Codifier la structure de voisinage

Auteurs

Marie-Pierre DE BELLEFON, Vincent LOONIS, Ronan LE GLEUT

INSEE

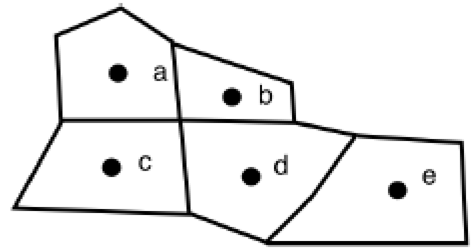
Pourquoi codifier la structure de voisinage ?

- *Objectif : définir des relations de voisinage **cohérentes** avec les véritables interactions spatiales entre les objets.*
  - *Définir le voisinage est indispensable pour appréhender de quelle manière les voisins s'influencent les uns les autres.*
  - *Plusieurs manières d'appréhender la notion de voisinage :*
    - *Contiguïté,*
    - *Distances entre observations, …*
- *Les relations spatiales sont multidirectionnelles et multilatérales.*
  - *A la différence des relations temporelles pour lesquelles les relations ne peuvent être que séquentielles le long de l'axe passé-présent-futur.*

Comment codifier la structure de voisinage ?

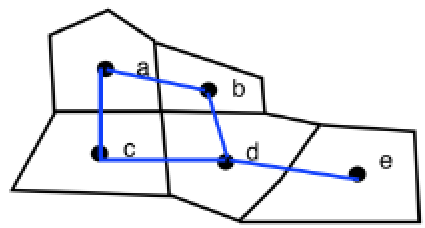
- *Transcrire la complexité de l'espace géographique en un ensemble fini de données analysables de manière systématique par un ordinateur.*

Carte



Carte avec les régions et leurs centroïdes

Spécification



Graphe de voisinage

	X	Y
a	x_a	y_a
b	x_b	y_b
c	x_c	y_c
d	x_d	y_d
e	x_e	y_e

Configuration spatiale

Source : TIEFELSDORF *1998*

Codification

	a	b	c	d	e
a	0	1	1	0	0
b	1	0	0	1	0
c	1	0	0	1	0
d	0	1	1	0	1
e	0	0	0	1	0

Matrice de contiguïté
Similarité

Seuil de coupure

	a	b	c	d	e
a	0	d_{ab}	d_{ac}	d_{ad}	d_{ae}
b	d_{ab}	0	d_{bc}	d_{bd}	d_{be}
c	d_{ca}	d_{cb}	0	d_{cd}	d_{ce}
d	d_{da}	d_{db}	d_{dc}	0	d_{de}
e	d_{ea}	d_{eb}	d_{ec}	d_{ed}	0

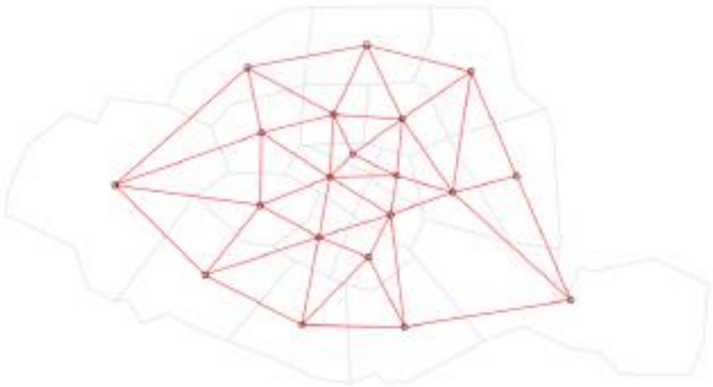
Matrice de distance
Dissimilarité

Définir les voisins en s'appuyant sur la distance

- *On peut calculer la distance qui sépare un ensemble de points répartis dans l'espace :*
  - *Lieux où l'information est observée.*
  - *Points représentatifs de chaque zone (ex : centroides).*
    - *On suppose que la répartition de la valeur de la variable d'intérêt au sein de chaque zone est suffisamment homogène pour que le point soit représentatif de la zone.*
- ***Graphes de voisinage*** *pour symboliser les liens entre les différentes entités*
  - *Ils doivent être tels que l'on se rapproche le plus possible de la structure spatiale sous jacente.*
- *Graphes de voisinage les plus fréquents :*
  - *Graphes fondés sur des notions géométriques.*
  - *Graphes fondés sur les plus proches voisins.*

