3. Indices d'autocorrélation spatiale

BOUAYAD AGHA SALIMA
GAINS (TEPP) et Crest
Le Mans Université

MARIE-PIERRE DE BELLEFON
Insee

Résumé

Les indices d'autocorrélation spatiale permettent de mesurer la dépendance spatiale entre les valeurs d'une même variable en différents endroits de l'espace. Plus les valeurs des observations sont influencées par les valeurs des observations qui leur sont géographiquement proches, plus l'autocorrélation spatiale est élevée.

Ce chapitre définit l'autocorrélation spatiale, puis décrit les indices d'autocorrélation spatiale au niveau global et local : principes, propriétés, mise en œuvre pratique avec R et interprétation de leur significativité.

R La lecture préalable des chapitres 1 : "Analyse spatiale descriptive" et 2 : "Codifier la structure de voisinage" est recommandée.

Très souvent, les variables pour lesquelles on dispose d'informations géolocalisées se caractérisent par des dépendances spatiales qui sont d'autant plus fortes que les localisations sont plus proches. Ainsi, l'accès de plus en plus fréquent à des données spatialisées permet de mieux prendre en compte les interactions et les externalités spatiales dans l'analyse des décisions économiques des agents. Une analyse des structures spatiales comprises dans les données est indispensable pour traiter, si cela s'avère nécessaire, la violation de l'hypothèse d'indépendance spatiale des variables. D'autre part, en termes d'interprétation, l'analyse de l'autocorrélation spatiale permet une analyse quantifiée de la structure spatiale du phénomène considéré. Les indices d'autocorrélation spatiale permettent de mesurer la dépendance spatiale entre les valeurs d'une même variable en différents endroits de l'espace.

3.1 Qu'est-ce que l'autocorrélation spatiale ?

L'autocorrélation mesure la corrélation d'une variable avec elle-même, lorsque les observations sont considérées avec un décalage dans le temps (autocorrélation temporelle) ou dans l'espace (autocorrélation spatiale). On définit l'autocorrélation spatiale comme la corrélation, positive ou négative, d'une variable avec elle-même du fait de la localisation spatiale des observations. Cette autocorrélation spatiale peut, d'une part, 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 : des phénomènes d'interaction (entre les décisions des agents par exemple) ou de diffusion (comme les phénomènes de diffusion technologique) dans l'espace sont autant de phénomènes qui peuvent produire de l'autocorrélaton spatiale. D'autre part, dans le contexte de la spécification de modèles économétriques, la mesure de l'autocorrélation spatiale peut être envisagée comme un outil de diagnostic et de détection d'une mauvaise spécification (variables omises spatialement corrélées, erreurs sur le choix de l'échelle à laquelle le phénomène spatial est analysé, etc.)

D'un point de vue statistique, de nombreuses analyses (analyse des corrélations, régressions linéaires, etc.) reposent sur l'hypothèse d'indépendance des variables. Lorsqu'une variable est spatialement autocorrélée, l'hypothèse d'indépendance n'est plus respectée, remettant ainsi en cause la validité des hypothèses sur la base desquelles ces analyses sont menées. D'autre part, l'analyse de l'autocorrélation spatiale permet une analyse quantifiée de la structure spatiale du phénomène étudié.
On insistera sur le fait que structure spatiale et autocorrélation spatiale ne peuvent pas exister indépendamment l'une de l'autre (TIEFELSDORF 1998) :

— on désigne par structure spatiale l'ensemble des liens grâce auxquels le phénomène autocorrélé va se diffuser ;
— sans la présence d'un processus autocorrélé significatif, la structure spatiale ne peut être empiriquement observée.

La distribution spatiale observée est alors considérée comme la manifestation du processus spatial sous-jacent.

3.1.1 Observation empirique de l'autocorrélation spatiale

En présence d'autocorrélation spatiale, on observe que la valeur d'une variable pour une observation est liée aux valeurs de cette même variable pour les observations voisines.

— L'autocorrélation spatiale est positive lorsque des valeurs similaires de la variable à étudier se regroupent géographiquement.

— L'autocorrélation spatiale est négative lorsque des valeurs dissemblables de la variable à étudier se regroupent géographiquement : des lieux proches sont plus différents que des lieux éloignés. On retrouve généralement ce type de situation en présence de concurrence spatiale.
— En l'absence d'autocorrélation spatiale, on peut considérer que la répartition spatiale des observations est aléatoire.

Les indices d'autocorrélation spatiale permettent d'évaluer la dépendance spatiale entre les valeurs d'une même variable en différents endroits de l'espace et de tester la significativité de la structure spatiale identifiée. Pour la mettre en évidence, les indices prennent en compte deux critères :

— la proximité spatiale ;
— la ressemblance ou la dissemblance des valeurs de la variable pour les unités spatiales considérées.

Attention : si les données sont agrégées suivant un découpage qui ne respecte pas le phénomène sous-jacent, on surestimera ou sous-estimera la force du lien spatial.

On distingue la mesure de l'autocorrélation spatiale globale d'une variable dans un espace donné et celle de l'autocorrélation locale dans chaque unité de cet espace. Celle-ci mesure l'intensité et la significativité de la dépendance locale entre la valeur d'une variable dans une unité spatiale et les valeurs de cette même variable dans les unités spatiales voisines (plus ou moins proches, selon le critère de voisinage retenu).

3.1.2 Le diagramme de Moran

Le diagramme de Moran permet une lecture rapide de la structure spatiale. Il s'agit d'un nuage de points avec les valeurs de la variable y centrée en abscisse et les valeurs moyennes de la variable pour les observations voisines Wy en ordonnée (où W est la matrice de poids normalisée). Les deux propriétés y centrée et W normalisée impliquent que la moyenne empirique de Wy est égale à celle de y et donc à 0. On trace également la droite de régression linéaire de Wy en fonction de y et les droites d'équation $y = 0$ et $Wy = 0$ qui délimitent des quadrants.

Si les observations sont réparties aléatoirement dans l'espace, il n'y a pas de relation particulière entre y et Wy. La pente de la droite de régression linéaire est nulle, et les observations sont réparties uniformément dans chacun des quadrants. Si au contraire les observations présentent une structure spatiale particulière, la pente de la régression linéaire est non nulle puisqu'il existe une corrélation entre y et Wy. Chacun des quadrants définis par $y = 0$ et $Wy = 0$ correspond à un type d'association spatiale particulier (figures 3.1 et 3.2).

— Les observations situées en haut à droite (quadrant 1) présentent des valeurs de la variable plus élevées que la moyenne, dans un voisinage qui leur ressemble (autocorrélation spatiale positive et valeur de l'indice élevé ; structure high-high).
— En bas à gauche (quadrant 3), les observations présentent des valeurs de la variable plus faibles que la moyenne, dans un voisinage qui leur ressemble (autocorrélation spatiale positive et valeur de l'indice faible ; structure low-low).
— Les observations situées en bas à droite (quadrant 2) ont des valeurs de la variable plus élevées que la moyenne dans un voisinage qui ne leur ressemble pas (autocorrélation spatiale négative et valeur de l'indice élevé ; structure high-low).
— En haut à gauche (quadrant 4), les observations présentent des valeurs de la variable plus basses que la moyenne dans un voisinage qui ne leur ressemble pas (autocorrélation spatiale négative et valeur de l'indice faible ; structure low-high).

La densité de points dans chacun des quadrants permet de visualiser la structure spatiale dominante. Le diagramme de Moran permet aussi de visualiser les points atypiques qui s'éloignent de cette structure spatiale.

