Statistique spatiale

Université Paris-Est Créteil
Serge Lhomme

Maître de conférences en géographie
http://sergelhomme.fr/
serge.lhomme@u-pec.fr

L'autocorrélation spatiale

Présentation

Compte tenu du caractère inégalitaire de nombreuses distributions, des objets géographiques se ressemblent plus que d'autres. Une question d'analyse spatiale que l'on est alors en droit de se poser est la suivante :

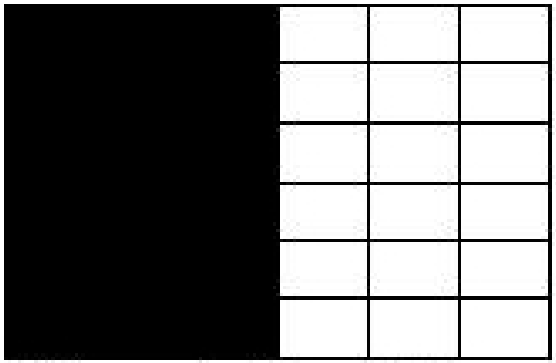
Est-ce que les objets géographiques qui sont proches se ressemblent plus que les objets géographiques qui sont éloignés ? C'est la question de l'autocorrélation spatiale.

Mesurer l'autocorrélation spatiale d'un phénomène (d'une distribution) revient à déterminer s'il semble exister une organisation spatiale sous-jacente à ce phénomène (à cette distribution) et donc qu'il (qu'elle) ne se repartit pas de façon aléatoire au sein du territoire étudié.

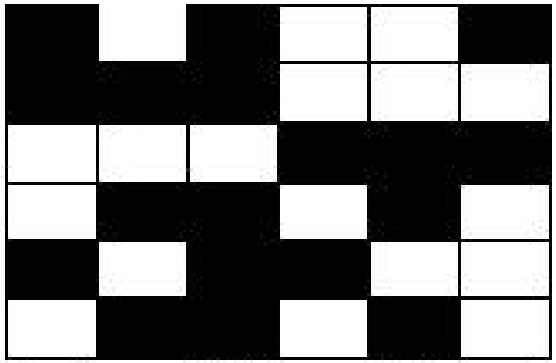
Par exemple : Les personnes riches se regroupent-elles ? Les communes très peuplées côtoient-elles des communes très peu peuplées ?

