (géostatistiques)

Définition :

Données qui **varient de façon continue dans l'espace**, mais qui sont **mesurées uniquement en un nombre discret de points**.

Objectifs de l'analyse spatiale des données continues :

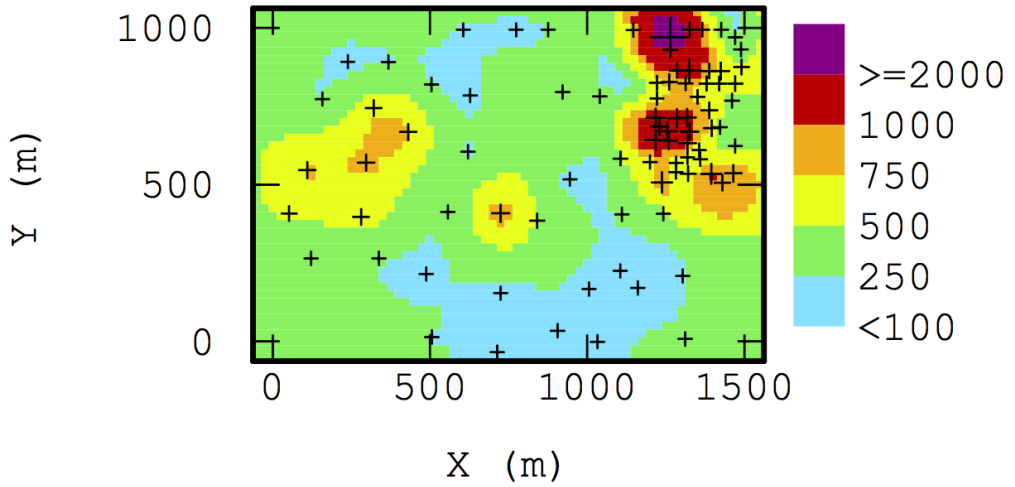
Prédire la valeur d'une variable en un point où elle n'a pas été échantillonnée, ainsi que la fiabilité de cette prédiction.

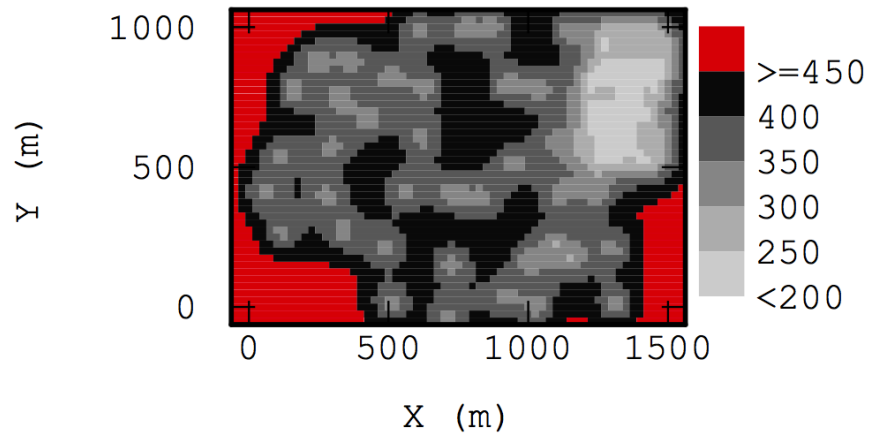
Exemples de données :

- Composition chimique du sol $\Rightarrow$ industrie minière
- Qualité de l'eau ou de l'air $\Rightarrow$ étude de la pollution
- Variables météorologiques $\Rightarrow$ prévisions météo

Données spatiales continues : exemple d'étude

Avec quelle précision prédire la teneur en polluant d'un sol lorsqu'on n'a qu'un échantillon fini de points de mesure ?





Prédiction de teneur en polluant (mg/kg/m^2) - Ecart type de la prédiction
Source : Manuel GéoSiPol - Mines de Paris

Données spatiales ponctuelles

Définition :

Données dont **la localisation est la variable aléatoire** qu'on étudie.

Objectifs de l'analyse spatiale des données ponctuelles :

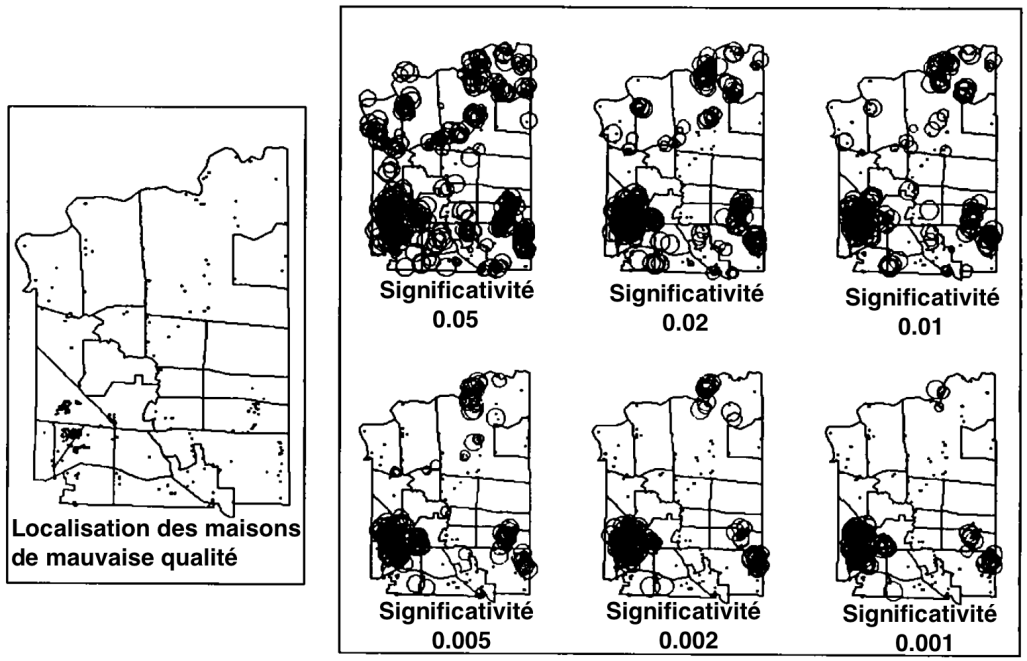
Quantifier l'écart entre la distribution spatiale des observations et une **distribution complètement aléatoire dans l'espace** : dans quelle mesure les observations sont-elles regroupées dans l'espace ?

Exemples de données :

- Ecologie $\Rightarrow$ distribution spatiale de deux espèces d'arbres.
- Epidémiologie $\Rightarrow$ distribution spatiale des cas d'apparition d'un virus.

Données spatiales ponctuelles : exemple d'étude

Détection des clusters et évaluation de leur significativité.
Les maisons inconfortables sont-elles réparties aléatoirement sur le territoire ?



Source : Fotheringham - Zhan, 1996

Données spatiales surfaciques

Définition :

La localisation des observations est considérée comme fixe, c'est **leur valeur qui est modélisée selon un processus aléatoire**.
Peuvent être agrégées sur une partition du territoire ou réparties en des points précis.