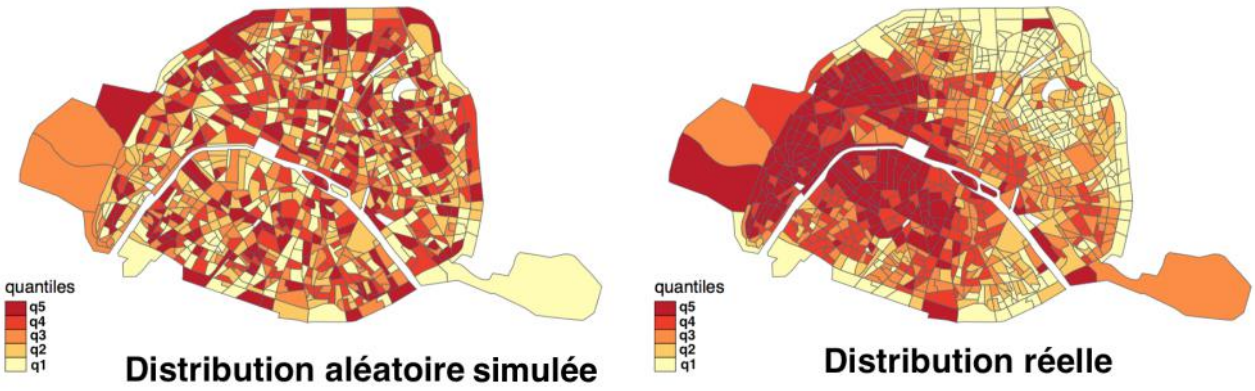


FIGURE 3.1 – Illustration, sur les Iris parisiens, de l'écart entre une distribution aléatoire et une distribution autocorrélée spatialement
Source : *Insee, Revenus Fiscaux Localisés 2010*

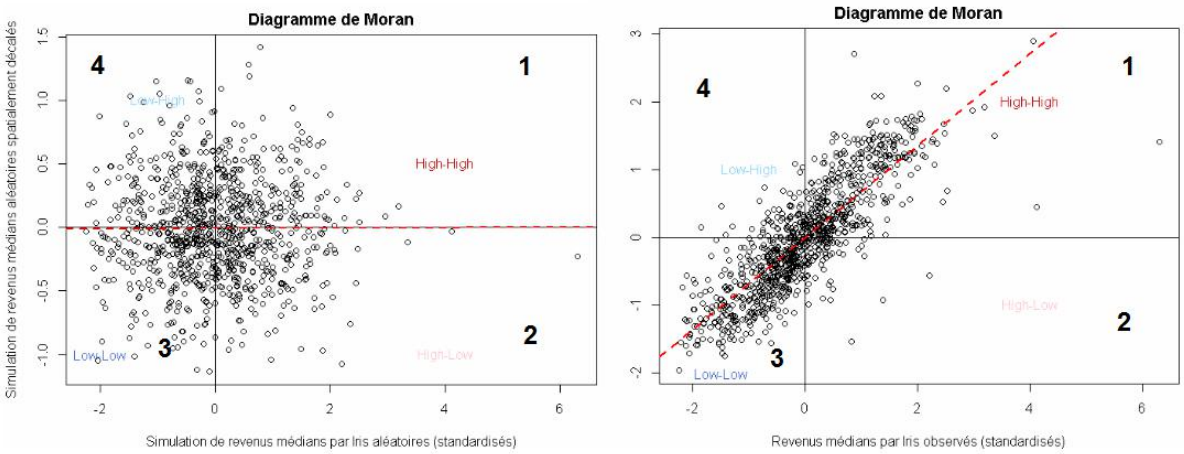


FIGURE 3.2 – Diagramme de Moran des revenus médians par Iris standardisés et d'une simulation de répartition aléatoire des revenus médians par Iris, pour les Iris parisiens
Source : *Insee, Revenus Fiscaux Localisés 2010*

Pour comprendre la façon dont l'autocorrélation spatiale est visible sur le diagramme de Moran, on a simulé une autocorrélation spatiale croissante des revenus par Iris (figures 3.3 et 3.4). Le paramètre ρ qui définit l'autocorrélation spatiale correspond à la pente du diagramme de Moran. À part pour les valeurs extrêmes, il est difficile d'identifier le signe et la force de l'autocorrélation spatiale en regardant simplement les cartes des différentes valeurs. En revanche, les diagrammes de Moran permettent d'identifier clairement les différents cas de figure.

3.2 Mesurer la dépendance spatiale globale

3.2.1 Indices d'autocorrélation spatiale

Lorsque le diagramme de Moran met en avant une structure spatiale particulière, le calcul des indices d'autocorrélation spatiale a pour objectif de répondre à deux questions :

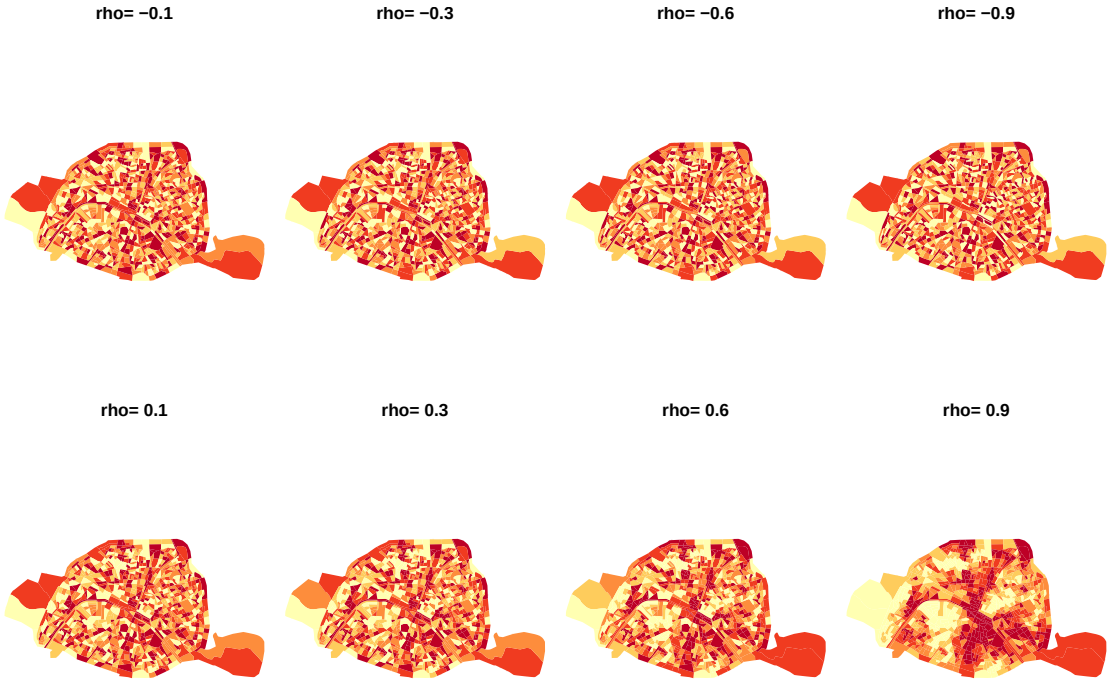


FIGURE 3.3 – Simulation d'une autocorrélation spatiale croissante des revenus par Iris parisiens
Source : *Insee, Revenus Fiscaux Localisés 2010*

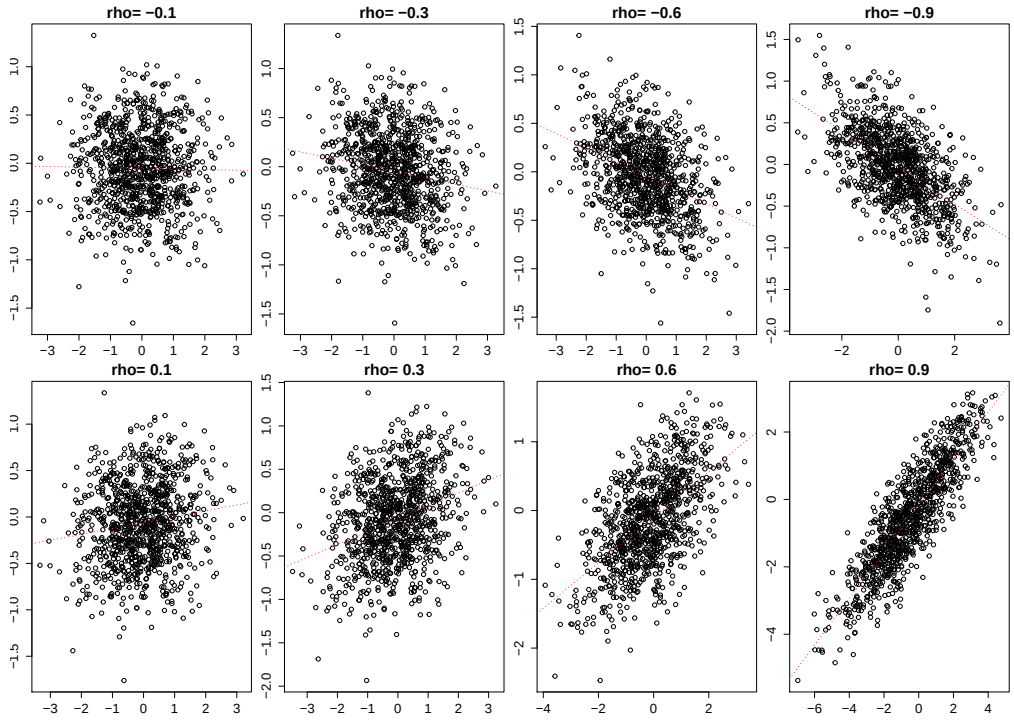


FIGURE 3.4 – Diagrammes de Moran associés aux simulations de revenus autocorrélés, pour les Iris parisiens
Source : *Insee, Revenus Fiscaux Localisés 2010*

— Les valeurs prises par les observations voisines auraient-elles pu être aussi comparables (ou aussi dissemblables) par le simple fait du hasard ?
— Si tel n'est pas le cas, il y a de l'autocorrélation spatiale : quels en sont le signe et la force ?