L'autocorrélation spatiale

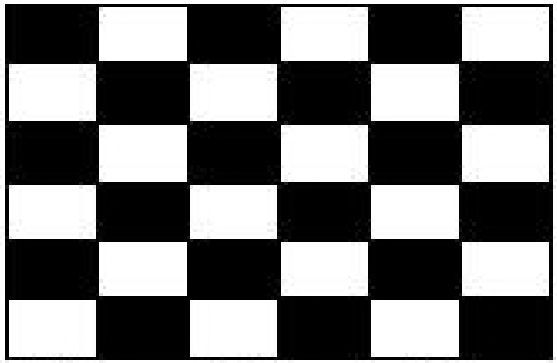
Présentation



Autocorrélation spatiale positive



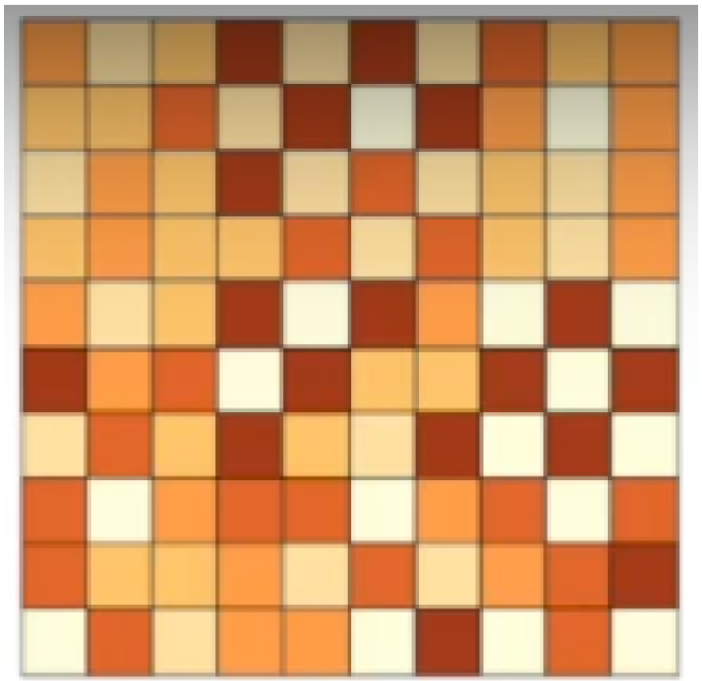
Autocorrélation spatiale nulle

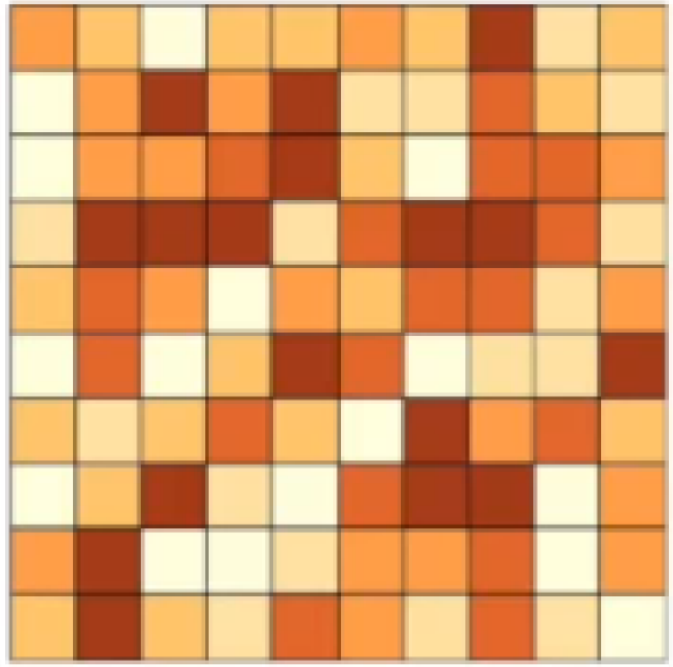


Autocorrélation spatiale négative

L'autocorrélation spatiale

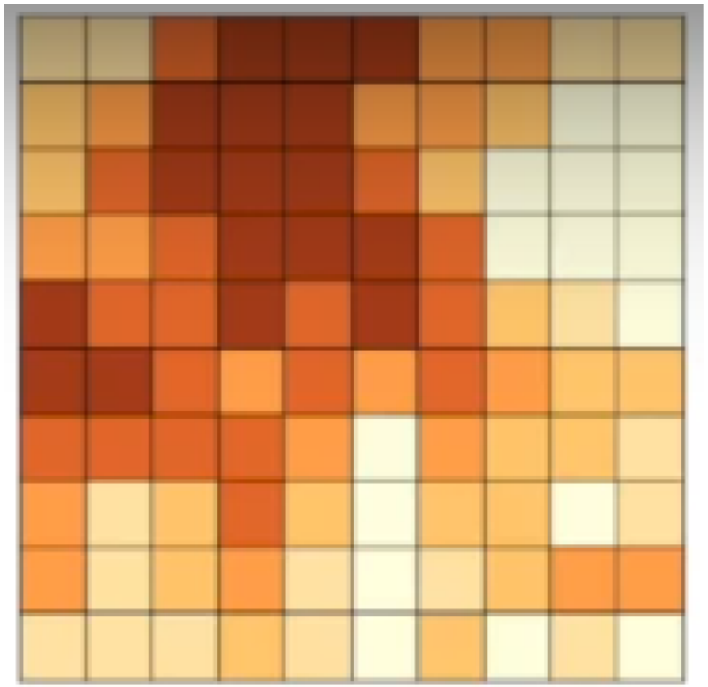
Présentation : aléatoire ou pas

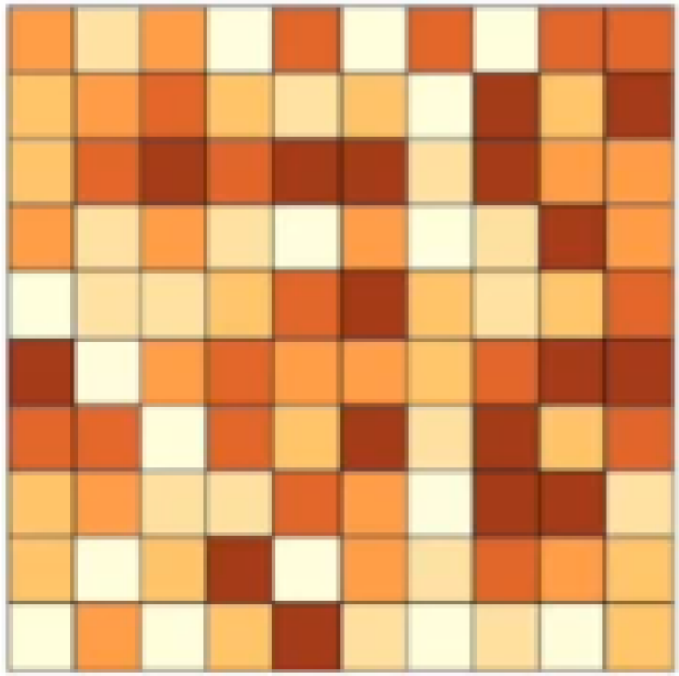




L'autocorrélation spatiale

Présentation : aléatoire ou pas





L'autocorrélation spatiale

Présentation statistique

Les coefficients d'autocorrélation spatiale sont alors construits statistiquement de telle manière qu'il soit possible de répondre à la question suivante :

La variation d'un caractère entre unités voisines (proches) est-elle plus ou moins grande que la variation de ce même caractère pour l'ensemble du territoire ou plus précisément entre unités non-voisines (éloignées) ?

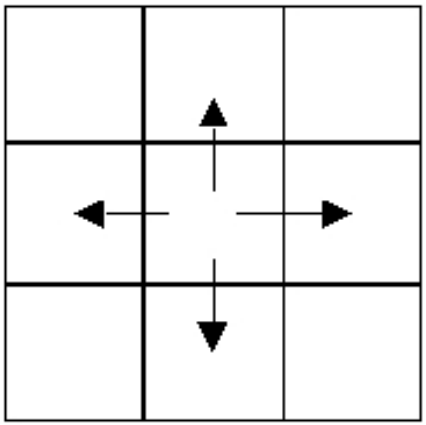
Il convient dès lors de définir ce qui est proche, de définir ce qui est voisin. Le plus simple est de le déterminer de manière binaire en s'appuyant par exemple sur la notion de contiguïté.

Il existe plusieurs indicateurs pour mesurer l'autocorrélation spatiale. Les deux principaux, c'est-à-dire les plus couramment utilisés, sont les indices de Moran et de Geary.

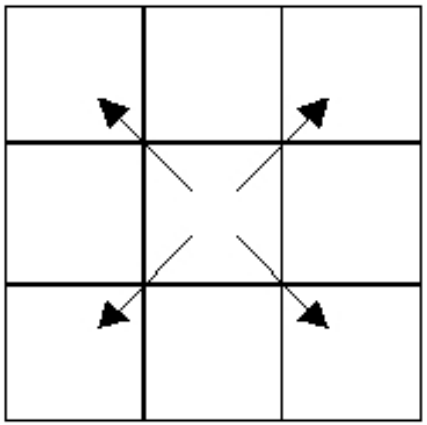
L'autocorrélation spatiale

La contiguïté

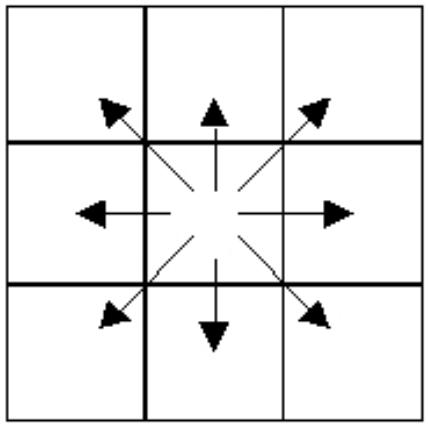
Rooks Case



Bishops Case



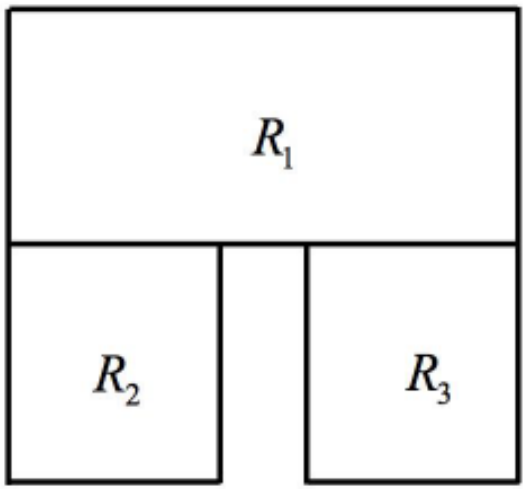
Queen's (Kings) Case

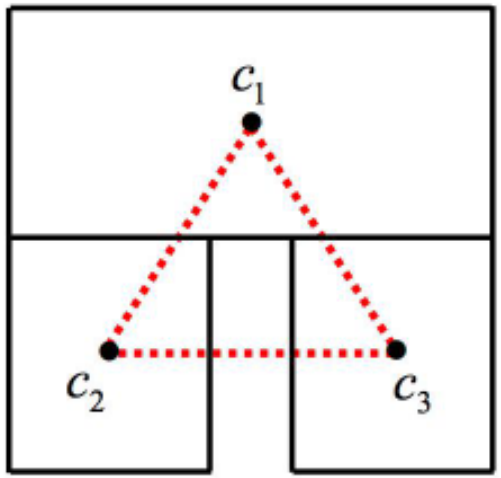


L'autocorrélation spatiale

La distance

Comme on travaille souvent avec des entités géographiques zonales, on passera souvent par les centroïdes pour évaluer la proximité, ici les trois entités sont alors équidistantes.