Objectifs de l'analyse spatiale des données surfaciques :

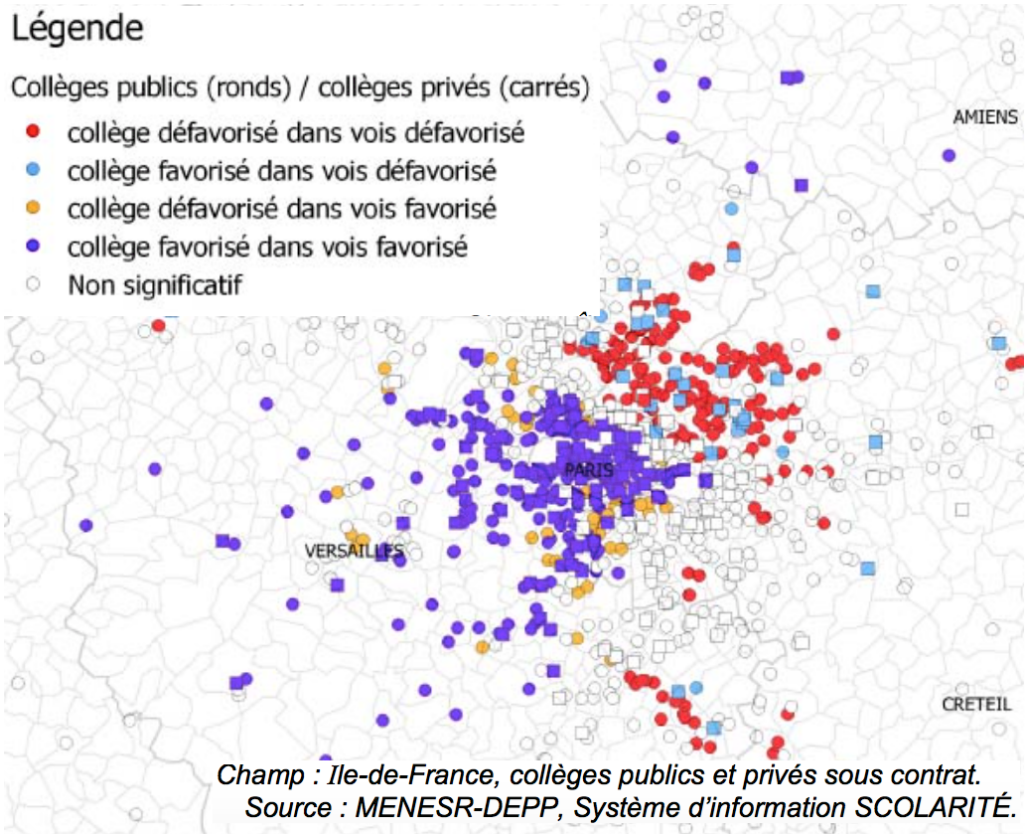
- Définir la structure de voisinage des observations.
- Quantifier l'influence qu'exercent les observations sur leurs voisines.

Exemples de données :

- Santé $\Rightarrow$ taux de malades par région.
- Economie $\Rightarrow$ croissance du PIB par région.

Données spatiales surfaciques : exemple d'étude

Les collèges favorisés sont-ils toujours situés dans un environnement favorisé ?



Source : Givord et al 2015

Sommaire

Données spatiales surfaciques

Dans la suite de cette présentation, on étudie les données surfaciques

Représentation sous forme d'un polygone

- La région d'étude peut être recouverte exhaustivement par des polygones mutuellement exclusifs.
- Deux polygones adjacents sont séparés par une frontière commune.
- Les frontières peuvent naître de discontinuités spatiales : limites administratives, barrières naturelles...
- Ou être les polygones de Voronoï issus de points particuliers

Polygone de Voronoï associé au point x_i :

La région de l'espace qui est plus proche de x_i que de tout autre point de l'ensemble d'étude $\mathbf{x}$

Données spatiales surfaciques

Représentation sous forme d'un point

- Les données "surfaciques" peuvent aussi être des **points fixes** du territoire :
- définis par les coordonnées géographiques d'un point particulier : lycée, mairie, point culminant d'une région,...
- ou définis géométriquement comme les centroïdes d'un polygone.

Centroïde c d'une surface R

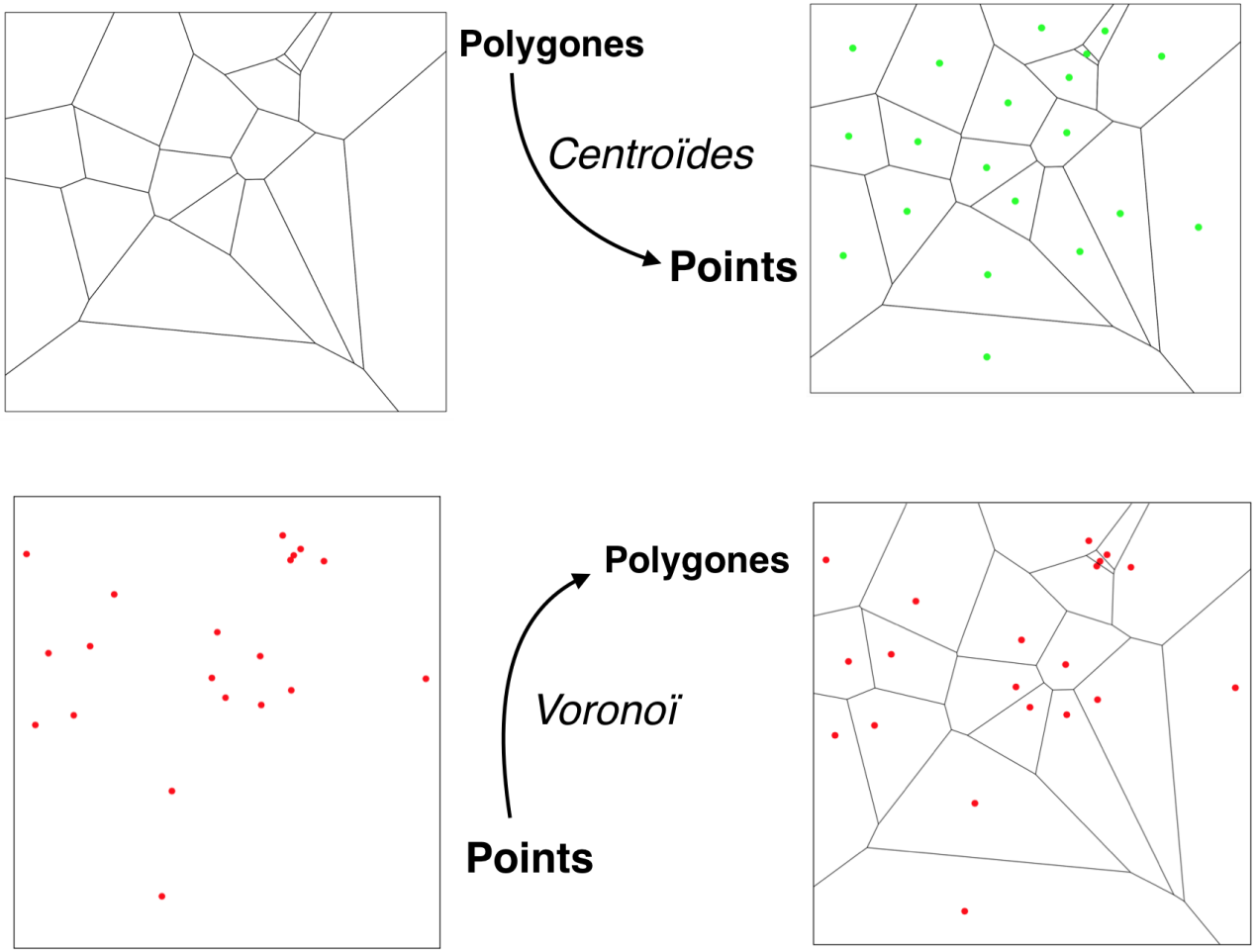
Minimise la distance quadratique moyenne à tous les points de R :