Graphes de voisinage fondés sur des notions géométriques

Triangulation de Delauney



Graphe de la sphère d'influence

Graphe de Gabriel

Graphe des voisins relatifs

Graphes de voisinage fondés sur les plus proches voisins

Plus proche voisin

Deux plus proches voisins

Trois plus proches voisins

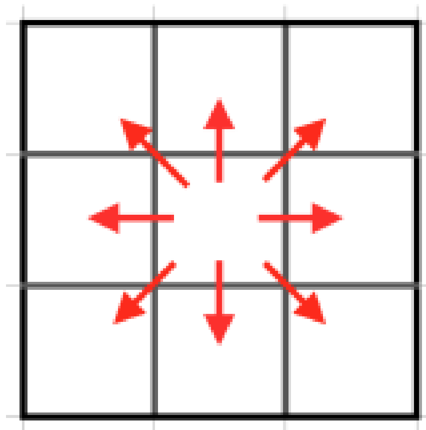
Voisins à la distance minimale

Définir les voisins en s'appuyant sur la contiguïté

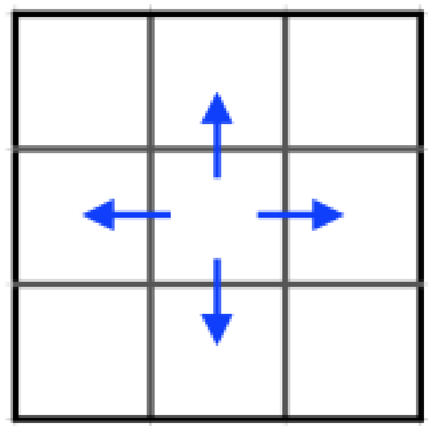
Contiguïté Queen : les voisins partagent un point de frontière commune.

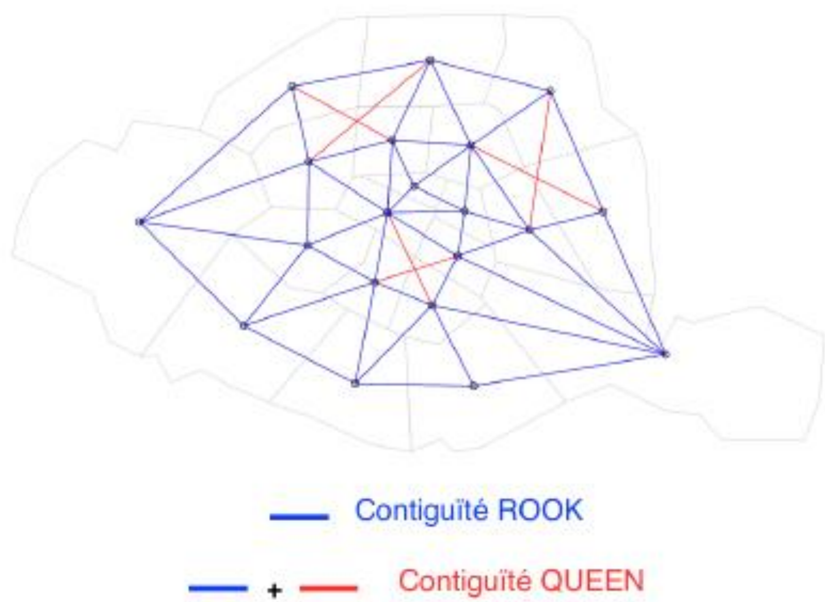
Déplacements de la Reine du jeu d'échecs

Contiguïté QUEEN



Contiguïté ROOK





Contiguïté Rook : les voisins possèdent au moins un segment de frontière commune.

Déplacements de la Tour du jeu d'échec

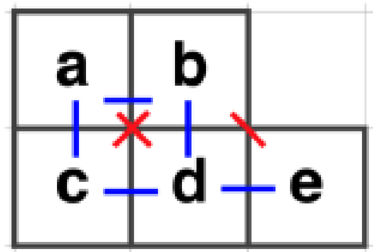
Accorder des poids aux voisins : la matrice de pondération spatiale

- ***Définition*** *: matrice carrée de dimension (N*N) de terme général w_{ij} non stochastique, non négatif, fini*
- ***Forme la plus simple*** *: matrice binaire symétrique ou asymetrique*
  - $w_{ij}=1$ *si i et j sont « voisins »*
    - *i et j sont contigus*
    - *i et j sont plus proches voisins*
    - d_{ij} *la distance qui sépare i et j est inférieure à D (distance critique)*
  - $w_{ij}=0$ *sinon*
  - *Par convention* $w_{ij}=0$ *si i=j*
- ***Standardisation en ligne***
  - *Moyenne pondérée des valeurs voisines*

$$w^s_{ij} = w_{ij} / \sum_j w_{ij} \text{ tels que } \sum_j w^s_{ij} = 1$$

  - *Les coefficients d'autocorrélation spatiale sont alors comparables d'un échantillon à l'autre*
  - *Si la matrice était symétrique, elle ne l'est plus après standardisation*