Répondre à la première question revient à tester l'hypothèse d'absence d'autocorrélation spatiale pour une variable brute y.
— H_0 : absence d'autocorrélation spatiale
— H_1 : présence d'autocorrélation spatiale
Pour mener à bien ce test, il faut préciser quelle est la distribution de la variable d'intérêt y, en l'absence d'autocorrélation spatiale (sous H_0). Dans ce contexte, l'inférence statistique est généralement menée en considérant l'une ou l'autre des deux hypothèses suivantes :

Hypothèse de normalité : chacune des valeurs de la variable, soit y_i, est le résultat d'un tirage indépendant dans la **distribution normale propre à chaque zone géographique i sur laquelle est mesurée cette variable**.

Hypothèse de randomisation : l'inférence sur le I de Moran est généralement menée sous l'hypothèse de randomisation. L'estimation de la statistique obtenue à partir des données est comparée avec **la distribution de celle obtenue en réordonnant au hasard (permutations) les données**. L'idée est simplement que si l'hypothèse nulle est vraie, alors toutes les combinaisons possibles des données sont équiprobables. Les données observées sont alors seulement l'une des réalisations parmi toutes celles également possibles. Dans le cas de l'autocorrélation spatiale, l'hypothèse nulle est toujours qu'il n'y a pas d'association spatiale et l'on affecte au hasard les valeurs de la variable aux unités spatiales afin de calculer la statistique du test. Si l'hypothèse nulle est rejetée, c'est-à-dire s'il y a de l'autocorrélation spatiale, on peut alors calculer l'intervalle de valeurs qui encadre l'indice d'autocorrélation spatiale et répondre ainsi à la question sur le signe et la force de l'autocorrélation spatiale : plus cet indice se rapproche de 1 en valeur absolue et plus la corrélation est élevée (cet intervalle dépend de la matrice de poids et peut parfois varier en dehors de l'intervalle $[-1;1]$, d'où l'intérêt de calculer les bornes de l'intervalle).

De manière très générale, les indices d'autocorrélation spatiale permettent de caractériser la corrélation entre les mesures géographiquement voisines d'un phénomène mesuré. Si l'on désigne par WY le vecteur des moyennes de la variable Y (où W désigne la matrice de pondération) dans le voisinage de chaque unité spatiale, les indices d'autocorrélation spatiale se mettent sous la forme :

$$Corr(Y,WY) = \frac{Cov(Y,WY)}{\sqrt{Var(Y).Var(WY)}} \tag{3.1}$$

À partir de cette formulation très générale, pour des variables quantitatives, deux indices sont principalement utilisés pour tester la présence d'autocorrélation spatiale : l'indice de Moran et l'indice de Geary. Le premier considère les variances et covariances en prenant en compte la différence entre chaque observation et la moyenne de toutes les observations. L'indice de Geary, lui, prend en compte la différence entre les observations voisines. Dans la littérature, l'indice de Moran est souvent préféré à celui de Geary en raison d'une stabilité générale plus grande (voir notamment UPTON et al. 1985).

Indice de Moran

$$I_W = \frac{n}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_i (y_i - \bar{y})^2} \quad i \neq j \tag{3.2}$$

— H_0 : Les voisins ne **co-varient** pas d'une façon particulière.
— $I_W > 0$ => autocorrélation spatiale positive.

Indice de Geary

$$c_W = \frac{n-1}{2\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij}(y_i - y_j)^2}{\sum_i (y_i - \bar{y})^2} \quad i \neq j \tag{3.3}$$

— H_0 : Les **différences** entre voisins n'ont pas de structure particulière.

— $c_W < 1$ => autocorrélation spatiale positive.

Selon la distribution retenue pour la variable en l'absence d'autocorrélation spatiale, le calcul de la variance des indices est modifié. En revanche, les équations qui donnent les expressions de l'espérance des indices (3.4) et la statistique du test (3.5) restent identiques. Ces relations permettent ainsi d'évaluer la significativité de l'autocorrélation spatiale.

$$\mathbb{E}(I_W) = \mathbb{E}(c_w) = -\frac{1}{n-1} \tag{3.4}$$

$$\frac{I_W - \mathbb{E}(I_W)}{\sqrt{Var(I_W)}} \sim \frac{c_W - \mathbb{E}(c_W)}{\sqrt{Var(c_W)}} \sim \mathcal{N}(0,1) \tag{3.5}$$

La mesure de l'autocorrélation spatiale reposant sur une comparaison de la valeur d'une variable pour un individu avec celle de ses voisins, la définition du voisinage va donc avoir un effet non négligeable sur la mesure de l'autocorrélation spatiale. Comme cela a été explicité dans le chapitre 2 "Codifier la structure de voisinage", plus le voisinage envisagé est large, plus on considère un nombre élevé de voisins, et plus la probabilité que leur moyenne se rapproche de la moyenne totale de la population va augmenter, ce qui risque de conduire à une valeur relativement faible de l'autocorrélation spatiale.

Un changement d'échelle peut également avoir des implications sur la mesure de l'autocorrélation spatiale. On désigne par MAUP (Modifiable Areal Unit Problem ; OPENSHAW et al. 1979b) l'influence du découpage spatial sur les résultats de traitements statistiques ou de modélisation. Plus précisément, les formes irrégulières et les limites des maillages administratifs qui ne reflètent pas nécessairement la réalité des distributions spatiales étudiées sont un obstacle à la comparabilité des unités spatiales inégalement subdivisées. Selon OPENSHAW 1984, le MAUP est une combinaison de deux problèmes distincts mais proches :

— le problème de l'échelle est lié à une variation de l'information engendrée lorsqu'un jeu d'unités spatiales est agrégé afin de former des unités moins nombreuses et plus grandes pour les besoins d'une analyse ou pour des questions de disponibilité des données ;
— le problème de l'agrégation (ou de zonage) est lié à un changement dans la diversité de l'information, engendré par les différents schémas possibles d'agrégation à une même échelle. Cet effet est caractéristique des découpages administratifs (particulièrement électoraux) et vient s'ajouter à l'effet d'échelle.

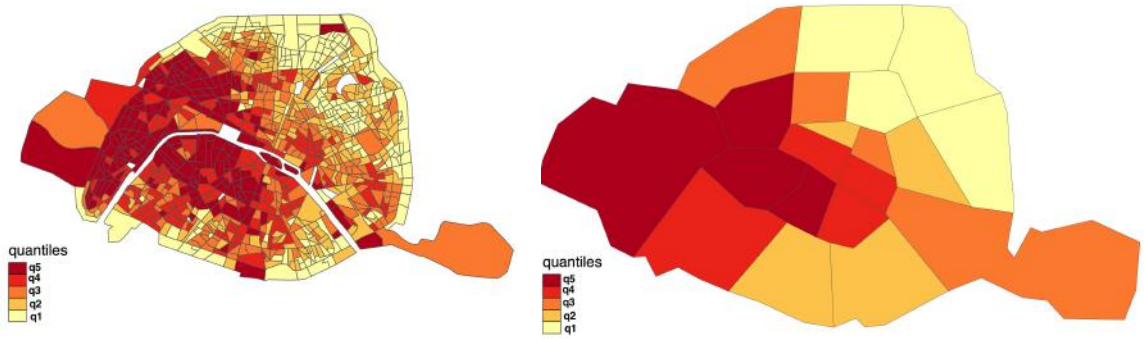
■ **Exemple 3.1 — Autocorrélation spatiale des revenus médians à Paris.** Quelle est l'intensité de l'autocorrélation spatiale des revenus parisiens ? Est-elle significative ? Dans quelle mesure dépend-elle de la spécification des relations spatiales (type de voisinage, échelle d'agrégation) ?