$$\min_c \frac{1}{a(R)} \int_R ||x - c||^2 dx$$

$$c = \frac{1}{a(R)} \int_R x dx$$

Coordonnées de c : moyenne des coordonnées de **tous les points de R**

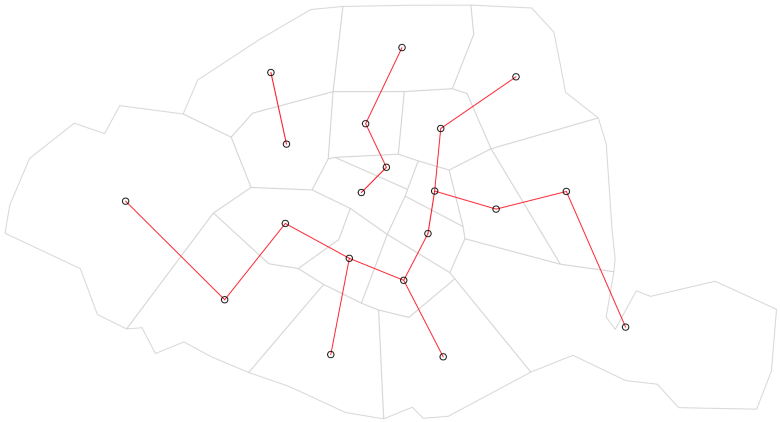
Données spatiales surfaciques



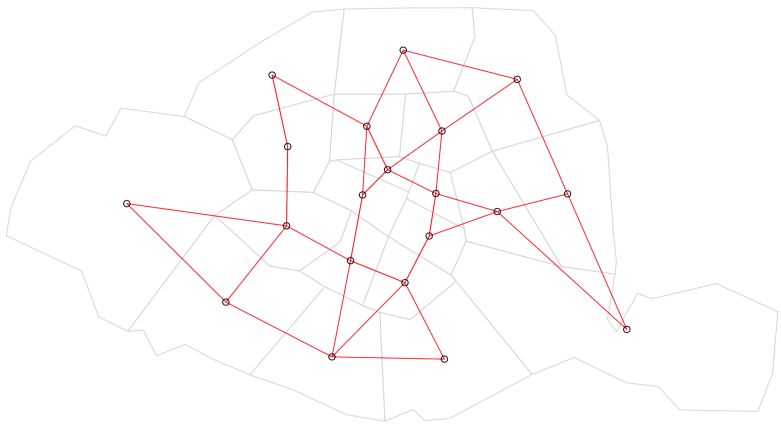
Relations entre objets spatiaux

- Multidirectionnelles et multilatérales
- Différent des relations temporelles où les liens sont orientés sur l'axe passé -> présent -> futur
- Spécifiées par un graphe de voisinage
- Voisinage défini en fonction de la **distance** (plus proches voisins), de la **contiguïté** (Queen, Rook) ou de notions **géométriques** (Delaunay, Gabriel)
- Codé dans une matrice de contiguïté

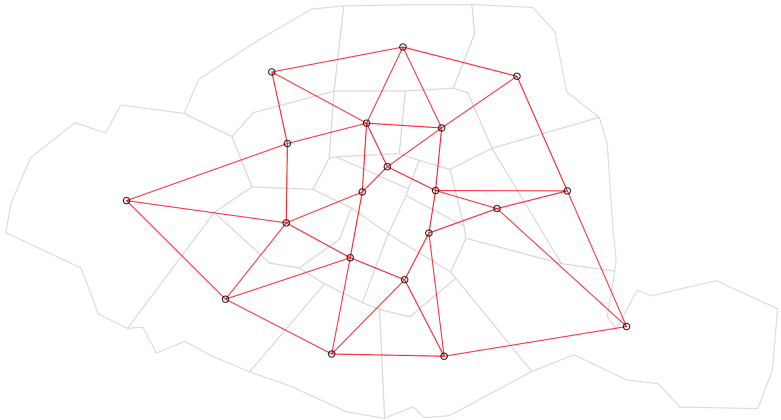
Distance : k plus proches voisins



Plus proches voisins



Deux plus proches voisins



Trois plus proches voisins

Remarques sur le voisinage basé sur la distance

k plus proches voisins

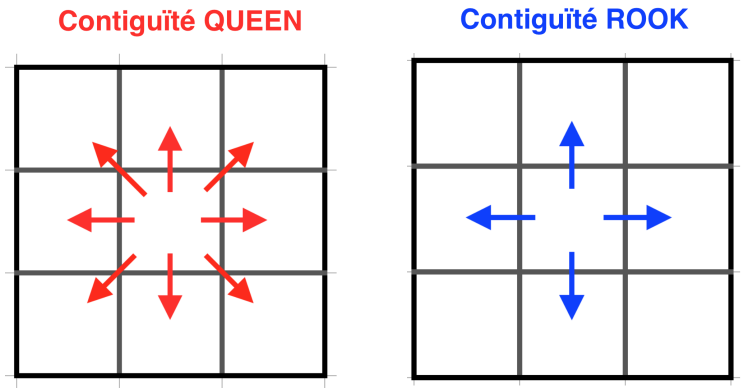
- Fait l'hypothèse que le degré d'influence est le même pour les *k* voisins les plus proches
- Ne prend pas en compte la complexité de la géographie des frontières
- Peut se situer à l'extérieur d'un polygone
- Pas symétrique

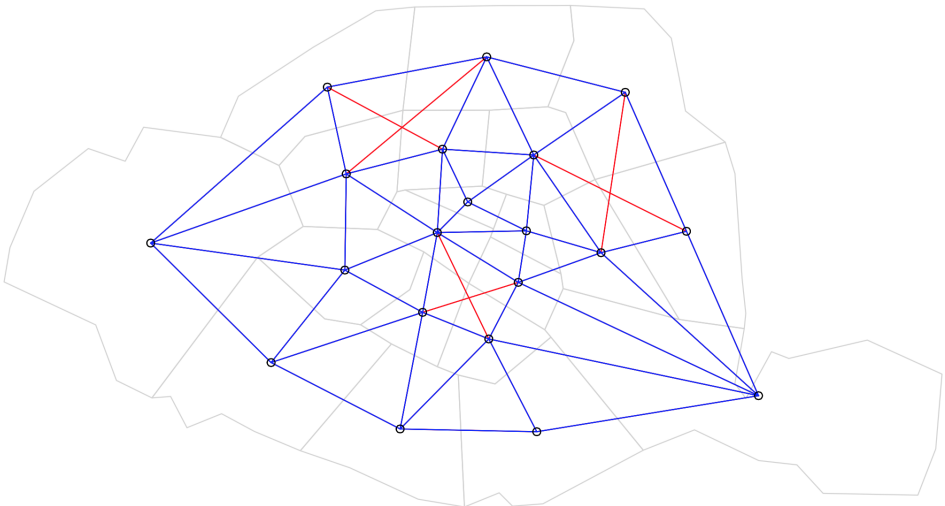
Voisins à distance radiale

Les voisins de i sont les points j tels que $0 \leq d_{ij} \leq d$

Voisinage basé sur la contiguïté

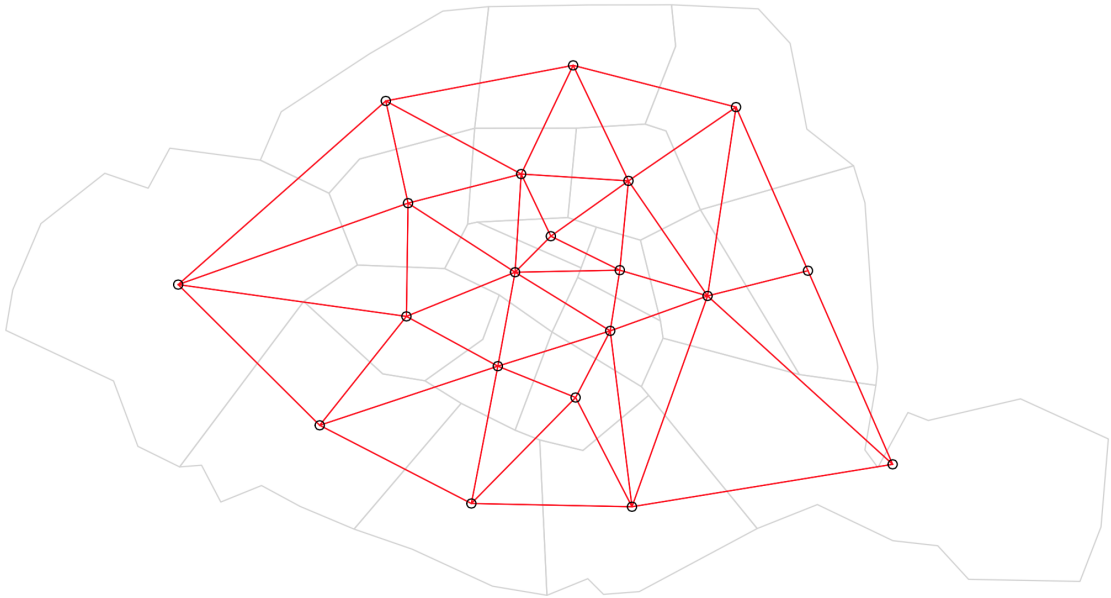
Plus adapté aux données démographiques et sociales qu'aux données environnementales.