Exemple de matrice de pondération : la matrice de contiguïté



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

Matrices de pondérations spatiales fondées sur la distance

- ***Matrices de distance minimale (symétrique)***

  $w_{ij} = 1$ si $d_{ij} < D$ seuil de distance ou distance critique

  où d_{ij} est la distance entre i et j

  $w_{ij} = 0$ sinon

  $w_{ij} = 0$ si $i = j$ par convention

- ***Avantages***
  - *Plus de polygones sans voisins.*
- ***Inconvénients***
  - *Inefficace pour des données irrégulièrement réparties dans l'espace car la distance minimale nécessaire pour qu'un point relativement isolé ait au moins un voisin est beaucoup plus élevée que la distance du plus proche voisin d'un point situé dans une zone dense. Il y aura donc de grandes disparités dans le nombre de voisins*
  - *Potentiellement, nombre élevé de connexions qui ne sont pas pertinentes politiquement. Le choix des points (centroïdes ou capitales a de l'importance et doit être justifié).*

Matrices de pondérations spatiales fondées sur les plus proches voisins

- ***Matrices des k plus proches voisins (asymétrique)***

$w_{ij} = 1$ si i et j sont plus proche voisin
$w_{ij} = 0$ sinon
$w_{ij} = 0$ si $i = j$ par convention

- ***Chaque unité territoriale i a exactement k voisins quelle que soit la distance qui séparent ces voisins de i***
- ***Implique une distorsion de l'espace***
- ***Avantages***
  - *Plus de polygones sans voisins.*
  - *Moins « polluant » pour le calcul de la matrice de pondération*
- ***Inconvénients***
  - *Les paramètres retenus pour le calcul (nombre de voisins) peuvent ne pas refléter le degré de connectivité et/ou d'isolement.*

Indices d'autocorrélation spatiale

Auteurs

Salima BOUAYAD AGHA

Le Mans Université et CREST

Marie-Pierre DE BELLEFON

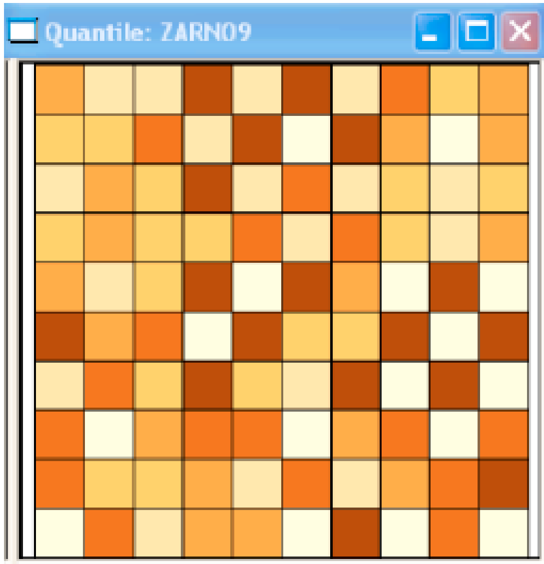
INSEE

Qu'est-ce que l'autocorrélation spatiale ?

- ***Corrélation, positive ou négative, d'une variable avec elle même du fait de la localisation spatiale des observations.***

- *Peut être le résultat de processus inobservés ou difficilement quantifiables qui associent des localisations différentes et qui de ce fait se traduisent par une structuration spatiale des activités :*
  - *Phénomènes d'interaction (entre les décisions des agents par exemple)*
  - *Phénomènes de diffusion (comme les phénomènes de diffusion technologique).*

Représentation de l'autocorrélation positive et de l'autocorrélation négative



Autocorrélation spatiale négative

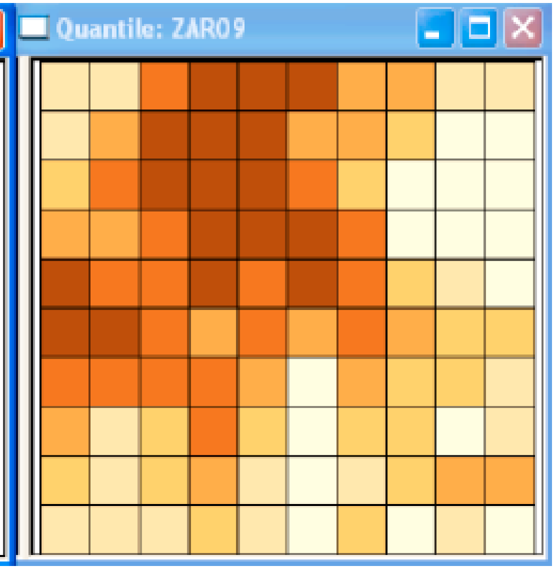
structure en échiquier : schéma de concurrence spatiale