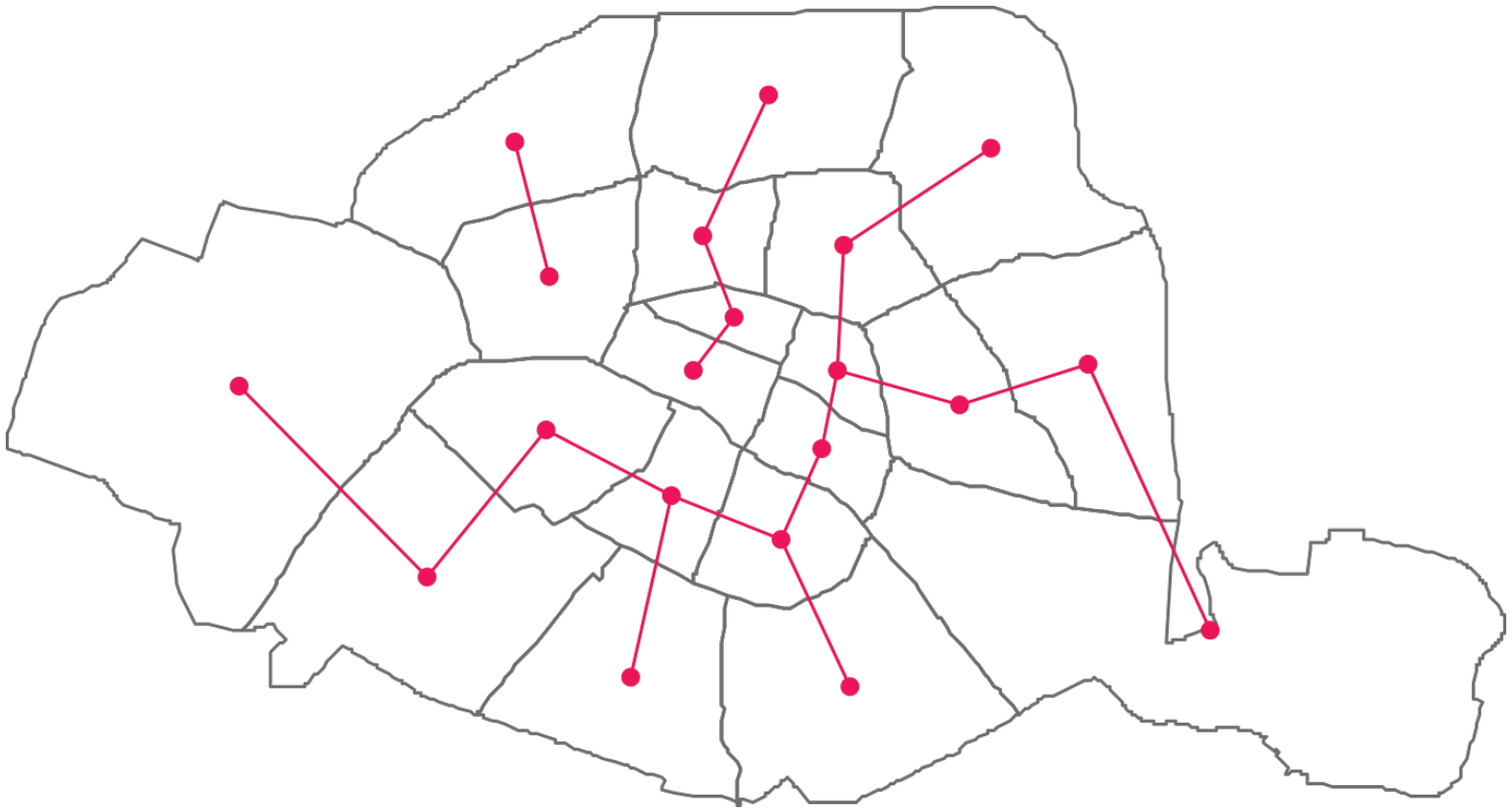




L'autocorrélation spatiale

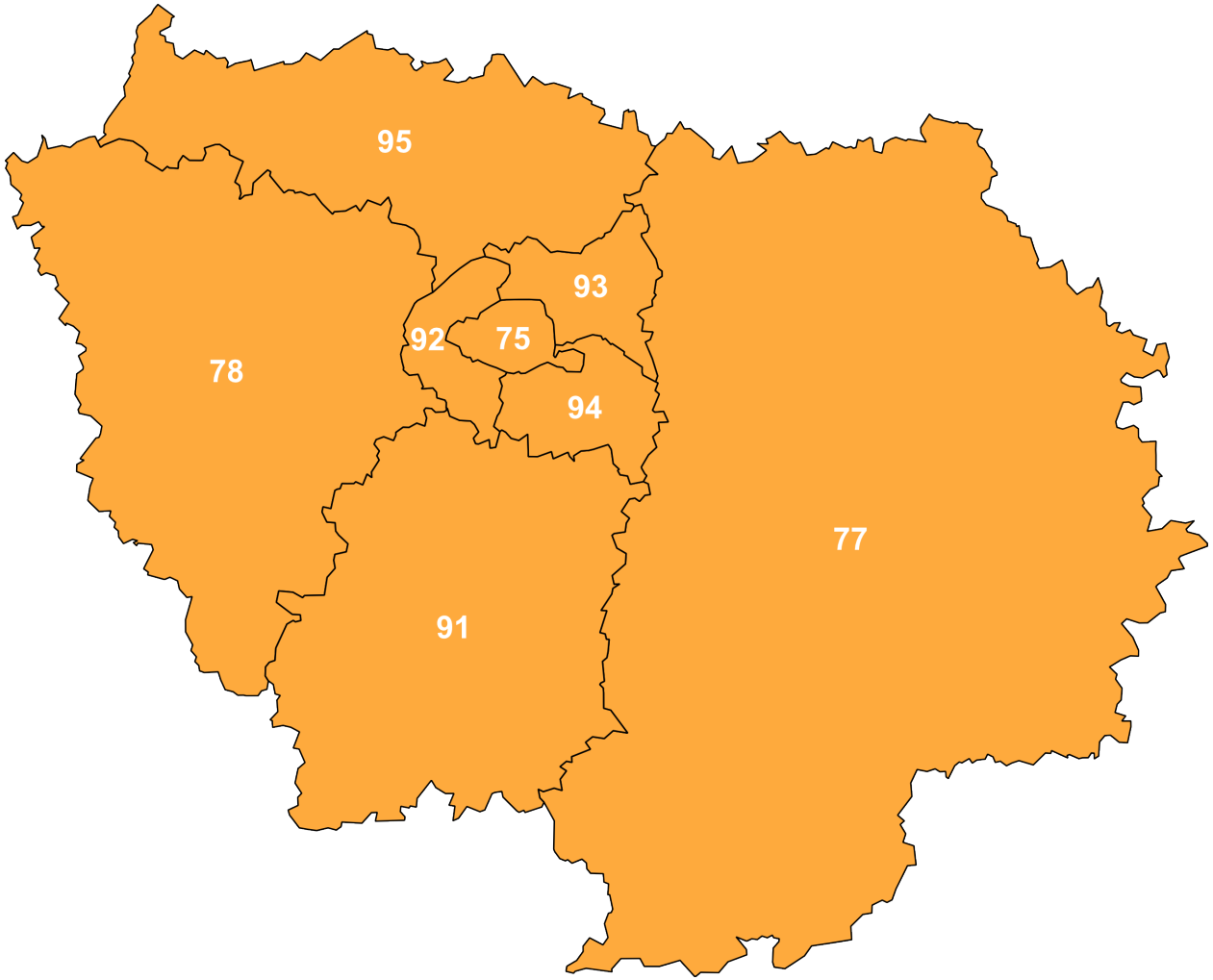
La distance

A partir de ces centroïdes, on peut alors créer des graphes de voisinage fondés sur la distance, par exemple au plus proche voisin.