Voisinage basé sur des notions géométriques

- Triangulation de Delaunay : maximise l'angle minimal des triangles dont les sommets sont les points d'observation.

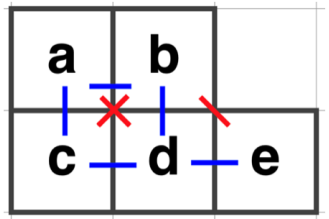


La matrice de poids : W

"L'expression formelle de la dépendance spatiale entre observations" (Anselin, 1988)

Matrice binaire : w_{ij}=1 si i voisin de j, 0 sinon.

- Inconvénient : la somme des poids des voisins d'une zone dépend du nombre de ses voisins.



Contiguïté ROOK

	a	b	c	d	e	Somme des poids des voisins
a	0	1	1	0	0	2
b	1	0	0	1	0	2
c	1	0	0	1	0	2
d	0	1	1	0	1	3
e	0	0	0	1	0	1

Contiguïté QUEEN

	a	b	c	d	e	Somme des poids des voisins
a	0	1	1	1	0	3
b	1	0	1	1	1	4
c	1	1	0	1	0	3
d	1	1	1	0	1	4
e	0	1	0	1	0	2

Normalisation de la matrice de poids

Normalisation par ligne : la plus courante

- Pour une zone, le poids accordé à chaque voisin est divisé par le nombre total de ses voisins.
- Avantage : $\sum_j w_{ij}x_j$: moyenne de x sur tous les voisins de x_i
- Inconvénient : trop de poids accordé aux observations avec peu de voisins (en bordure de la région étudiée).
- Alternatives : normalisation globale (poids divisés par nombre total de zones),...

Réflexions sur le choix de la matrice de poids

- "Le choix des poids est souvent arbitraire [...] et le résultat des études varie considérablement en fonction de la définition des poids spatiaux" Bhattacharjee et Jensen-Butler (2006)
- "Le choix de la matrice de poids est le plus grand mythe de l'économétrie spatiale" : l'impact de ce choix sur les résultats ne serait pas si crucial. Lesage et Pace (2012)
- Pour pallier l'exogénéité du choix de la matrice de poids, certains utilisent les données spatiales elles mêmes pour la spécifier.

Sommaire

Structure spatiale

- Soit y_i la variable d'intérêt et v_{ij} le lien entre i et j.
  Un processus spatial est autocorrélé si
  $\mathbb{P}(Y_i \leq y_i | Y_j \leq y_j, \forall j \quad tq \quad v_{ij} > 0) \neq \mathbb{P}(Y_i \leq y_i)$
- Structure spatiale et autocorrélation spatiale ne peuvent pas exister indépendamment l'une de l'autre (Tiefelsdorf, 1998).
  - La structure spatiale est l'ensemble des liens grâce auxquels le phénomène autocorrélé va se diffuser.
  - Sans la présence d'un processus autocorrélé significatif, la structure spatiale n'est pas visible empiriquement.
- La distribution spatiale observée est une manifestation du processus spatial sous-jacent.

Autocorrélation spatiale : observation empirique

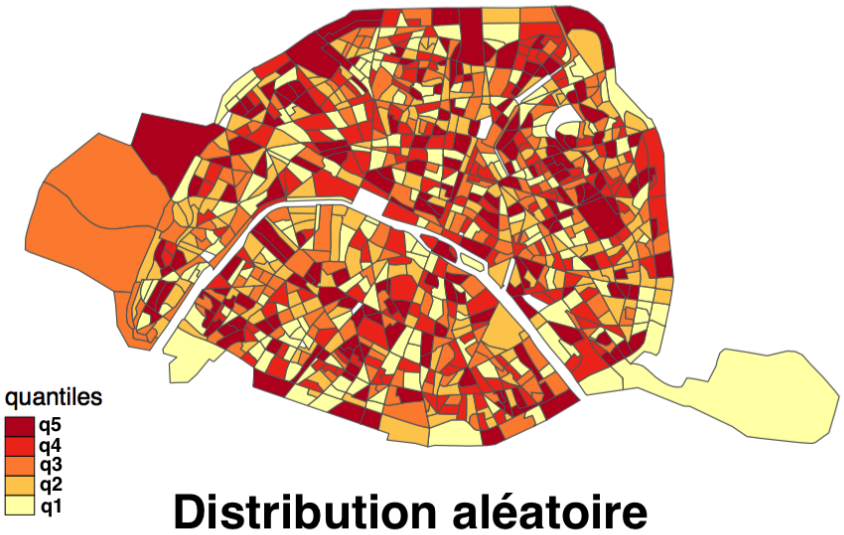
- La valeur d'une observation est liée aux valeurs des observations voisines.
- L'autocorrélation spatiale est positive lorsque des valeurs similaires de la variable à étudier se regroupent géographiquement.
- Le niveau d'autocorrélation spatiale mesure la force du lien entre deux entités.

Indices d'autocorrélation spatiale

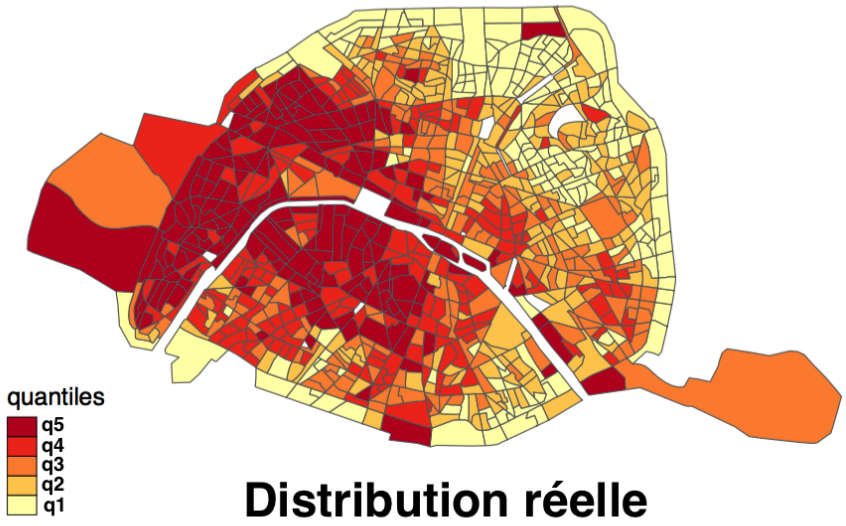
- Testent la significativité de la structure spatiale identifiée.
- Attention : si les données sont agrégées suivant un découpage qui ne respecte pas le phénomène sous-jacent, on surestimera la force du lien spatial.

Autocorrélation spatiale : observation empirique

Distribution spatiale du revenu médian par IRIS *(source : RFL, INSEE)*



Distribution aléatoire



Distribution réelle

Le regroupement spatial des valeurs similaires est-il significatif ?
Dans quelle mesure aurait-il pu être observé si les observations étaient réparties aléatoirement ?