Regroupement géographique de valeurs dissemblables de la variable à étudier

Pas d'autocorrélation spatiale

invariance par permutation aléatoire des individus

Répartition aléatoire de la variable à étudier



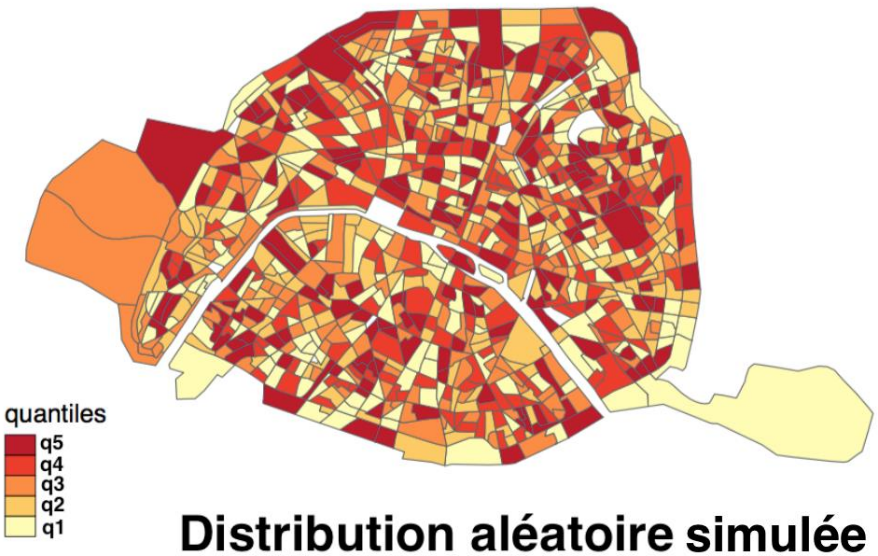
Autocorrélation spatiale positive

concentration spatiale de valeurs élevées et faibles

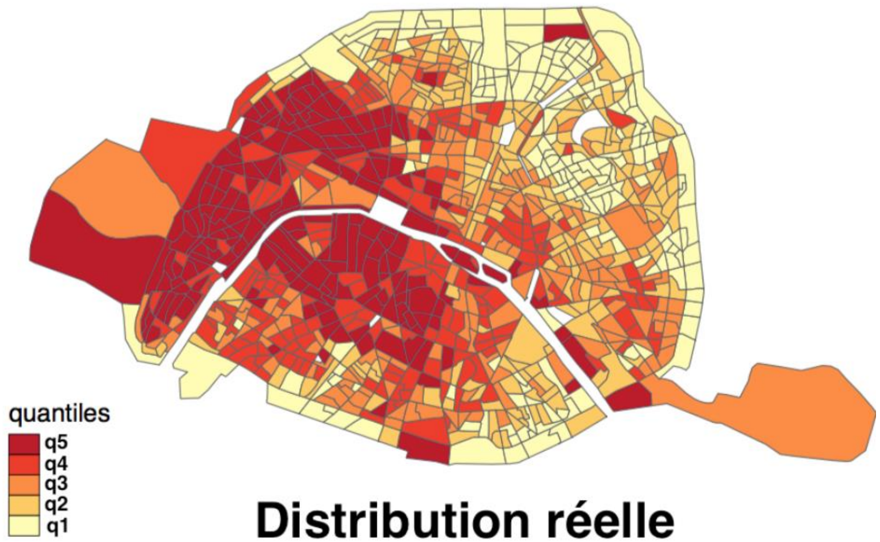
Regroupement géographique de valeurs similaires de la variable à étudier