L'autocorrélation spatiale

Contiguïté Vs Distance



L'autocorrélation spatiale

Les indices de Moran et Geary

$$G = \frac{N-1}{2L} \times \frac{\sum_{i,j} l_{ij} \times (X_i - X_j)^2}{\sum_i (X_i - \bar{X})^2}$$

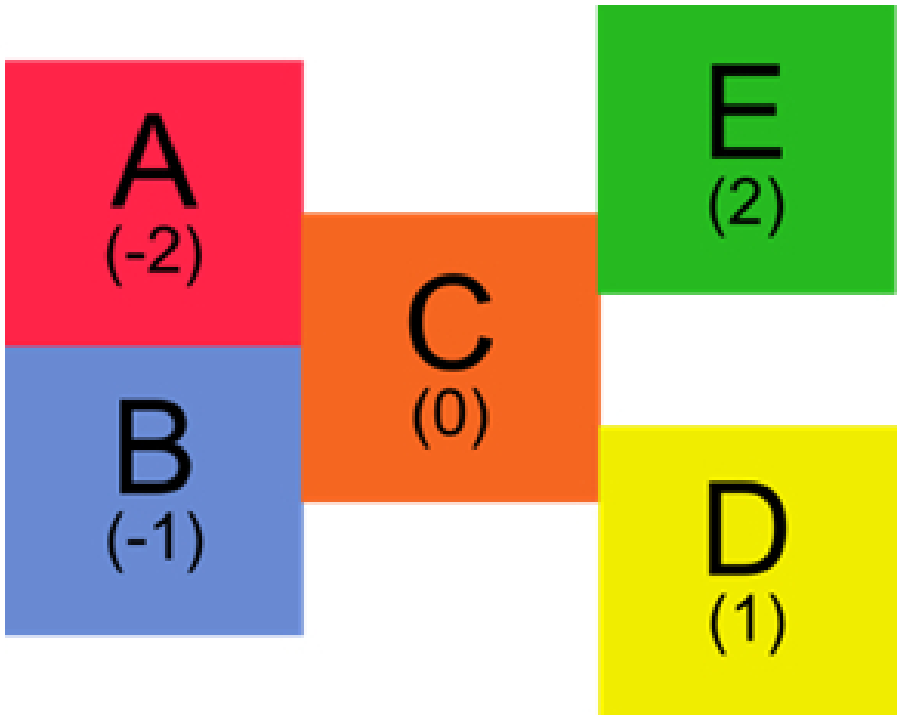
Les valeurs de l'indice de Geary s'étendent de 0 à 2. La valeur 1 signifie qu'aucune autocorrélation spatiale n'est présente dans les mesures effectuées. Une valeur plus petite que 1 signifie une autocorrélation spatiale positive.

$$I = \frac{N}{L} \times \frac{\sum_{i,j} l_{ij} \times (X_i - \bar{X}) \times (X_j - \bar{X})}{\sum_i (X_i - \bar{X})^2}$$

Les valeurs de l'indice de Moran s'étendent de -1 (corrélation négative) à +1 (corrélation positive). Une valeur nulle correspond à un modèle spatial parfaitement aléatoire.

L'autocorrélation spatiale

Exemple Geary



L'autocorrélation spatiale

Exemple Geary

$$\bar{X} = \frac{(-2-1+0+2+1)}{5} = 0$$

$$\sum_i (X_i - \bar{X})^2 = (-2)^2 + (-1)^2 + (0)^2 + (1)^2 + (2)^2 = 10$$

$$N = 5 \text{ et } L = 10$$

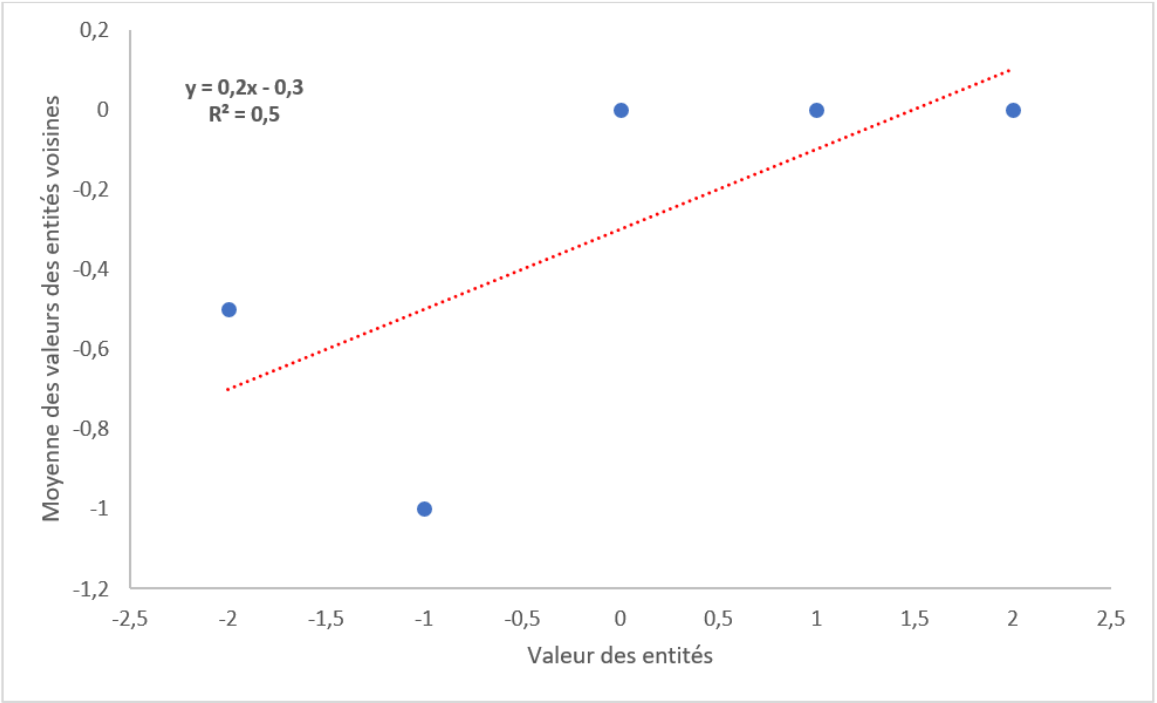
$A \rightarrow B$	$(-2-(-1))^2 = 1$	$A \rightarrow C$	$(-2-0)^2 = 4$	$B \rightarrow A$	$(-1-(-2))^2 = 1$
$B \rightarrow C$	$(-1-0)^2 = 1$	$C \rightarrow A$	$(0-(-2))^2 = 4$	$C \rightarrow B$	$(0-(-1))^2 = 1$
$C \rightarrow D$	$(0-1)^2 = 1$	$C \rightarrow E$	$(0-2)^2 = 4$	$D \rightarrow C$	$(1-0)^2 = 1$
$E \rightarrow$	$(2-0)^2 = 4$	**Total**		*22*	