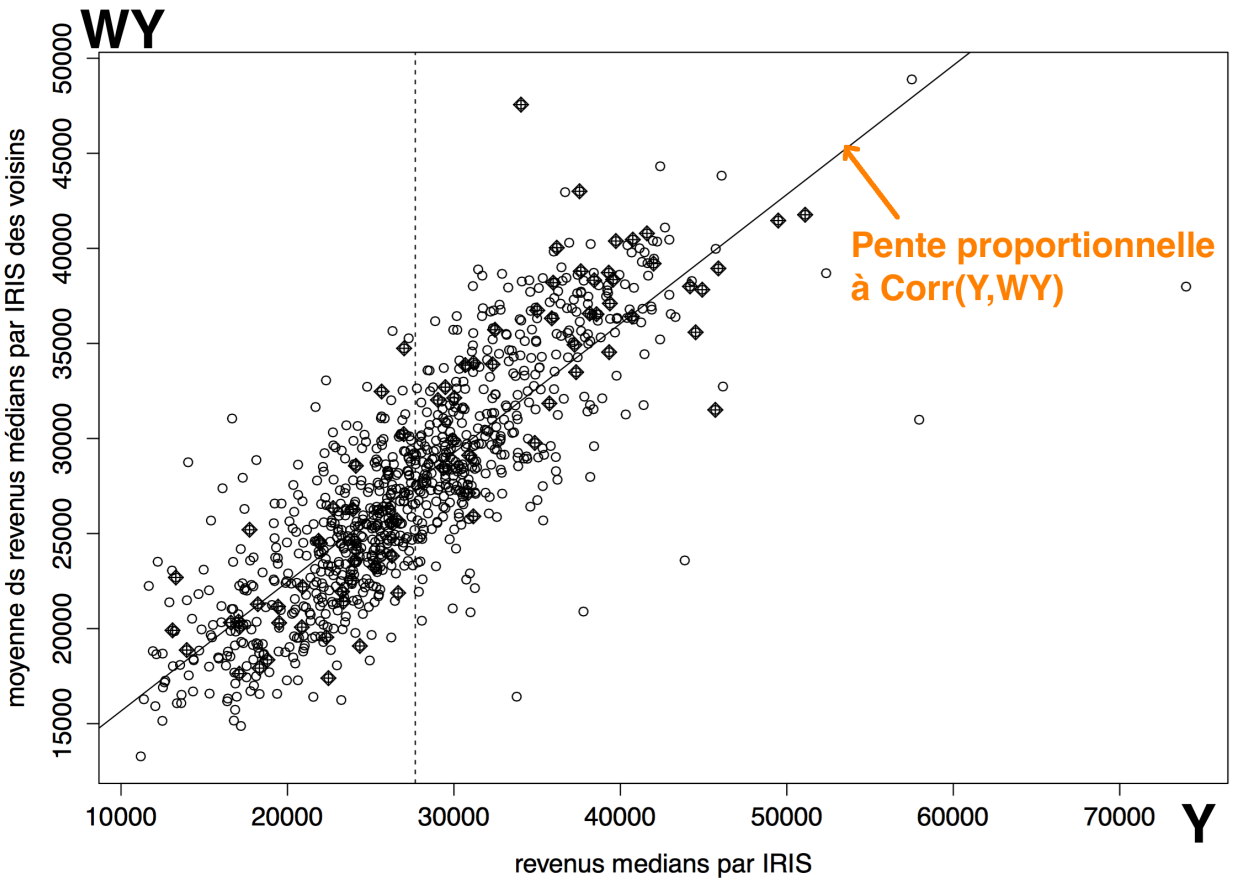
Indices de Moran et de Geary

Principe

- Mesure la corrélation entre une variable et sa valeur sur les entités voisines.
- $Corr(Y, WY) = \frac{Cov(Y, WY)}{\sqrt{Var(Y), Var(WY)}}$ W : matrice de pondération
- Hypothèses sur sa distribution, qui permettent de tester sa significativité.

Indices de Moran et de Geary

Corrélation entre les revenus médians par IRIS et les revenus médians des IRIS voisins. (Source : INSEE, RFL 2010)



Indice de Moran

$$I_W = \frac{n}{\sum_{i \neq j} w_{ij}} \frac{\sum_i \sum_j w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_i (y_i - \bar{y})^2} \quad i \neq j$$

H_0 : Les voisins ne **co-varient** pas de la même façon.
$I_W > 0$ => autocorrélation spatiale positive

Indice de Geary

$$c_W = \frac{n-1}{2W} \frac{\sum_i \sum_j w_{ij}(y_i - y_j)^2}{\sum_i (y_i - \bar{y})^2} \quad i \neq j$$

H_0 : Les **différences** entre voisins n'ont pas de structure particulière.
$c_W < 1$ => autocorrélation spatiale positive.

Propriétés de la distribution du I de Moran

Hypothèses sur la distribution de la variable d'intérêt y en l'absence d'autocorrélation spatiale :

Hypothèse normale

Chaque valeur y est une réalisation choisie alétoirement dans la **distribution normale propre à chaque zone géographique**.

Hypothèse uniforme

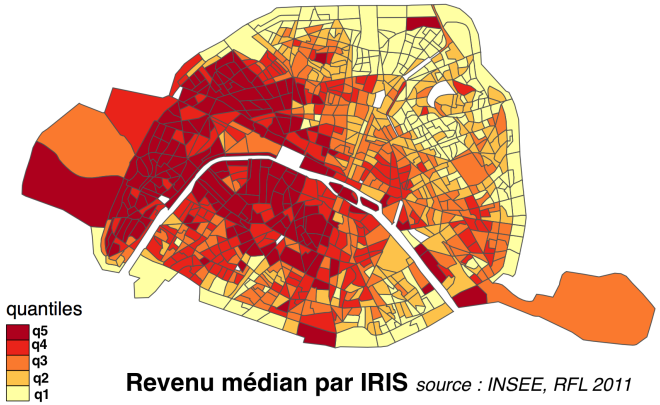
Les valeurs y sont différentes réalisations d'une **distribution uniforme identique sur toutes les zones**.

Distribution du I de Moran

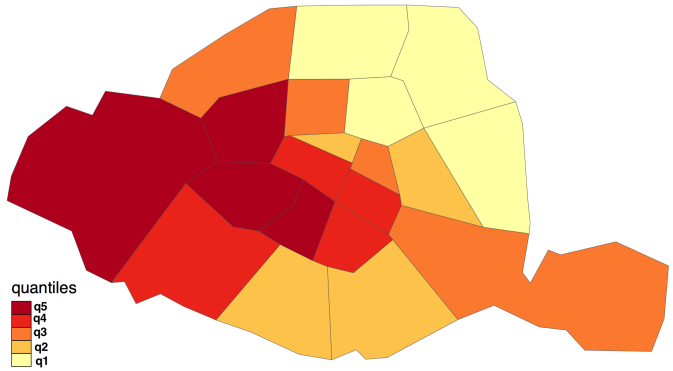
- *min(valeurs propres* $(\frac{W+W^T}{2})) < I_W <$*max(v. p.*$(\frac{W+W^T}{2}))$
- $E(I_W)$ ne dépend pas des hypothèses sur la distribution de y
- $E(I_W) = -\frac{1}{n-1}$
- $Var(I)$ dépend des hypothèses sur la distribution de y
- $\frac{I - E(I)}{\sqrt{Var(I)}} \sim \mathcal{N}(0,1)$ => calcul des p-values
- La vitesse de convergence vers la distribution normale dépend de la disposition spatiale des zones d'étude
  => Simulation Monte Carlo pour estimer la distribution.

Autocorrélation spatiale des revenus médians à Paris

- Quelle est l'intensité de l'autocorrélation spatiale des revenus parisiens ?
- Est-elle significative ?
- Dans quelle mesure dépend-elle de la spécification des relations spatiales (échelle d'agrégation, voisinage) ?