Observation empirique

Distribution spatiale du revenu médian par IRIS *(source : RFL 2010, Insee)*

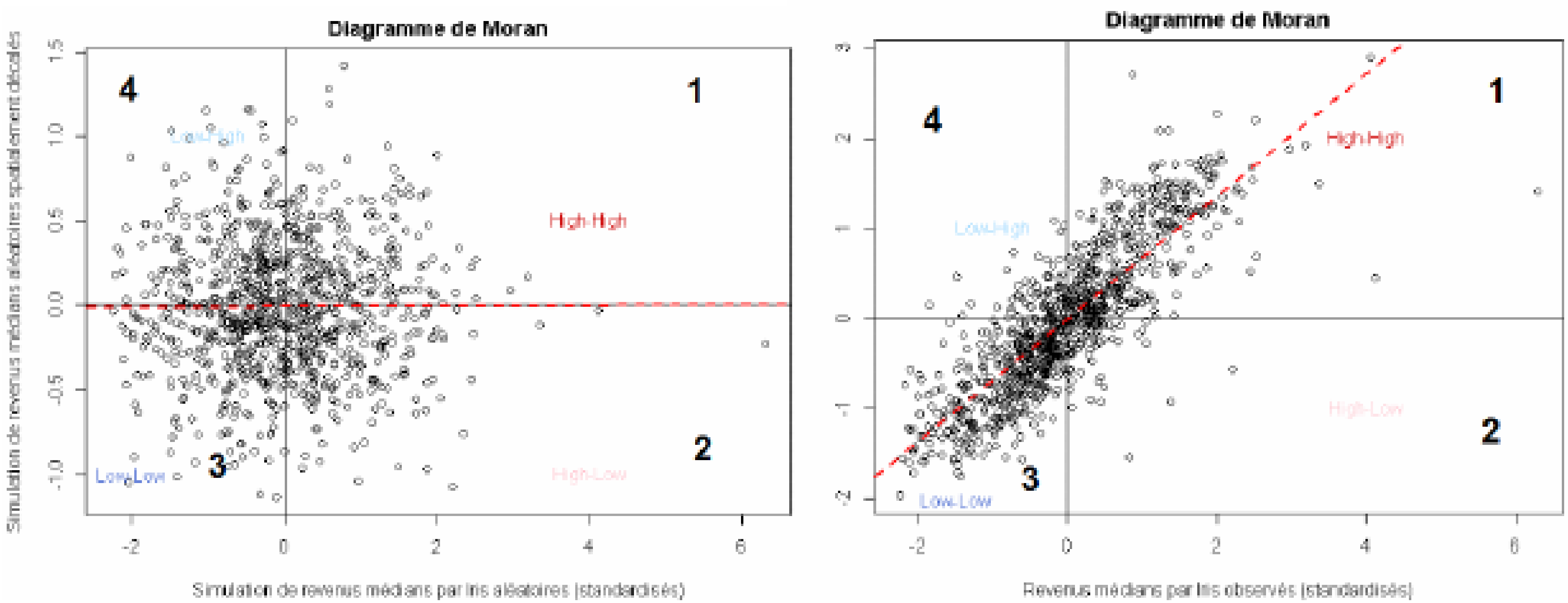


Distribution aléatoire simulée



Distribution réelle

Observation empirique : diagramme de Moran



On représente la valeur standardisée de la variable x (abscisse) et son décalage spatial Wx (ordonnée).

Les 4 quadrants représentent quatre formes d’association spatiale.

La densité des quadrants permet de voir le processus spatial dominant

Mesurer et tester la dépendance spatiale globale

- *Les indices d'autocorrélation spatiale permettent de caractériser la corrélation entre les mesures géographiquement voisines d'un phénomène mesuré.*
  - *Indice de Moran*
  - *Indice de Geary*
- *Test de l'hypothèse d'absence d'autocorrélation.*
  - *Tester si les valeurs prises par les observations voisines auraient pu être aussi comparables (ou aussi dissemblables) du simple fait du hasard.*
- *L'inférence est généralement menée sous l'hypothèse de:*
  - *normalité*
  - *randomisation*