$$G = \frac{(5-1)\times 22}{2\times 10\times 10} = 0,44$$

L'autocorrélation spatiale

Exemple Moran : Le diagramme de Moran

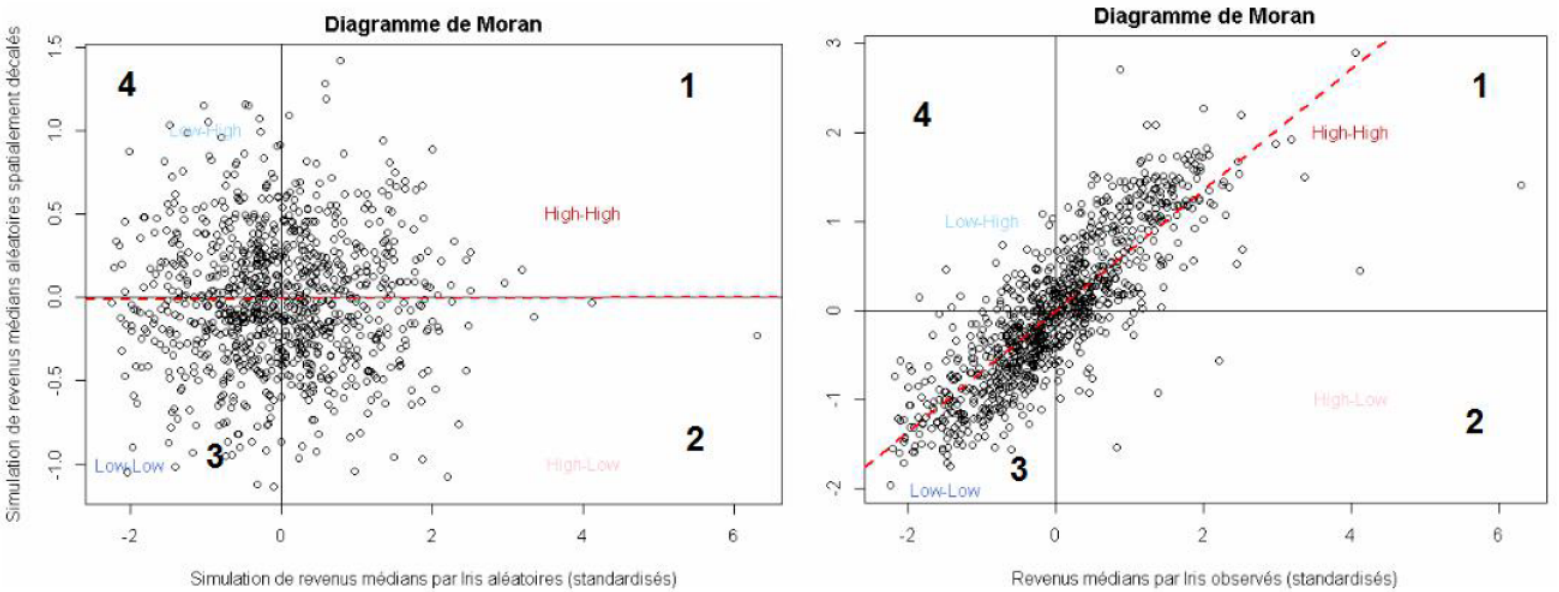
Derrière sa formulation mathématique qui peut paraitre peu intuitive, l'indice de Moran revient simplement à mesurer la pente de la relation linéaire entre les valeurs prises par les entités géographiques et leurs entités voisines ou proches (en prenant pour celles-ci une valeur moyenne). Le diagramme de Moran est alors fondamental.



L'autocorrélation spatiale

Exemple Moran : Le diagramme de Moran

On aura tendance à présenter le diagramme de Moran avec des valeurs standardisées et les moyennes correspondantes.