Revenu médian par IRIS *source : INSEE, RFL 2011*



Revenu médian par arrondissement *Source : INSEE, RFL 2010*

Autocorrélation spatiale des revenus médians à Paris

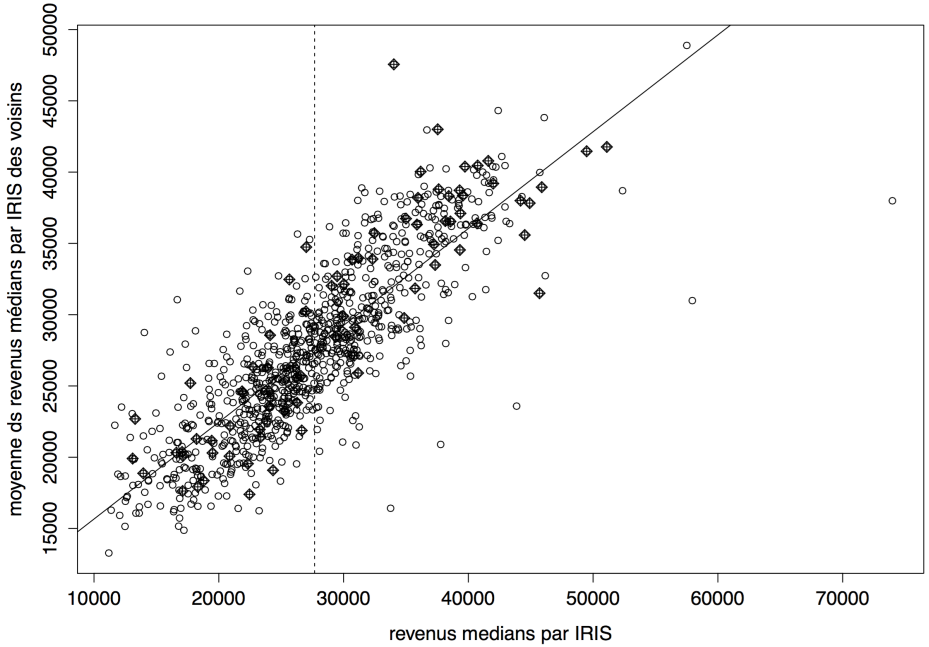
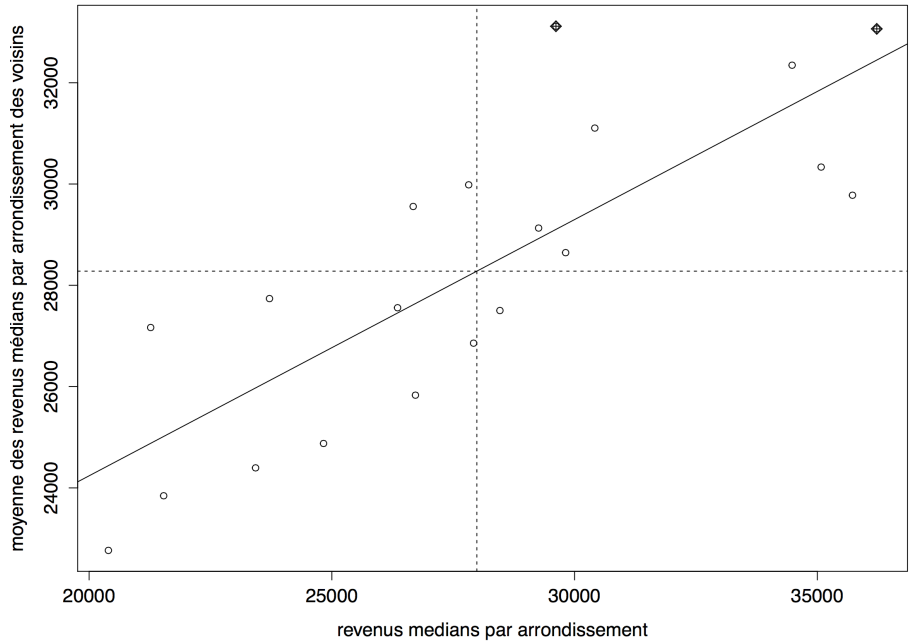


Diagramme de Moran : IRIS



Arrondissement

Echelle d'agrégation	I	p value	H0	bornes de I
IRIS	0.68	$< 2.10^{-16}$	rejetée	[-1.06,1.06]
Arrondissement	0.51	$< 9.10^{-9}$	rejetée	[-0.53,1.01]

Autocorrélation spatiale des revenus médians à Paris

Type de voisinage	I	p value	H0
QUEEN	0.51	3.10^{-6}	rejetée
ROOK	0.57	2.10^{-6}	rejetée
1NN	0.30	0.07	rejetée
3NN	0.58	9.10^{-6}	rejetée
Delauney	0.57	6.10^{-7}	rejetée

- Quelle que soit l'échelle d'agrégation ou la définition du voisinage, l'autocorrélation spatiale des revenus parisiens est positive et significative.
- La force de l'autocorrélation spatiale varie légèrement en fonction du cadre de mesure choisi.

Sommaire

Pourquoi un indicateur local ?

Limites d'un indicateur global

- Statistiques globales font l'**hypothèse de stationnarité du processus spatial** : l'autocorrélation spatiale serait la même dans tout l'espace.
- Hypothèse souvent non réaliste, d'autant moins que le nombre d'observations est élevé.

Doubles objectifs d'un indicateur local

- Détecter les regroupements significatifs de valeurs identiques autour d'une localisation particulière (clusters).
- Repérer les zones de non-stationnarité spatiale, qui ne suivent pas le processus global.

LISA : Local Indicator of Spatial Association

Définition (Anselin, 1995)

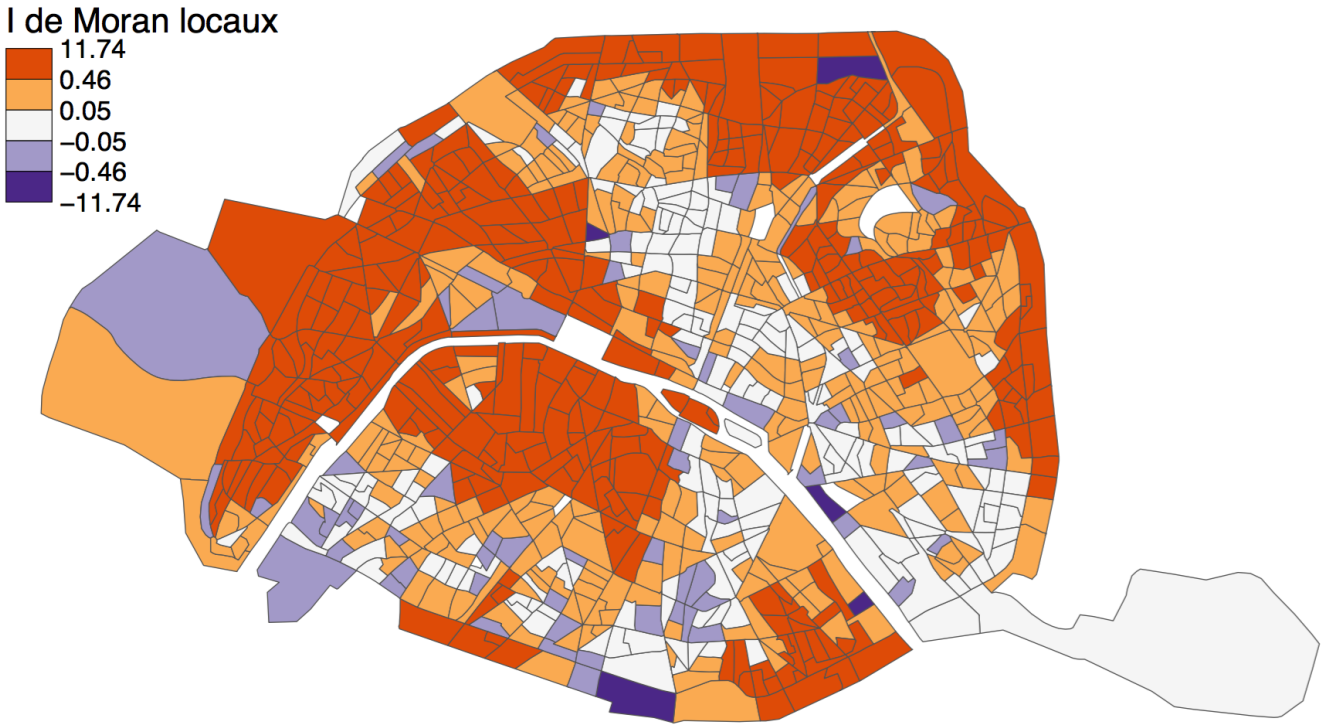
- Le LISA de chaque observation indique l'intensité du regroupement spatial de valeurs similaires autour de cette observation.
- La somme de tous les LISA est proportionnelle à un indicateur global d'association spatiale.