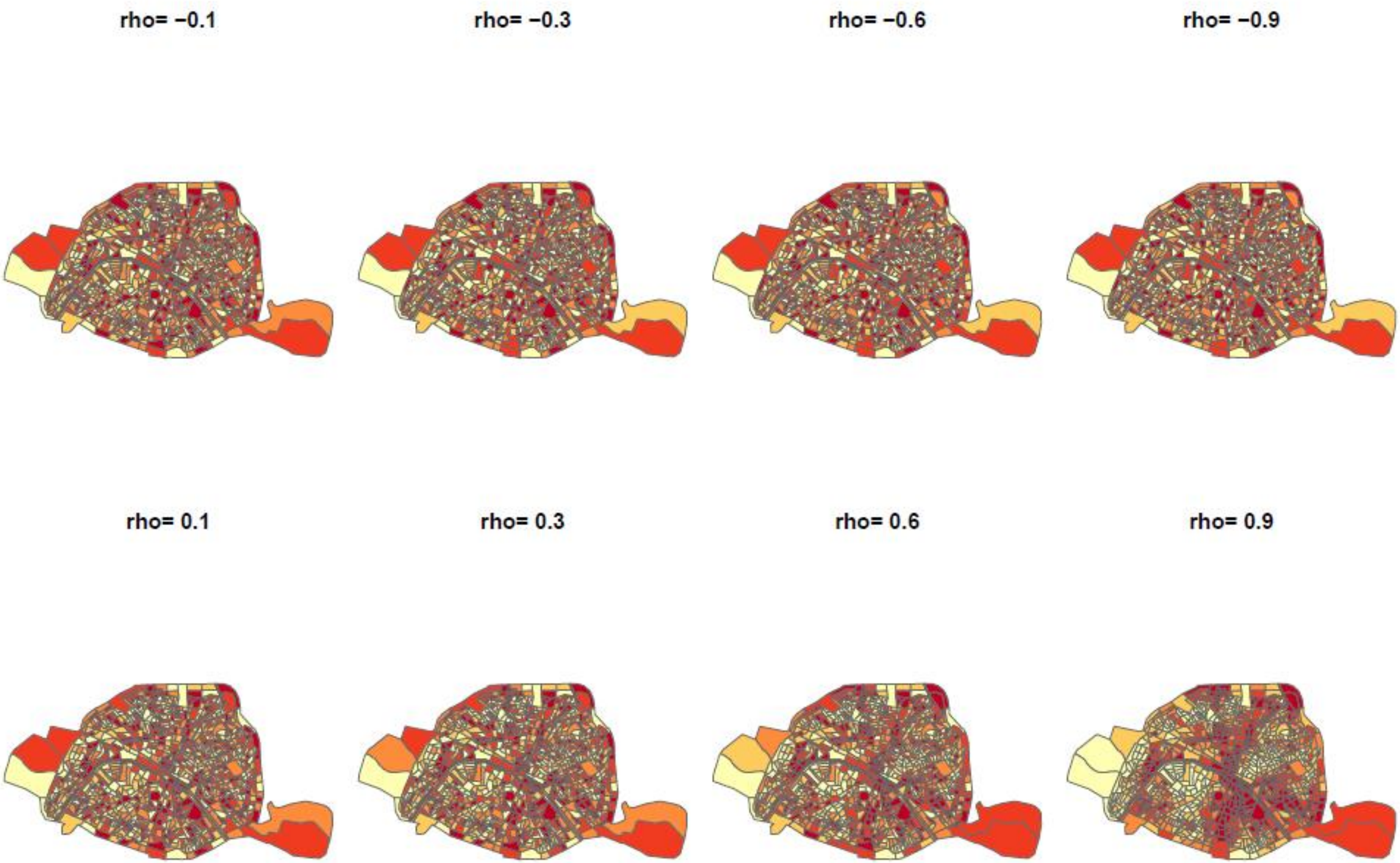
Le I de Moran et le C de Geary

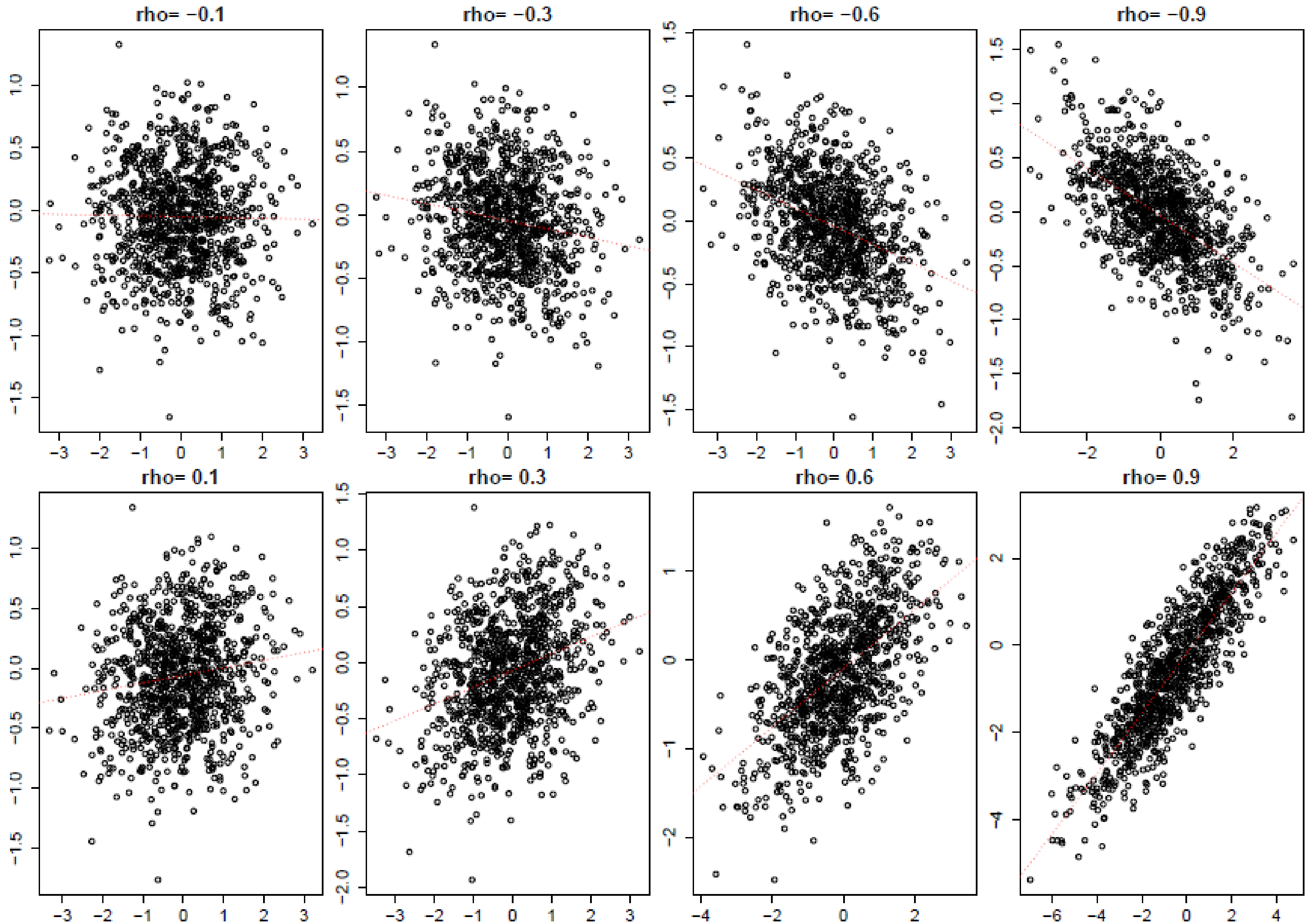
- *rapport entre la covariance pondérée entre les observations voisines et la variance totale observée en tenant compte des pondérations spatiales :*
- *Pour Moran, les produits croisées sont basés sur les écarts à la moyenne pour les deux localisations*