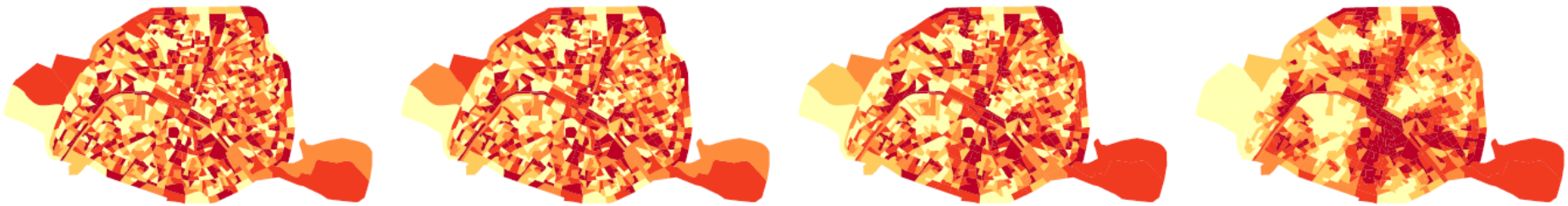


L'autocorrélation spatiale

Exemple Moran : Le diagramme de Moran

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

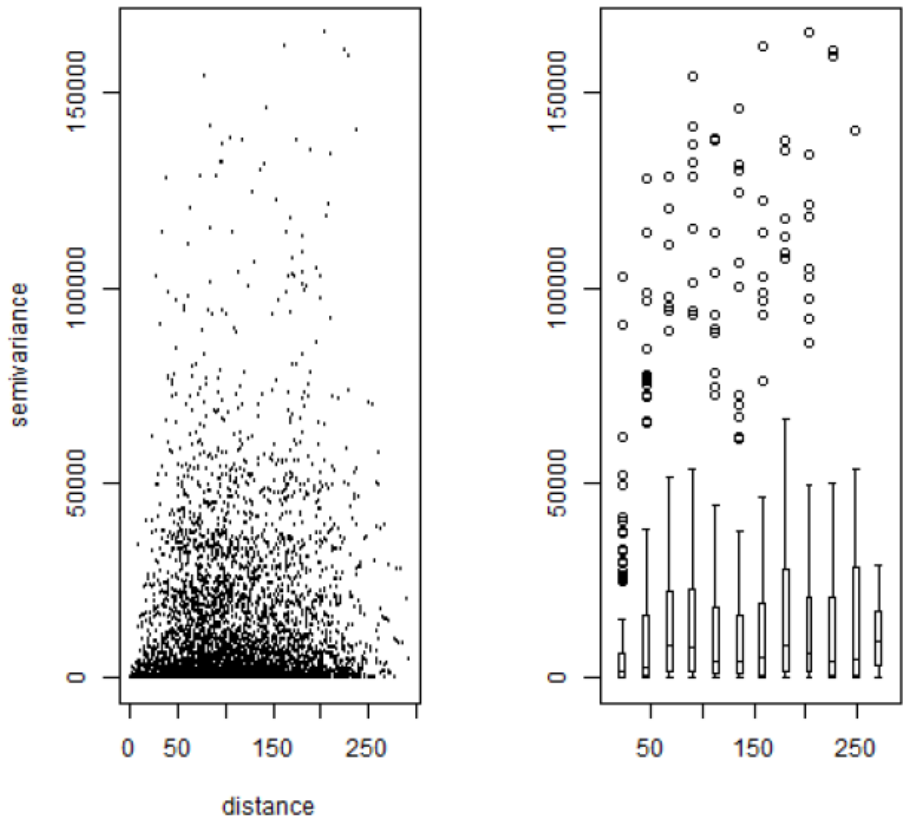
rho= 0.1 rho= 0.3 rho= 0.6 rho= 0.9



L'autocorrélation spatiale

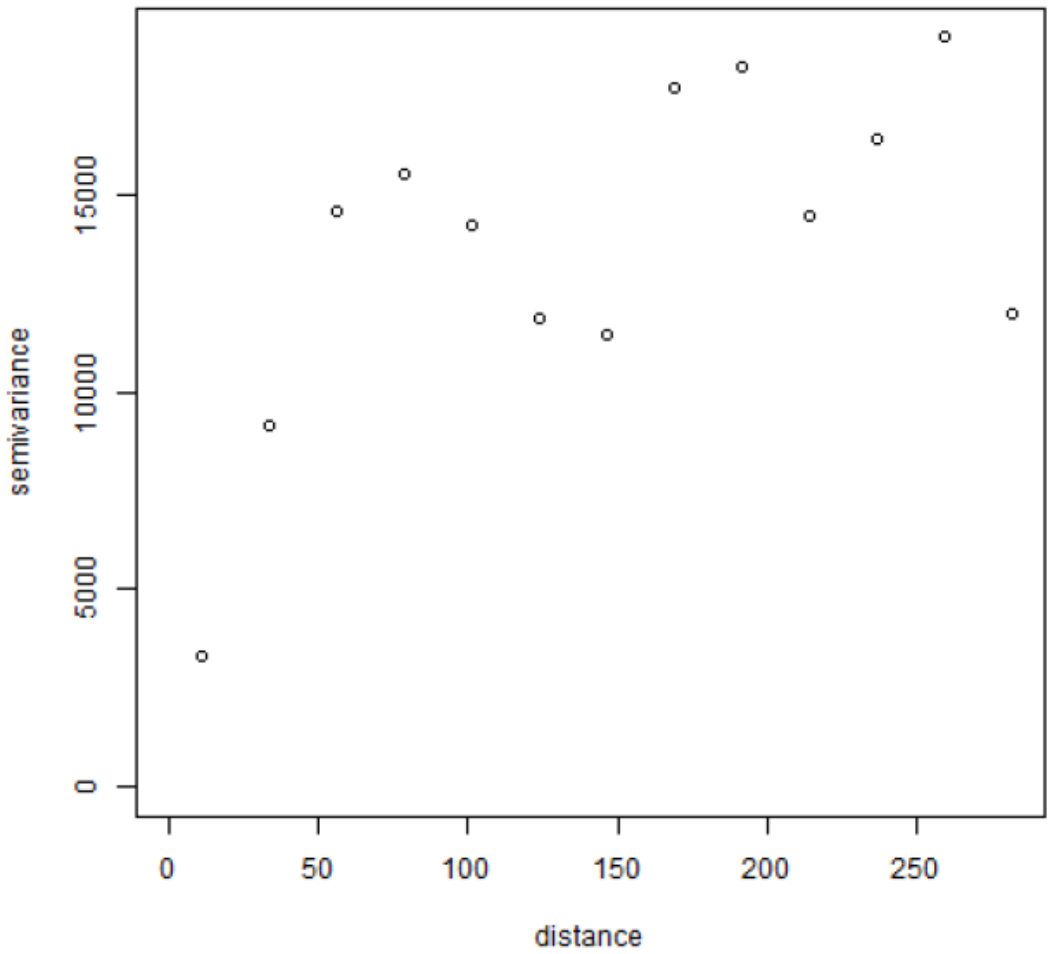
Une variante du diagramme de Moran : le semi-variogramme

Une manière a priori plus intuitive d'entrevoir l'autocorrélation spatiale est simplement d'étudier pour chaque paire de lieux la variation de la variable en fonction de la distance. On obtient alors une nuée variographique généralement peu lisible qu'il convient de résumer.



L'autocorrélation spatiale

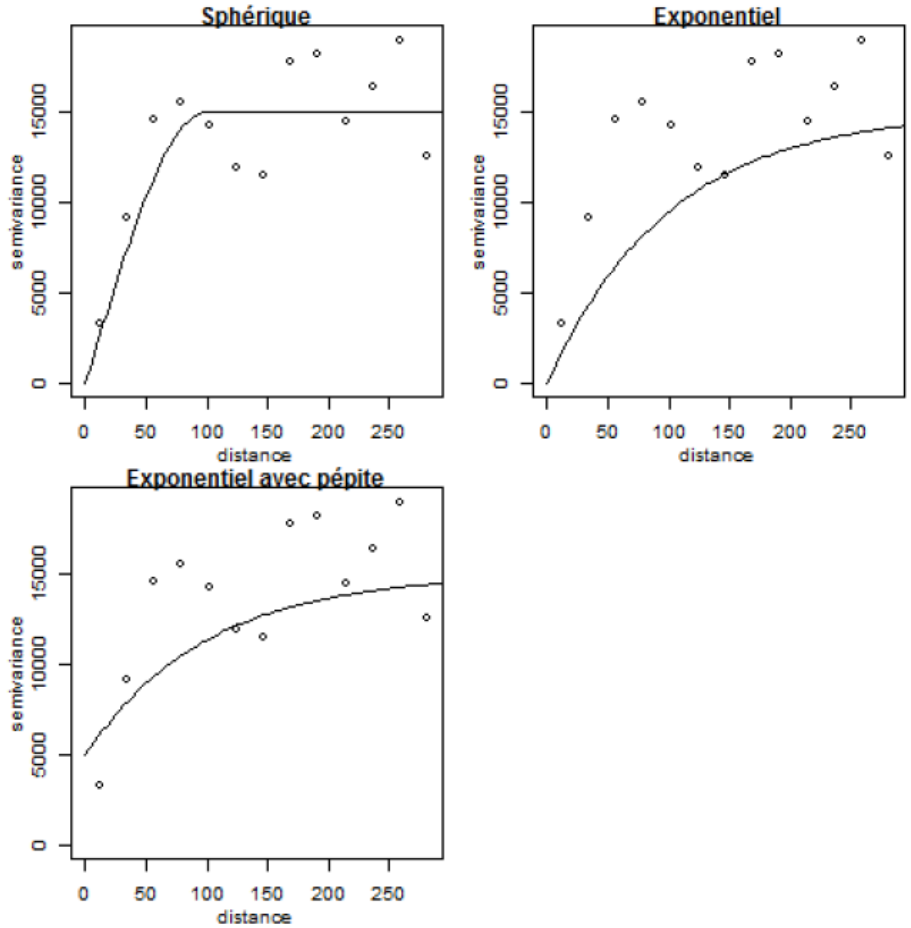
Une variante du diagramme de Moran : le semi-variogramme



Pour résumer cette nuée variographique, on retient la moyenne d'une classe de distance : c'est le semi-variogramme ou variogramme expérimental.