I de Moran local

- $I_i = (y_i - \bar{y}) \sum_j w_{ij}(y_j - \bar{y})$
- $I = constante * \sum_i I_i$

Illustration : I de Moran locaux - revenus à Paris

- $I_i > 0$: regroupement de valeurs similaires (plus élevées ou plus faibles que la moyenne)
- $I_i < 0$: regroupement de valeurs dissimilaires (par ex : valeurs élevées entourées de valeurs faibles)

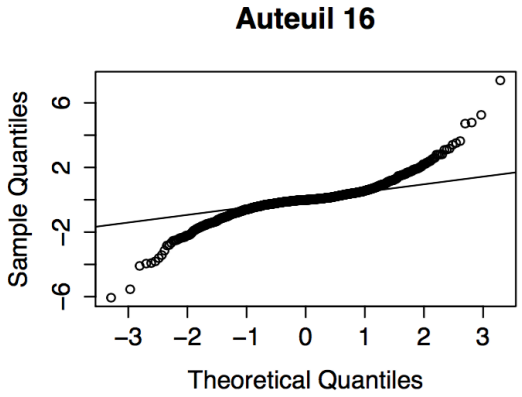


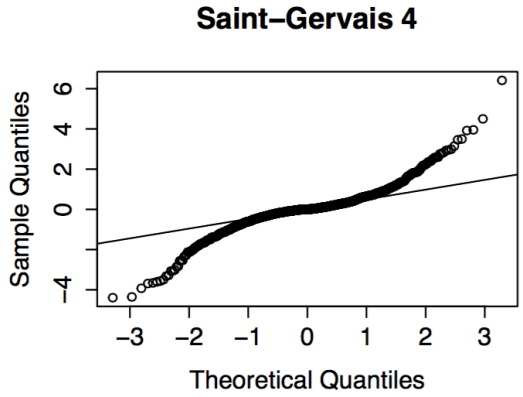
I de Moran locaux des revenus médians parisiens, *source : INSEE, RFL 2010*

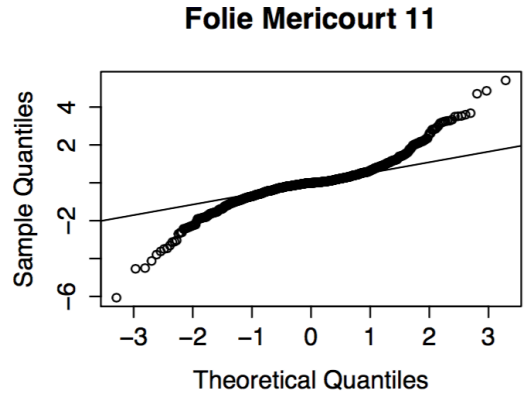
Test de l'hypothèse de normalité

Test de l'hypothèse de normalité de la distribution des I_i

- Méthode : répartition aléatoire des revenus, puis calcul des I locaux.
- Résultats : les quantiles extrêmes de la distribution des I locaux sont plus élevés que ceux d'une distribution normale.
- Conséquence : attention à l'interprétation des p-values !

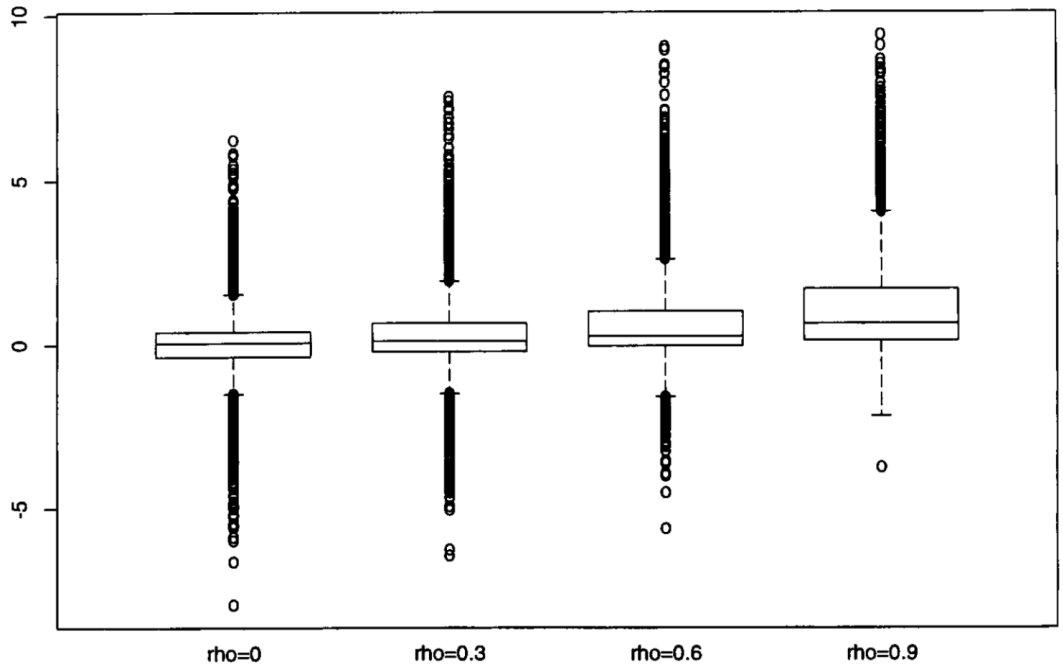






Normalité en présence d'autocorrélation globale

- Si l'hypothèse normale est vérifiée, $z(I_i) = \dfrac{I - E(I_i)}{\sqrt{Var(I_i)}} \sim \mathcal{N}(0,1)$
- **En présence d'autocorrélation spatiale globale, l'hypothèse de normalité des I_i n'est plus vérifiée.**



Box-Plot des z(Ii)
en présence d'autocorrélation spatiale.
n= 24, 10000 simulations, source : Anselin (1995)

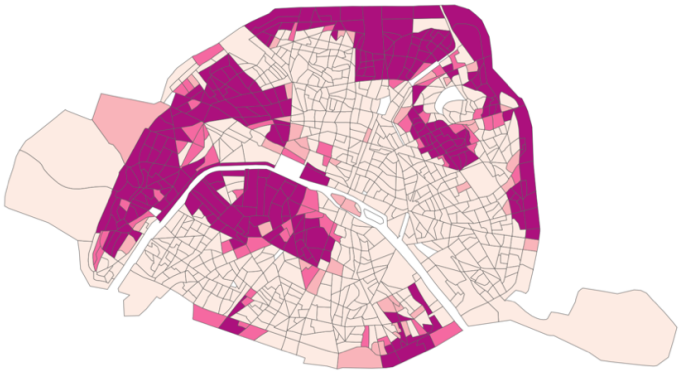
Ajustement des p-values

- Objectif : Diminuer la probabilité de rejeter H_0 à tort.
- Bonferroni :
  - Probabilité de ne pas rejeter à tort H_0 : $1-p$ par polygone $(1-p)^n$ pour toute la zone, avec n le nombre de polygones.
  - Probabilité de rejeter au moins une fois à tort H_0 : $1-(1-p)^n \approx np = p*$
  - $p \approx \frac{p*}{n}$: un regroupement est significatif à 0.05%

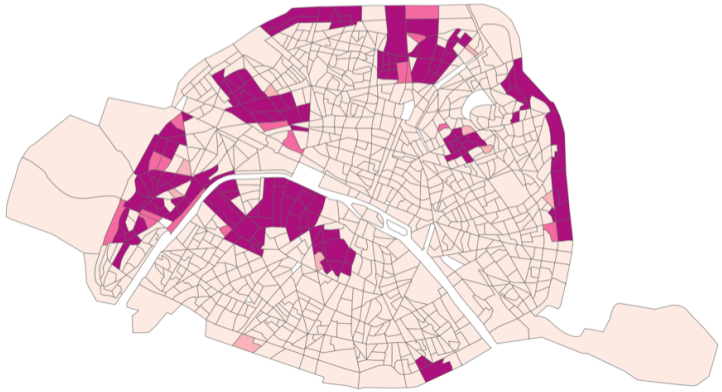
    si sa pvalue vaut $\frac{0.05}{n}$
- Limites : Risque de rejeter à tort des regroupements significatifs
- Alternatives : distributions simulées en tirant les valeurs des voisins dans une loi uniforme.