Source	I_W	c_w	p value	H0	bornes de I_W
Revenu : répartition observée	0.68	0.281	3.10^{-6}	rejetée	[-1.06,1.06]
Revenu : répartition aléatoire simulée	0.0027	1.0056	> 0.5	acceptée	[-1.06,1.06]

TABLE 3.1 – Indices de Moran et Geary des revenus médians des Iris parisiens : distribution réelle et simulée
Source : *Insee, Revenus Fiscaux Localisés 2010*

Type de voisinage	I_W	p value	H0
QUEEN	0.68	3.10^{-6}	rejetée
ROOK	0.57	2.10^{-6}	rejetée
1NN	0.30	0.07	rejetée
3NN	0.58	9.10^{-6}	rejetée
Delaunay	0.57	6.10^{-7}	rejetée

TABLE 3.2 – Indices de Moran et Geary des revenus médians des Iris parisiens en fonction de la définition du voisinage (voir chapitre 2 : "Codifier la structure de voisinage")
Source : *Insee, Revenus Fiscaux Localisés 2010*



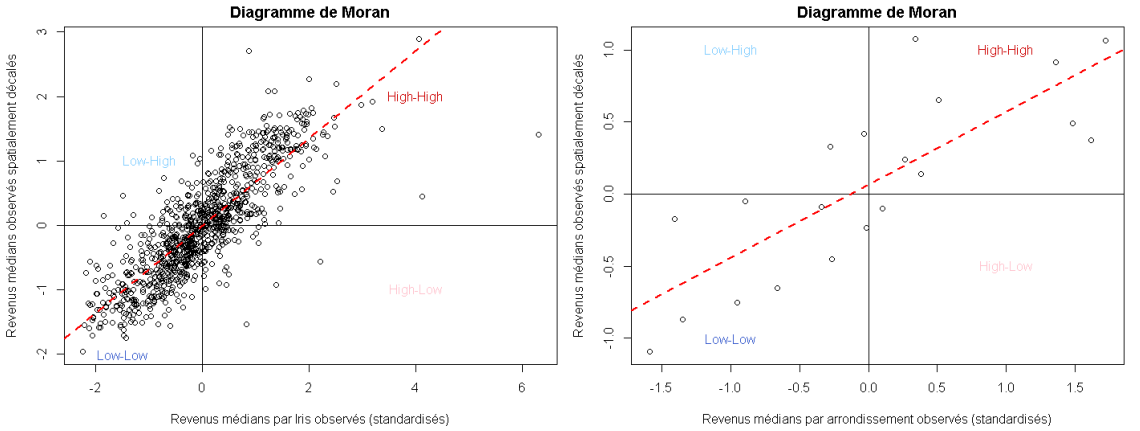
(a) à l'Iris

(b) à l'arrondissement

FIGURE 3.5 – Agrégation des revenus à Paris
Source : *Insee, Revenus Fiscaux Localisés 2010*

échelle d'agrégation	I_W	p value	H0	bornes de I_W
Iris	0.68	3.10^{-6}	rejetée	[-1.06,1.06]
Arrondissement	0.51	$< 9.10^{-9}$	rejetée	[-0.53,1.01]

TABLE 3.3 – Valeur et significativité du I de Moran en fonction de l'échelle d'agrégation choisie
Source : *Insee, Revenus Fiscaux Localisés 2010*



(a) à l'IRIS (b) à l'arrondissement

FIGURE 3.6 – Diagrammes de Moran pour la distribution des revenus à Paris
Source : *Insee, Revenus Fiscaux Localisés 2010*

Dans cet exemple (voir tables 3.1, 3.2, 3.3 et figures 3.5, 3.6), quelle que soit la définition du voisinage ou l'échelle d'agrégation, l'autocorrélation spatiale des revenus parisiens est positive et significative. La force de l'autocorrélation spatiale varie légèrement selon le type de voisinage retenu (en particulier, dans cet exemple, considérer uniquement les plus proches voisins diminue un peu la force de l'autocorrélation spatiale mesurée). ■

Application avec R

Le package *spdep* permet de calculer les indices d'autocorrélation spatiale et leur significativité grâce aux fonctions `moran.test` et `geary.test`.
Par défaut, la distribution de la variable d'intérêt sous l'hypothèse nulle est obtenue par randomisation. L'argument `randomisation = FALSE` permet de supposer qu'il s'agit d'une distribution normale.

Encadré 3.2.1 — Si certaines entités n'ont pas de voisins. Pour que les fonctions du package *spdep* acceptent des matrices de poids dans lesquelles certaines entités n'ont pas de voisins, il est nécessaire de spécifier l'option : `zero.policy=TRUE`. Par défaut, la taille de la matrice est diminuée pour exclure les observations n'ayant pas de voisins. On peut spécifier le contraire avec l'option : `adjust.n=FALSE`. Dans ce cas, la valeur absolue de la statistique de test augmente, et la valeur absolue de son espérance et de sa variance diminuent (BIVAND et al. 2013a). De façon générale, les indices d'autocorrélation spatiale ont été développés en supposant que toutes les entités avaient des voisins, et les avis ne sont pas unanimes sur l'attitude à adopter lorsque tel n'est pas le cas.

Comme vu précédemment, il existe deux approches pour estimer la significativité de ces indices : une solution analytique qui s'appuie sur l'hypothèse de normalité et une solution de Monte Carlo qui s'appuie sur l'hypothèse de randomisation. La solution analytique (utilisée par la fonction `moran.test`) fait l'hypothèse que la statistique du test suit asymptotiquement une loi normale de moyenne 0 et de variance 1. Cela peut ne pas toujours s'avérer être la mesure la plus précise de la significativité car la convergence vers cette loi peut dépendre de l'arrangement des polygones. On peut utiliser à la place la fonction `moran.mc` qui permet de choisir le nombre de permutations pour calculer la distribution simulée du I de Moran. Comparer les seuils de significativité calculés à partir des fonctions `moran.mc` et `moran.test` permet de s'assurer de la robustesse des conclusions.

```
library(spdep)

#####################################
# Préparation des données  ##########
#####################################

#Extraction de la liste des voisins (au sens Queen par défaut)
iris75.nb <- poly2nb(iris75)
#Création de la matrice de poids
iris75.lw <- nb2listw(iris75.nb,zero.policy=TRUE)
#Calcul des revenus médians standardisés
iris75.data <- as.data.frame(iris75)
iris75.data$med_revenu_std <- scale(iris75.data$med_revenu)

######################
# Diagramme de MORAN
######################

moran.plot(iris75.data$med_revenu_std,iris75.lw,labels=FALSE,
 xlab='revenus medians par IRIS',ylab='moyenne des revenus médians par IRIS
      des voisins')

######################
# Test du I de Moran
######################

moran.test(iris75.data$med_revenu_std,iris75.lw,zero.policy=TRUE,
    randomisation=FALSE)

#Calcul des intervales du I de Moran :
moran.range <- function(lw) {
    wmat <- listw2mat(lw)
    return(range(eigen((wmat+t(wmat))/2)$values))
}

moran.range(iris75.lw)
```

3.2.2 Autocorrélation spatiale des variables catégorielles

Lorsque la variable d'intérêt n'est pas continue, mais catégorielle, on mesure le degré d'association locale grâce à une analyse des statistiques des *join count* (ZHUKOV 2010).