$$I = \left(\frac{n}{\left(\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}\right)} \right) \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

- *Pour Geary, les produits croisés utilisent les valeurs réelles pour chaque localisation*

$$C = \frac{(n-1) \sum_i \sum_j w_{ij}(y_i - y_j)^2}{2\left(\sum_i \sum_j w_{ij}\right) \sum_i (y_i - \bar{y})^2}$$


rho= −0.1
rho= −0.3
rho= −0.6
rho= −0.9
rho= 0.1
rho= 0.3
rho= 0.6
rho= 0.9

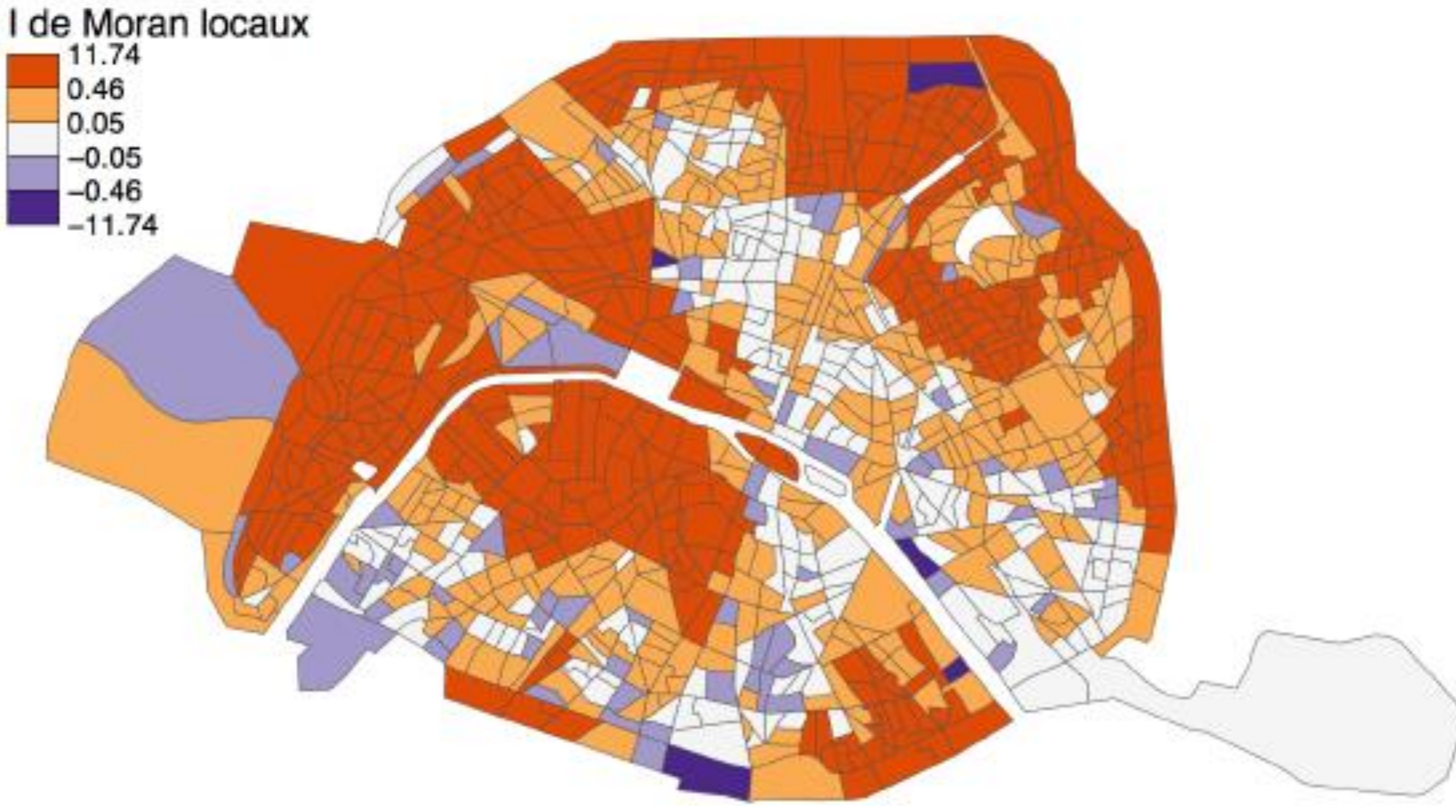

rho= −0.1
rho= −0.3
rho= −0.6
rho= −0.9
rho= 0.1
rho= 0.3
rho= 0.6
rho= 0.9

Analyse globale versus analyse locale

- ***Analyse globale***
  - *La statistique **I** de Moran caractérise la totalité de l'échantillon*
  - *c'est un indicateur unique du schéma d'autocorrélation globale*
  - *Elle met en évidence une concentration spatiale dans l'échantillon*
  - *Hypothèse d'homogénéité de l'échantillon*
- ***Analyse locale***
  - *Statistiques spécifiques à chaque observation/localisation*
  - *Elle met en évidence les localisations atypiques, les poches de non stationnarité locale, les outliers spatiaux comme le diagramme de Moran, mais elle permet de plus d'évaluer leur significativité (inférence statistique)*