Significativité des LISA

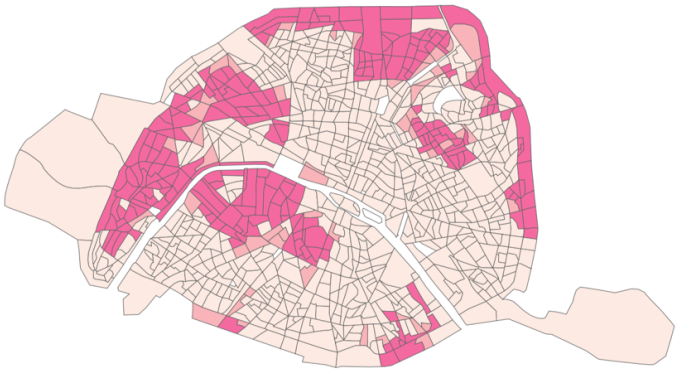
pvalue des I de Moran locaux



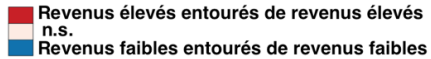
pvalue ajustée : Bonferroni



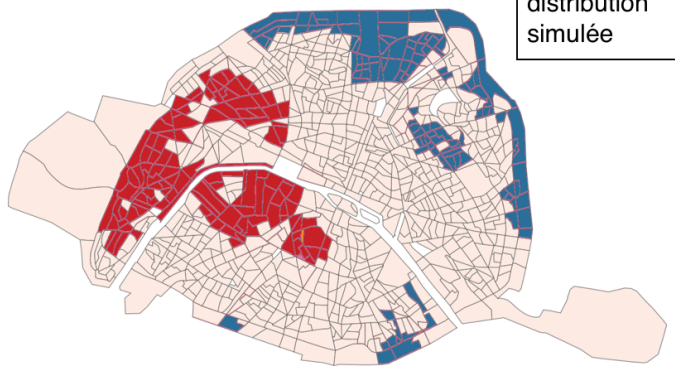
pvalue ajustée : simulation MonteCarlo



I de Moran locaux significatifs



pvalue issues d'une distribution simulée



Interprétation des LISA

En l'absence d'autocorrélation spatiale globale

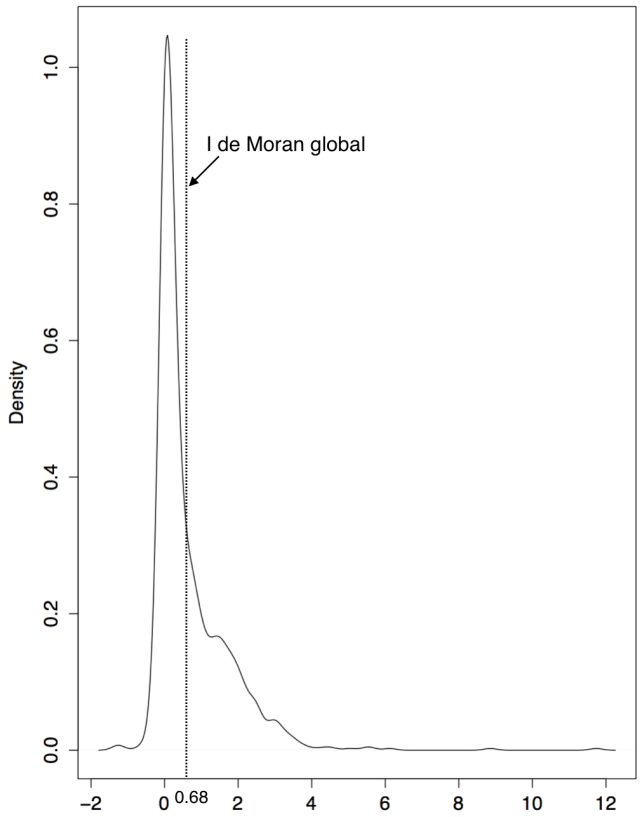
- Les p-values ajustées limitent le risque de rejeter H_0 à tort.
- Les LISA permettent de **détecter les zones où des valeurs similaires se regroupent de façon significative**.
- Structure spatiale locale : les liens entre voisins sont particulièrement forts.

En présence d'autocorrélation spatiale globale

- Même ajustées, les p-values risquent d'être trop faibles, puisque la distribution des I_i s'éloigne de la normale.
- Plus l'autocorrélation globale augmente, plus le nombre de valeurs extrêmes augmente.
- $\Rightarrow$ Deuxième interprétation des LISA.

Distribution des I de Moran locaux

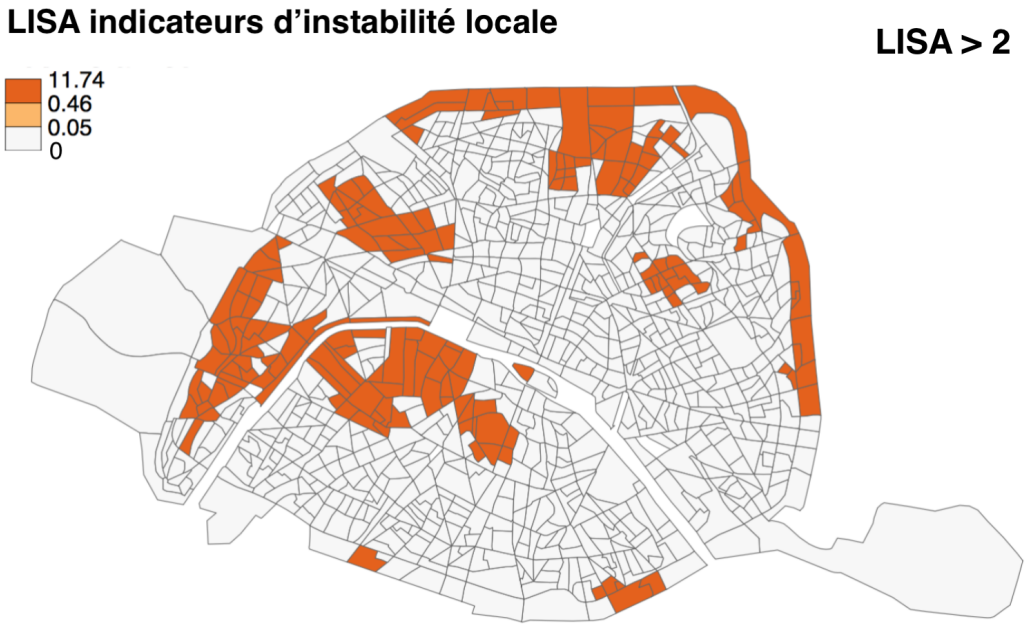
- Distribution non centrée sur le I de Moran global.
- **Certaines zones ont une structure d'association spatiale significativement différente du processus global.**



I de Moran locaux des revenus médians par IRIS

LISA comme indicateur d'instabilité locale

- **Indique les zones qui influent particulièrement sur le processus global** (autocorrélation locale plus marquée que l'autocorrélation globale), **ou au contraire qui s'en démarquent** (plus faible autocorrélation).



Conclusion

Analyse Spatiale Exploratoire

- Trois types de données spatiales aux propriétés différentes.
- Importance de la définition de la structure spatiale : voisins, matrice de poids, échelle d'agrégation.
- I de Moran, c de Geary pour mesurer la dépendance spatiale globale.
- LISA pour détecter les regroupements significatifs de valeurs semblables (quand pas d'autocorrélation globale) et les non-stationnarités locales (quand autocorrélation globale).

L'étape d'après :

- Analyser les déterminants des phénomènes spatiaux...
- ... grâce aux modèles d'économétrie spatiale.