L'autocorrélation spatiale

Une variante du diagramme de Moran : le semi-variogramme



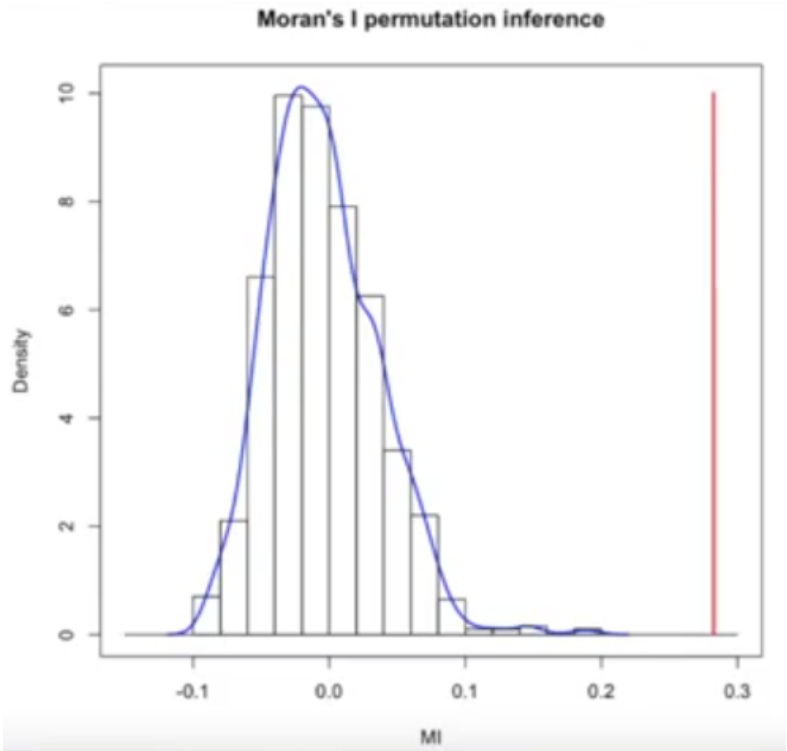
Pour résumer davantage ce variogramme expérimental, on peut chercher à l'approximer par une fonction mathématique qui tente de reproduire sa portée et l'effet de pépite : c'est le variogramme théorique.

L'autocorrélation spatiale

Aléatoire ou pas : la p-value

La p-value calcule la probabilité que la valeur d'autocorrélation soit obtenue par une distribution aléatoire de la variable étudiée.

Il existe deux manières de calculer la p-value : de manière analytique ou par simulation (des permutations).



999 permutations et une valeur de i = 0.289

L'autocorrélation spatiale

Aléatoire ou pas : la p-value

Attention, il ne faut pas confondre p-value et z-value.

La z-value est une valeur qui peut être comparée à d'autres z-value, contrairement au i de Moran qui ne sont pas comparables.

Elle est dépendante de la manière dont on calcule la p-value...

L'autocorrélation spatiale

La pondération spatiale

La mesure de l'auto-corrélation spatiale requiert une précaution essentielle, ne pas accorder trop d'importance à des entités géographiques qui auraient plus de voisins que les autres.

En effet, à partir des graphes de voisinage il est facile de produire une matrice de poids. Néanmoins, la somme de chaque ligne traduit alors l'importance accordée à chaque entité. Une matrice de poids devra toujours être normalisée.

	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

L'autocorrélation spatiale

Normalisation de la matrice de poids

Il existe alors plusieurs méthodes de normalisation :

- Normalisation en ligne (schéma de codage "W") : pour une zone, le poids accordé à chaque voisin est divisé par la somme des poids de ses voisins. Cette standardisation facilite l'interprétation de la matrice de poids, puisqu'in fine on obtient la moyenne de la variable x calculée sur tous les voisins de l'observation i.
- Normalisation globale (schéma de codage "C") : les poids sont standardisés de sorte que la somme de tous les poids soit égale au nombre total d'entités.
- Normalisation par stabilisation de la variance (schéma de codage "S") : elle permet de réduire l'hétérogénéité dans les poids liée aux différences de taille et de nombre de voisins entre les zones. En effet, la normalisation en ligne donne plus de poids aux observations situées en bordure de la zone d'étude, avec un faible nombre de voisins.

L'autocorrélation spatiale

L'autocorrélation spatiale locale : les LISA

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. Les données spatiales sont souvent caractérisées par de l'hétérogénéité spatiale.

Le diagramme de Moran montre même qu'il existe des endroits où l'autocorrélation aurait tendance à être négative alors même que la tendance globale est positive et inversement.

Deux questions se posent alors : Comment mesurer une valeur d'autocorrélation spatiale locale ? Ces valeurs sont-elles significatives ?

Pour chaque observation, ces indicateurs indiquent l'intensité du regroupement de valeurs similaires (ou de tendance opposée) autour de cette observation et la somme des indices locaux sur l'ensemble des observations doit être proportionnelle à l'indice global correspondant.

L'autocorrélation spatiale

L'autocorrélation spatiale locale : les LISA

Concrètement, il est possible pour chaque lieu de calculer uniquement le numérateur de l'indicateur de Moran pour le décliner localement.

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

$$I = \frac{N}{L} \times \frac{\sum_i \sum_j I_{ij} \times (X_i - \bar{X}) \times (X_j - \bar{X})}{\sum_i (X_i - \bar{X})^2}$$