Pour illustrer le calcul de ces statistiques, on considère une variable binaire qui représente deux couleurs, Blanc (B) et Noir (N) de sorte qu'une liaison puisse être qualifiée de Blanc-Blanc, Noir-Noir ou Blanc-Noir. On observe :

- une autocorrélation spatiale positive si le nombre de liaisons Blanc-Noir est significativement **inférieur** à ce que l'on aurait obtenu à partir d'une répartition spatiale aléatoire ;
- une autocorrélation spatiale négative si le nombre de liaisons Blanc-Noir est significativement **supérieur** à ce que l'on aurait obtenu à partir d'une répartition spatiale aléatoire ;

— aucune autocorrélation spatiale si le nombre de liaisons Blanc-Noir est approximativement **identique** à ce que l'on aurait obtenu à partir d'une répartition spatiale aléatoire.

S'il y a n observations, n_b observations blanches et $n_n = n - n_b$ observations noires, la probabilité d'obtenir une observation blanche est : $P_b = \frac{n_b}{n}$ et la probabilité d'obtenir une observation noire est : $P_n = 1 - P_b$.
En l'absence d'autocorrélation spatiale, les probabilités d'obtenir des observations d'une même couleur dans deux cellules voisines sont : $P_{bb} = P_b * P_b = P_b^2$ et $P_{nn} = P_n * P_n = (1 - P_b)^2$.
La probabilité d'obtenir des observations de couleur différente dans deux cellules voisines est : $P_{bn} = P_b * (1 - P_b) + (1 - P_b) * P_b = 2P_b * (1 - P_b)$.

Comme $\frac{1}{2}\sum_i \sum_j w_{ij}$ mesure le nombre de liaisons existantes, sous l'hypothèse d'une répartition spatiale aléatoire des observations, on peut écrire :

$$\begin{aligned}
\mathbb{E}[bb] &= \frac{1}{2}\sum_i \sum_j w_{ij} P_b^2 \\
\mathbb{E}[nn] &= \frac{1}{2}\sum_i \sum_j w_{ij} (1 - P_b)^2 \\
\mathbb{E}[bn] &= \frac{1}{2}\sum_i \sum_j w_{ij} 2P_b * (1 - P_b)
\end{aligned} \tag{3.6}$$

Si l'on désigne par $y_i = 1$ lorsque l'observation est de couleur noire et par $y_i = 0$ dans le cas contraire (couleur blanche), les contre-parties empiriques (valeurs observées) de ces espérances mathématiques peuvent s'écrire :

$$\begin{aligned}
nn &= \frac{1}{2}\sum_i \sum_j w_{ij} y_i y_j \\
bb &= \frac{1}{2}\sum_i \sum_j w_{ij} (1 - y_i)(1 - y_j) \\
bn &= \frac{1}{2}\sum_i \sum_j w_{ij} (y_i - y_j)^2
\end{aligned} \tag{3.7}$$

Dans ce cas, la statistique de test permettant d'évaluer la significativité de l'autocorrélation spatiale repose sur l'hypothèse qu'en l'absence d'autocorrélation spatiale, les statistiques de *join count* (*bb, nn* et *bn*) suivent une loi normale. On peut alors écrire :

$$\frac{bn - \mathbb{E}(bn)}{\sqrt{Var(bn)}} \sim \mathcal{N}(0,1) \quad \frac{bb - \mathbb{E}(bb)}{\sqrt{Var(bb)}} \sim \mathcal{N}(0,1) \quad \frac{nn - \mathbb{E}(nn)}{\sqrt{Var(nn)}} \sim \mathcal{N}(0,1) \tag{3.8}$$

■ **Exemple 3.2 — Statistiques *join count* pour l'emploi des individus parisiens.** [1]

On considère la variable binaire qui vaut 1 si l'individu i est chômeur et 0 sinon. On cherche à déterminer si les chômeurs parisiens sont plus regroupés dans l'espace que s'ils étaient répartis aléatoirement. Les statistiques de *join count* permettent de répondre à cette question. À partir de la table 3.4 on peut constater que la localisation des chômeurs est significativement corrélée, tout comme l'est celle des actifs.

1. L'objectif de cet exemple n'est pas de détailler les résultats d'une étude économique, mais d'illustrer les techniques mises en œuvre. Il n'y a aucune interprétation à en tirer.

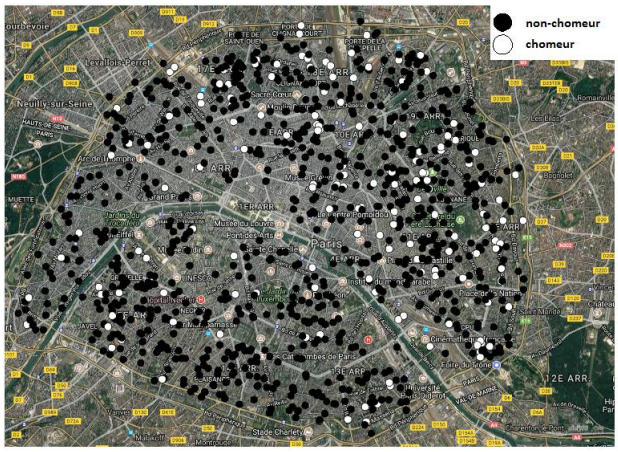


FIGURE 3.7 – Emploi d'un échantillon de 1000 individus parisiens
Source : *Insee, Revenus Fiscaux Localisés 2010*

Variable	p-value de la statistique joincount d'association spatiale	H0
Chômeur	$5.439.10^{-}3$	rejetée
Actif	$9.085.10^{-}5$	rejetée

TABLE 3.4 – Significativité de la statistique du joincount des chômeurs parisiens
Source : *Insee, Revenus Fiscaux Localisés 2010*

Application avec R

Cette statistique *joint count* est obtenue en mettant en œuvre la fonction `joincount.test` du package R *spdep*.

```
library(spdep)

#Conversion en facteur
menir10_d75_subset$chomage <- ifelse(menir10_d75_subset$ZCHOM>0,3,1)
chomage <- as.factor(menir10_d75_subset$chomage,levels=c("actif","chomeur")
    )

#Création des listes de voisins et matrices de poids
coordinates(menir10_d75_subset) <- c("PLG_X", "PLG_Y")
proj4string(menir10_d75_subset)<- CRS( "+init=epsg:27572 +proj=lcc +lat_
    1=46.8 +lat_0=46.8 +lon_0=0 +k_0=0.99987742 +x_0=600000 +y_0=2200000 +a
    =6378249.2 +b=6356515 +towgs84=-168,-60,320,0,0,0,0 +pm=paris +units=m
    +no_defs")
menir10_d75_subset <- spTransform (menir10_d75_subset, CRS ("+init=epsg
    :2154") )

menir75.nb<-  knn2nb(knearneigh(menir10_d75_subset,k=2))

#Mise en oeuvre du test
joincount.test(chomage,listw2U(nb2listw(menir75.nb)))
```

Dans le cas de plusieurs catégories, la fonction `joincount.multi` du package *spdep* permet

de tester la significativité, selon le même principe, de l'association spatiale de différentes variables.

■

3.3 Mesurer la dépendance spatiale locale

Les statistiques globales font l'**hypothèse de stationnarité du processus spatial** : l'autocorrélation spatiale serait la même dans tout l'espace. Or cette hypothèse est d'autant moins réaliste que le nombre d'observations est élevé.

3.3.1 Indice de Getis et Ord

Getis et Ord (GETIS et al. 1992) proposent un indicateur permettant de détecter les dépendances spatiales locales qui n'apparaissent pas dans l'analyse globale.

Indicateur de Getis et Ord

$$G_i = \frac{\sum_j w_{ij} y_j}{\sum_j w_{ij}} \tag{3.9}$$

$G_i > 0$ indique un regroupement de valeurs plus élevées que la moyenne.
$G_i < 0$ indique un regroupement de valeurs plus basses que la moyenne.

On peut tester la significativité de l'indicateur de Getis et Ord en faisant l'hypothèse, en l'absence de dépendance spatiale locale, d'une distribution normale.

$$z(G_i) = \frac{G_i - \mathbb{E}(G_i)}{\sqrt{Var(G_i)}} \sim \mathcal{N}(0,1) \tag{3.10}$$

Application avec R

La fonction `localG` du package *spdep* permet d'utiliser cet indicateur.