  - *Elle met en évidence l'hétérogénéité de l'échantillon*

Les statistiques spatiales locales

- *Les statistiques globales ne permettent pas d'analyser la structure de l'autocorrélation spatiale.*
- *LISA (« Local indicators of spatial association ») sont définis pour une zone i comme une fonction f(zi,zj) mettant en relation la valeur observée dans la zone et celle qui est observée dans son voisinage.*
- *Les statistiques spatiales locales permettent pour chaque région i, de déterminer s'il existe une autocorrélation locale dans un voisinage de i*
- *Souvent utilisé après un test global pour voir si :*
  - *Le phénomène Y est homogène dans l'espace (les statistiques locales sont des valeurs similaires).*
  - *Il existe des aberrations locales responsables de la significativité de la statistique globale.*
  - *On peut les calculer même s'il n'y a pas d'autocorrélation globale*



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

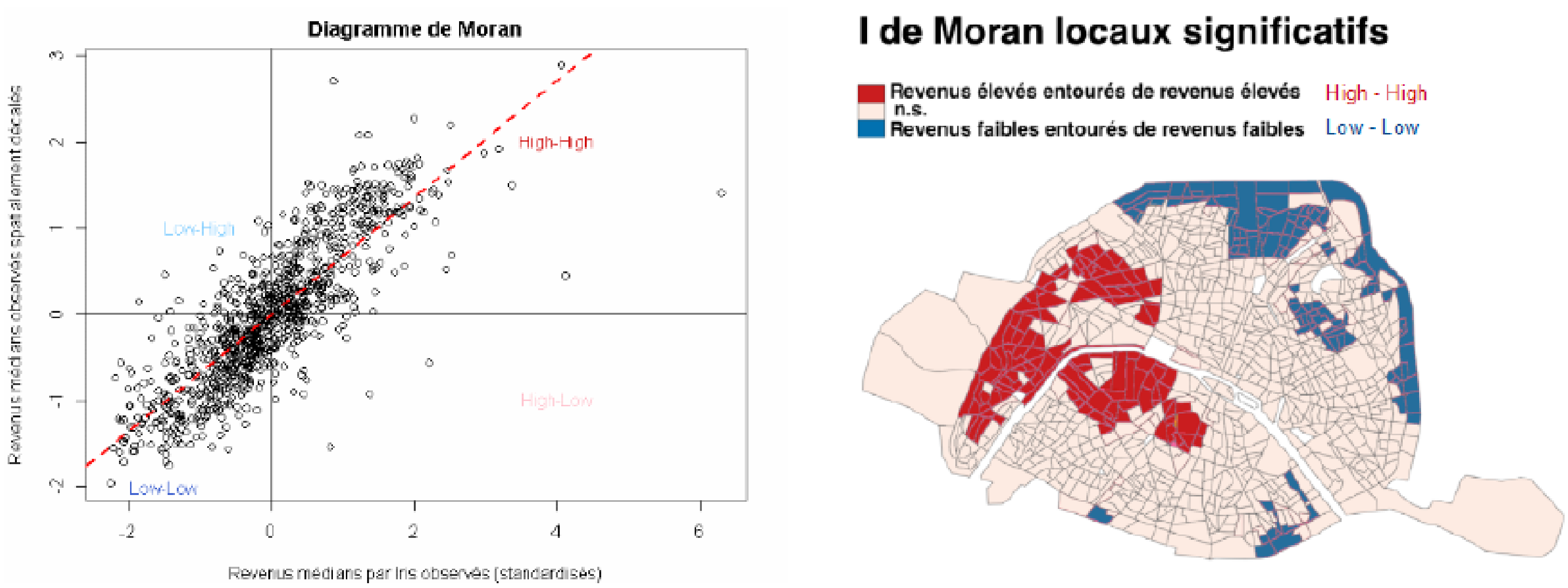


FIGURE 1.11 – I de Moran locaux significatifs

Merci de votre attention