3.3.2 Indicateurs d'autocorrélation spatiale locale

Anselin (ANSELIN 1995) développe les notions introduites par Getis et Ord en définissant des *indicateurs d'autocorrélation spatiale locale*. Ceux-ci doivent permettre de mesurer l'intensité et la significativité de la dépendance locale entre la valeur d'une variable dans une unité spatiale et les valeurs de cette même variable dans les unités spatiales environnantes. Plus précisément, ces indicateurs permettent de :

— 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.

Les indicateurs de Getis et Ord ne répondent qu'au premier de ces deux objectifs. Pour être considérés comme des mesures locales d'association spatiale (LISA ; *Local Indicators of Spatial Association*) telles qu'elles ont été définies par Anselin, ces indicateurs doivent vérifier les deux propriétés suivantes :

— pour chaque observation, ils indiquent l'intensité du regroupement de valeurs similaires (ou de tendance opposée) autour de cette observation ;
— la somme des indices locaux sur l'ensemble des observations est proportionnelle à l'indice global correspondant.

Le LISA le plus couramment utilisé est le I de Moran local.

I de Moran local

$$I_i = (y_i - \bar{y}) \sum_j w_{ij}(y_j - \bar{y}) \tag{3.11}$$

$$I_W = constante * \sum_i I_i \tag{3.12}$$

$I_i > 0$ indique un regroupement de valeurs similaires (plus élevées ou plus faibles que la moyenne). $I_i < 0$ indique un regroupement de valeurs dissimilaires (par exemple des valeurs élevées entourées de valeurs faibles).

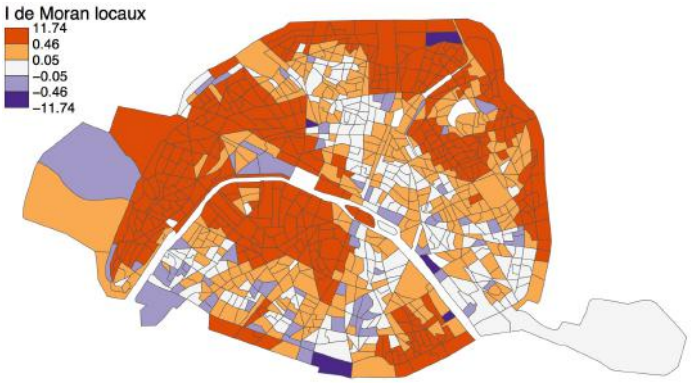


FIGURE 3.8 – Valeurs des I de Moran locaux, sur les Iris parisiens
Source : *Insee, Revenus Fiscaux Localisés 2010*

3.3.3 Significativité du I de Moran local

Des LISA significatifs correspondent à des regroupements de valeurs similaires ou dissimilaires plus marqués que ce que l'on aurait pu observer à partir d'une répartition spatiale aléatoire. Ces regroupements peuvent correspondre aux quatre types de regroupements spatiaux décrits en 3.1 et identifiables sur le diagramme de Moran (high-high, low-low, high-low ou low-low). Le test de significativité de chaque indicateur d'association locale repose sur une statistique supposée suivre asymptotiquement une loi normale sous l'hypothèse nulle. En effet, si l'on admet que l'hypothèse de normalité est vérifiée, $z(I_i) = \frac{I - \mathbb{E}(I_i)}{\sqrt{Var(I_i)}} \sim \mathcal{N}(0,1)$.

Pour tester la validité de l'hypothèse de normalité des LISA sous l'hypothèse nulle, on simule plusieurs répartitions aléatoires dans l'espace de la variable d'intérêt puis on calcule les indicateurs locaux associés à ces simulations.

En reprenant l'exemple des revenus parisiens, on observe (figure 3.10) que les quantiles extrêmes de la distribution des I locaux sont plus élevés que ceux d'une distribution normale. Les *p-values* calculées sous l'hypothèse de normalité devront donc être utilisées avec précaution. En effet, Anselin (ANSELIN 1995) montre, à partir de simulations (figure 3.11), **qu'en présence d'autocorrélation spatiale globale, l'hypothèse de normalité des I_i n'est plus vérifiée.**

D'autre part, le test de significativité des LISA soulève un problème que l'on rencontre à chaque fois que l'on effectue des comparaisons multiples. En effet, lorsque plusieurs tests statistiques sont réalisés simultanément à partir du même jeu de données, le risque global d'erreur de première espèce (probabilité de rejeter à tort l'hypothèse nulle) s'accroît. La répétition à chaque test du risque d'obtenir un résultat significatif par hasard augmente le risque global de conclure

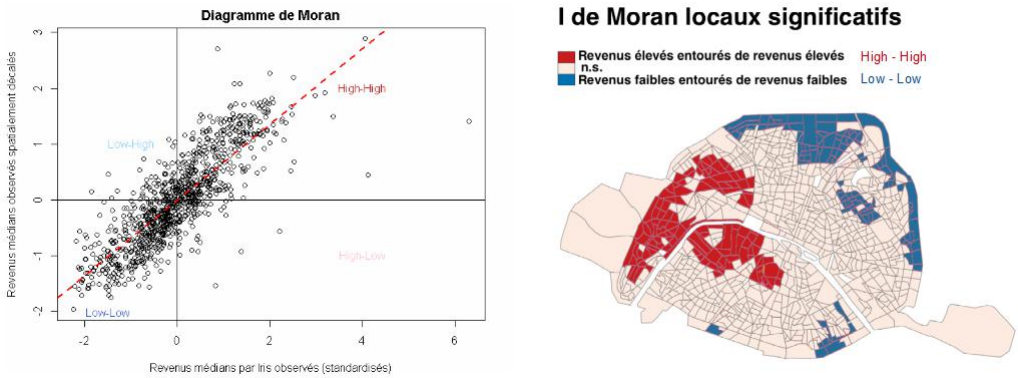


FIGURE 3.9 – I de Moran locaux significatifs, sur les Iris parisiens
Source : *Insee, Revenus Fiscaux Localisés 2010*

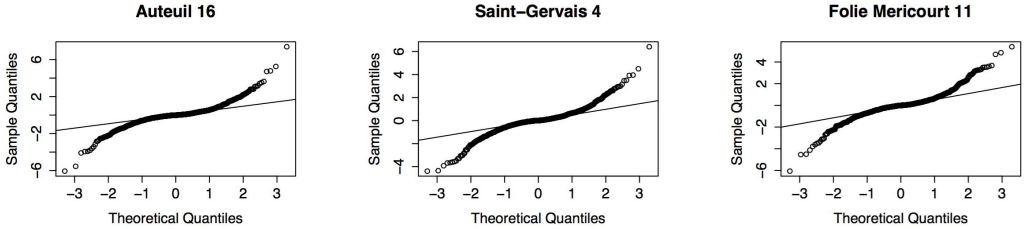


FIGURE 3.10 – Test de l'hypothèse de normalité de la distribution des I de Moran locaux sur trois Iris parisiens
Source : *Insee, Revenus Fiscaux Localisés 2010*

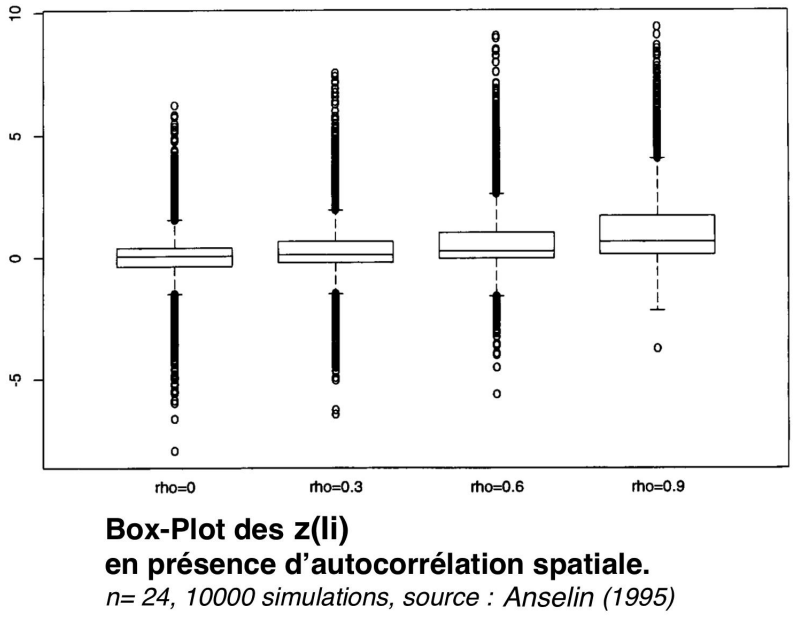


FIGURE 3.11 – Distribution des I de Moran locaux en présence d'autocorrélation spatiale globale
Source : *Insee, Revenus Fiscaux Localisés 2010*

à tort à la significativité de l'indice local. Ainsi, dans notre cas, on conclura à **l'existence** d'une autocorrélation spatiale locale si **au moins un** indice d'autocorrélation spatiale locale est significatif parmi tous les indices de la zone d'étude. S'il y a 100 indices d'autocorrélation spatiale locaux, on multiplie par 100 le risque d'en détecter au moins un significatif à tort (formule précisée en encadré 3.3.1). Il y a inflation du risque α (erreur de type I) : on augmente le risque de conclure à tort à l'existence d'une autocorrélation spatiale locale (ANSELIN 1995 ; ORD et al. 1995).

Différentes méthodes ont été développées pour éviter cette inflation du risque α lorsque plusieurs comparaisons statistiques sont nécessaires. Nous en détaillons quelques-unes dans ce qui suit.
Soit α le seuil de significativité retenu pour chaque indice local.

Encadré 3.3.1 — Méthode de Bonferroni : la méthode historique. La probabilité de ne pas rejeter à tort H_0 est $1-\alpha$ par polygone, donc $(1-\alpha)^n$ pour toute la zone, avec n le nombre de polygones.
La probabilité de rejeter au moins une fois à tort H_0 est $\alpha^* = 1-(1-\alpha)^n \approx n\alpha$.
Si l'on veut maintenir le risque global approximativement à α, on peut donc retenir $\alpha' \approx \frac{\alpha^*}{n}$ comme seuil de chaque test individuel. Ainsi pour $\alpha = 0.05$, un regroupement est significatif si sa p-value vaut $\frac{0.05}{n}$.
Le logiciel R permet de mettre en place cette méthode avec l'option `method='bonferroni'` de la fonction `p.adjust`.
On considère que cette méthode ne donne de bons résultats que lorsque le nombre de tests réalisés est petit. Dans le cas des LISA elle est un peu trop restrictive et l'on s'expose au risque, vu le nombre de comparaisons, de ne pas détecter certains LISA significatifs.

Encadré 3.3.2 — Méthode d'ajustement de Holm : permet de détecter l'existence d'un cluster spatial. La méthode d'ajustement de (HOLM 1979) prend en compte le fait que si parmi les n polygones, k sont vraiment des clusters spatiaux significatifs, la probabilité de rejeter à tort H_0 sur toute la zone n'est pas $(1-\alpha)^n$ mais $(1-\alpha)^{n-k}$ où α est le seuil de significativité souhaité.

La méthode de Holm classe les p-values de α_1 la plus faible à α_n la plus élevée.
Si $\alpha'_1 \sim n\alpha_1 < \alpha$, *i.e.* $\alpha_1 < \frac{\alpha}{n}$, on considère que cet indice local est effectivement significatif puisqu'il remplit le critère le plus restrictif. On regarde alors si $\alpha_2 < \frac{\alpha}{n-1}$, et ainsi de suite jusqu'à tester si $\alpha_k < \frac{\alpha}{n-k+1}$.
Le logiciel R permet de mettre en place cette méthode avec l'option `method='holm'` de la fonction `p.adjust`.
La méthode d'ajustement de Holm conduit à un plus grand nombre de clusters significatifs que la méthode de Bonferroni. Elle lui est donc le plus souvent préférée. Cependant, cette méthode se concentre aussi sur la détection **de l'existence d'au moins un cluster dans toute la zone**.

Encadré 3.3.3 — Méthode du False Discovery Rate : permet de localiser les clusters spatiaux. La méthode du False Discovery Rate (FDR) a été introduite par BENJAMINI et al. 1995. Avec cette méthode, le risque de juger - à tort - un cluster comme significatif est plus élevé, mais inversement le risque de juger - à tort - un cluster comme non significatif est plus faible. CALDAS DE CASTRO et al. 2006 prouvent l'intérêt de cette méthode pour **localiser** les clusters spatiaux significatifs.
La méthode du FDR classe les p-values de α_1 la plus faible à α_n la plus élevée.

Soit k le plus grand entier tel que $\alpha_k \leq \frac{k}{n}\alpha$. Benjamini et Hochberg expliquent qu'on peut rejeter l'hypothèse nulle d'absence d'autocorrélation spatiale locale pour tous les clusters dont les p-values sont inférieures ou égales à α_k.
Le logiciel R permet de mettre en place cette méthode avec l'option `method='fdr'` de la fonction `p.adjust`.

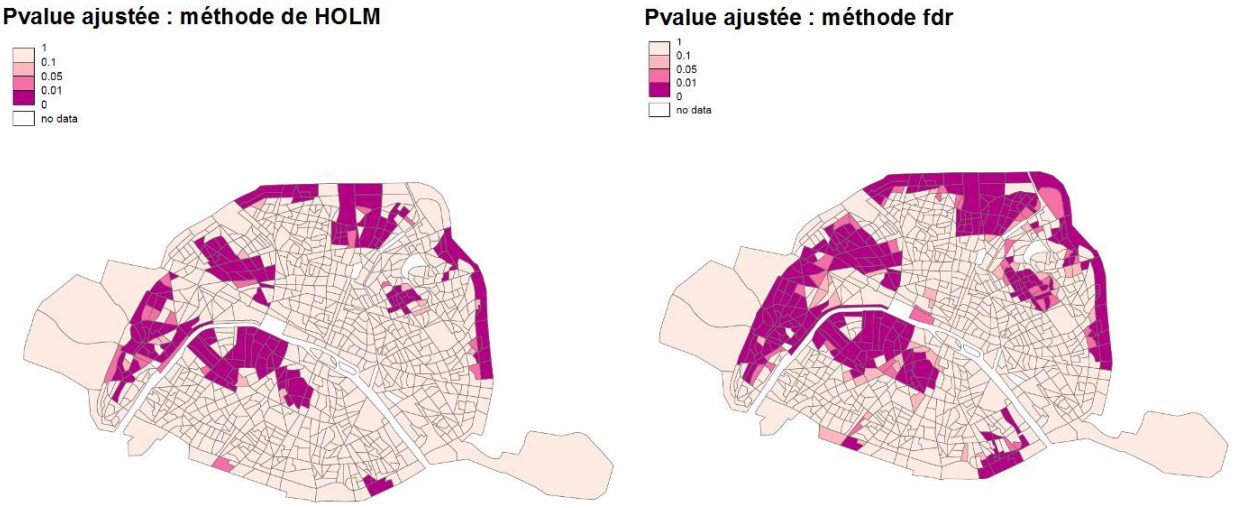


FIGURE 3.12 – Test de la significativité des I de Moran locaux, sur les Iris parisiens
Source : *Insee, Revenus Fiscaux Localisés 2010*

Dans l'exemple des revenus parisiens (figure 3.12), on observe bien que l'ajustement des p-values par la méthode de Holm conduit à moins de p-values significatives que l'ajustement par la méthode FDR. La méthode de Holm diminue en effet le risque de conclure à tort à l'**existence** d'une autocorrélation spatiale locale. En revanche, cette méthode augmente le risque de passer à côté d'un cluster local. Le choix de la méthode d'ajustement dépendra donc des objectifs de l'étude et des risques que l'on privilégie.

Application avec R

```
lisa_revenus<-  localmoran(iris75.data$med_revenu,iris75.lw,zero.policy=
    TRUE)

#Calcul des p-values ajustées
iris75.data.LISA$pvalue_ajuste<-
  p.adjust(iris75.data.LISA$pvalue_LISA,method='bonferroni')
```

3.3.4 Interprétation des indices locaux

En l'absence d'autocorrélation spatiale globale

Les LISA permettent de **détecter les zones où des valeurs similaires se regroupent de façon significative**. Il s'agit de zones où la structure spatiale locale est telle que les liens entre voisins sont particulièrement forts.

En présence d'autocorrélation spatiale globale

Les LISA **indiquent les zones qui influent particulièrement sur le processus global** (autocorrélation locale plus marquée que l'autocorrélation globale), **ou au contraire qui s'en dé-**

marquent (plus faible autocorrélation). Ainsi, dans l'exemple des revenus médians parisiens, on observe que la distribution des I de Moran locaux n'est pas centrée sur le I de Moran global (figure 3.13). Certaines zones ont une structure d'association spatiale significativement différente du processus global.

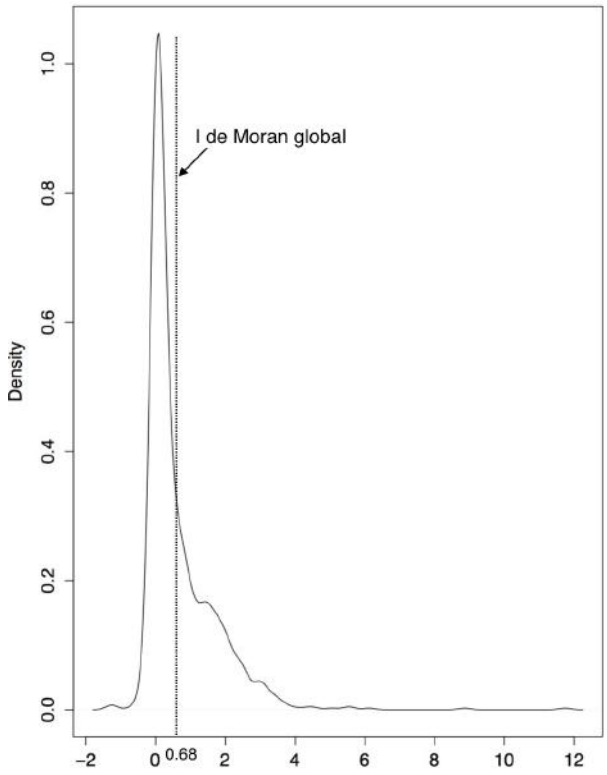


FIGURE 3.13 – Distribution des I de Moran locaux des revenus médians par Iris parisiens
Source : *Insee, Revenus Fiscaux Localisés 2010*

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. Les LISA élevés peuvent donc difficilement être interprétés comme des regroupements significatifs de valeur similaire. Dans ce cas, on interprète les LISA comme indicateurs d'une certaine **instabilité locale**.

3.4 Indices spatio-temporels

Il n'est pas rare qu'une base de données géolocalisée comporte des observations relevées à différents moments dans le temps, comme c'est le cas avec les bases de données recensant les transactions sur les biens immobiliers. Il peut être intéressant de comprendre comment un phénomène localisé s'est diffusé et a évolué dans l'espace et dans le temps et comment cela peut être lié aux conditions de l'environnement qui l'entoure. Dans ce cas, il faut être en mesure d'évaluer comment les structures spatiales sous-jacentes se modifient à différentes périodes de temps. Sur des données spatio-temporelles, l'exploration graphique préalable des données en coupe (I de Moran classique) peut permettre d'étudier l'existence et l'évolution des tendances de regroupement ou de dispersion qui sont statistiquement significativement différentes des modèles aléatoires.

De nombreux développements récents montrent un intérêt croissant pour l'analyse des données spatio-temporelles dans de nombreux domaines de recherche tels que la physique, la météorologie, l'économie et les études environnementales. En prolongeant l'indice de Moran pour y inclure des attributs temporels, il devient possible de calculer des indices globaux et localisés qui tiennent compte simultanément des autocorrélations spatiale et temporelle. Cela peut également se faire à partir de matrices de pondérations spatio-temporelles. Les travaux de MARTIN et al. 1975, WANG et al. p.d., LÓPEZ-HERNÁNDEZ et al. 2007 proposent des extensions du I de Moran traditionnellement utilisé pour mesurer la dépendance spatiale, pour calculer un I de Moran spatio-temporel. CHEN et al. 2013 développent une approche analytique améliorée fondée sur le I de Moran traditionnel et qui repose sur la stationnarité des données dans le temps. Comme le remarquent LEE et al. 2017, les séries temporelles géolocalisées sont généralement non stationnaires. Lorsque cette hypothèse n'est pas respectée, l'indice de Moran spatio-temporel proposé par CHEN et al. 2013 peut être fallacieux. LEE et al. 2017 proposent de contourner cette difficulté en appliquant une correction des fluctuations autour de la tendance (detrended fluctuation analysis, DFA) et suggèrent une nouvelle méthode de calcul de cet indice.

Conclusion

Les indices d'autocorrélation spatiale sont des outils de statistique exploratoire qui permettent de mettre en évidence l'existence d'un phénomène spatial significatif. Les sections 2 et 3 présentent des méthodes différentes pour prendre en compte ce phénomène spatial, au niveau global ou local, pour des variables quantitatives ou qualitatives. Il est important de savoir si l'autocorrélation est significative ou pas, mais également de mesurer la portée de l'autocorrélation afin de déterminer l'échelle de la dépendance spatiale. L'étude de l'autocorrélation spatiale est une étape indispensable avant d'envisager toute spécification des interactions spatiales dans un modèle approprié.

Références - Chapitre 3

ANSELIN, Luc (1995). « Local indicators of spatial association—LISA ». *Geographical analysis* 27.2, p. 93–115.

BENJAMINI, Yoav et Yosef HOCHBERG (1995). « Controlling the false discovery rate : a practical and powerful approach to multiple testing ». *Journal of the royal statistical society. Series B (Methodological)*, p. 289–300.

BIVAND, Roger S, Edzer PEBESMA et Virgilio GOMEZ-RUBIO (2013a). *Applied spatial data analysis with R*. T. 10. Springer Science & Business Media.

CALDAS DE CASTRO, Marcia et Burton H SINGER (2006). « Controlling the false discovery rate : a new application to account for multiple and dependent tests in local statistics of spatial association ». *Geographical Analysis* 38.2, p. 180–208.

CHEN, S.-K et al. (2013). « Analysis on Urban Traffic Status Based on Improved Spatiotemporal Moran's I ». *Acta Physica Sinica* 62.14.

GETIS, Arthur et J Keith ORD (1992). « The analysis of spatial association by use of distance statistics ». *Geographical analysis* 24.3, p. 189–206.

HOLM, Sture (1979). « A simple sequentially rejective multiple test procedure ». *Scandinavian journal of statistics*, p. 65–70.

LEE, Jay et Shengwen LI (2017). « Extending moran's index for measuring spatiotemporal clustering of geographic events ». *Geographical Analysis* 49.1, p. 36–57.

LÓPEZ-HERNÁNDEZ, Fernando A et Coro CHASCO-YRIGOYEN (2007). « Time-trend in spatial dependence : Specification strategy in the first-order spatial autoregressive model ». *Estudios de Economia Aplicada* 25.2.

MARTIN, Russell L et JE OEPPEN (1975). « The identification of regional forecasting models using space : time correlation functions ». *Transactions of the Institute of British Geographers*, p. 95–118.

OPENSHAW, Stan (1984). *The modifiable areal unit problem*. T. CATMOG 38. GeoBooks, Norwich, England.

OPENSHAW, Stan et Peter TAYLOR (1979b). « A million or so correlation coefficients ». *Statistical methods in the spatial sciences*, p. 127–144.

ORD, J Keith et Arthur GETIS (1995). « Local spatial autocorrelation statistics : distributional issues and an application ». *Geographical analysis* 27.4, p. 286–306.

TIEFELSDORF, Michael (1998). « Modelling spatial processes : The identification and analysis of spatial relationships in regression residuals by means of Moran's I (Germany) ». Thèse de doct. Université Wilfrid Laurier.

UPTON, Graham, Bernard FINGLETON et al. (1985). *Spatial data analysis by example. Volume 1 : Point pattern and quantitative data.* John W & Sons Ltd.

WANG, Y. F. et H. L. HE. « Spatial Data Analysis Method ». *Science Press, Beijing, China*.

ZHUKOV, Yuri M (2010). « Applied spatial statistics in R, Section 2 ». *Geostatistics.[Online] Available : http ://www. people. fas. harvard. edu/˜ zhukov/Spatial5. pdf*.