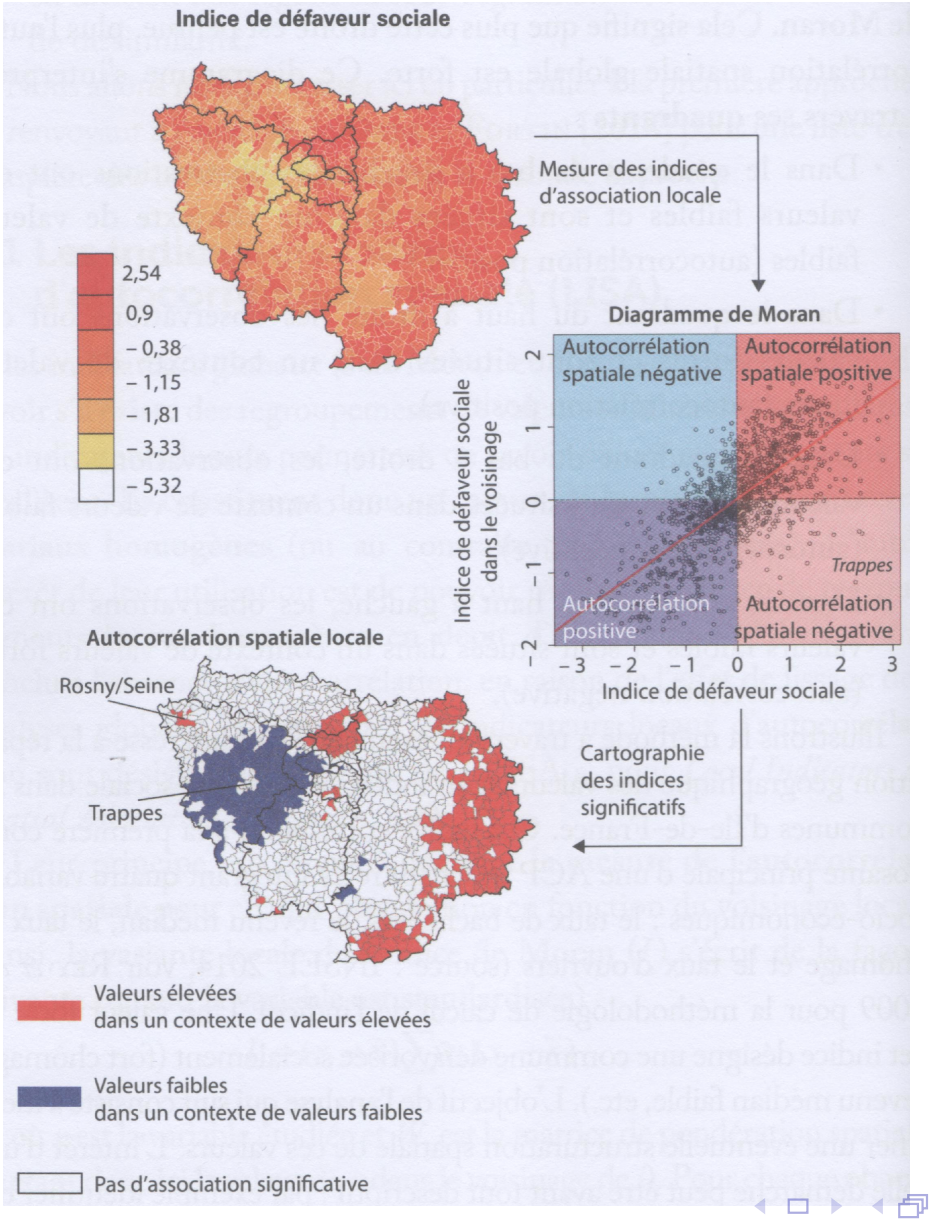
$$I_i = (X_i - \bar{X}) \sum_j I_{ij} \times (X_j - \bar{X})$$

Attention au calcul de la p-value

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. Différentes méthodes ont été développées pour éviter cette inflation du risque α.

L'autocorrélation spatiale

L'autocorrélation spatiale locale : les LISA



L'autocorrélation spatiale

Exemple sortie R Moran

```
        Moran I test under randomisation

data:  revenu
weights: irisqueenw

Moran I statistic standard deviate = 38.729, p-value < 2.2e-16
alternative hypothesis: greater
sample estimates:
Moran I statistic       Expectation          Variance
     0.7749098570     -0.0011587486      0.0004015296
```

L'autocorrélation spatiale

L'autocorrélation spatiale locale : Getis and Ord

L'indice de Getis-Ord est un autre moyen simple de détecter des clusters de valeurs fortes (hot spot) ou faibles (cold spot).

Pour cela, cet indice va simplement rapporter pour chaque observation, la valeur moyenne dans le voisinage à la somme des valeurs totales.

Une fois standardisée en un Z-score les valeurs négatives pourront être considérées comme des cold spot (notamment celles inférieures à 1,96) et les valeurs positives comme des hot spot (notamment celles supérieures à 1,96).

Cet indice présente l'inconvénient de ne pas détecter l'autocorrélation spatiale négative.

Cet indice est globalement moins robuste que les i locaux (LISA).

Régression spatiale, régression géographiquement pondérée

Les MCO remise en question par l'autocorrélation spatiale

L'autocorrélation spatiale resterait un simple outil de géographes si elle n'avait pas d'incidences pratiques. Or celle-ci remet justement en cause des méthodes couramment utilisées en économétrie ou pour faire de l'analyse territoriale. Elle remet notamment en question les régressions linéaires.

En effet la dépendance spatiale des observations peut se traduire soit par une perte d'efficacité des MCO, soit par des estimateurs biaisés.

Pour effectuer une régression linéaire certaines propriétés doivent être respectées : la normalité du terme d'erreur (distribution des erreurs) et l'indépendance des erreurs (graphique valeurs-résidus).

Or la dépendance spatiale des variables étudiées peut remettre en cause l'indépendance des erreurs. Il convient alors de tester l'autocorrélation spatiale des résidus. Si cette autocorrélation existe, la significativité des coefficients peut être remise en cause (elle est surestimée), car la redondance d'informations conduit à sous-estimer la variance calculée.

Régression spatiale, régression géographiquement pondérée

Les autres écueils des MCO

Pour traiter la question de la dépendance spatiale, on pourra utiliser les régressions spatiales. Néanmoins, la régression linéaire est aussi confrontée à trois autres spécificités des données en géographie :

- l'hétérogénéité spatiale (les phénomènes et les observations varient selon les lieux) ;
- les problèmes d'échelle et de zonage (les données peuvent être agrégées de différentes manières) ;
- et les effets de contexte (les observations dépendent des différents niveaux).

Pour étudier l'hétérogénéité spatiale, on pourra utiliser les régressions géographiquement pondérées. Pour les effets de contexte, il existe les régressions multiniveau.

Les problèmes d'échelle et de zonage (modifiable areal unit problem) restent ouverts, comme la prise en compte de toutes ces spécificités conjointement.

Régression spatiale, régression géographiquement pondérée

Les autres écueils des MCO

500	100	700
200	400	100
300	600	200

population

25	40	100
100	80	6
30	150	85

population pauvre

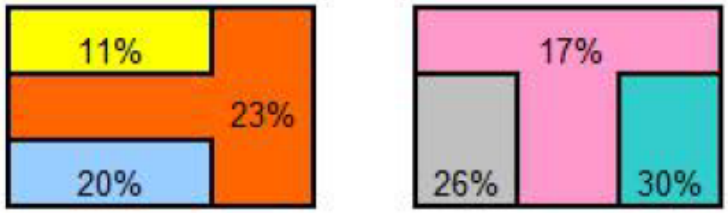
population et population pauvre dans 9 unités géographiques

5%	40%	14%
50%	20%	6%
10%	25%	43%

13%
27%
24%

taux de pauvreté avec 2 échelles différentes

MAUP : Effet d'échelle

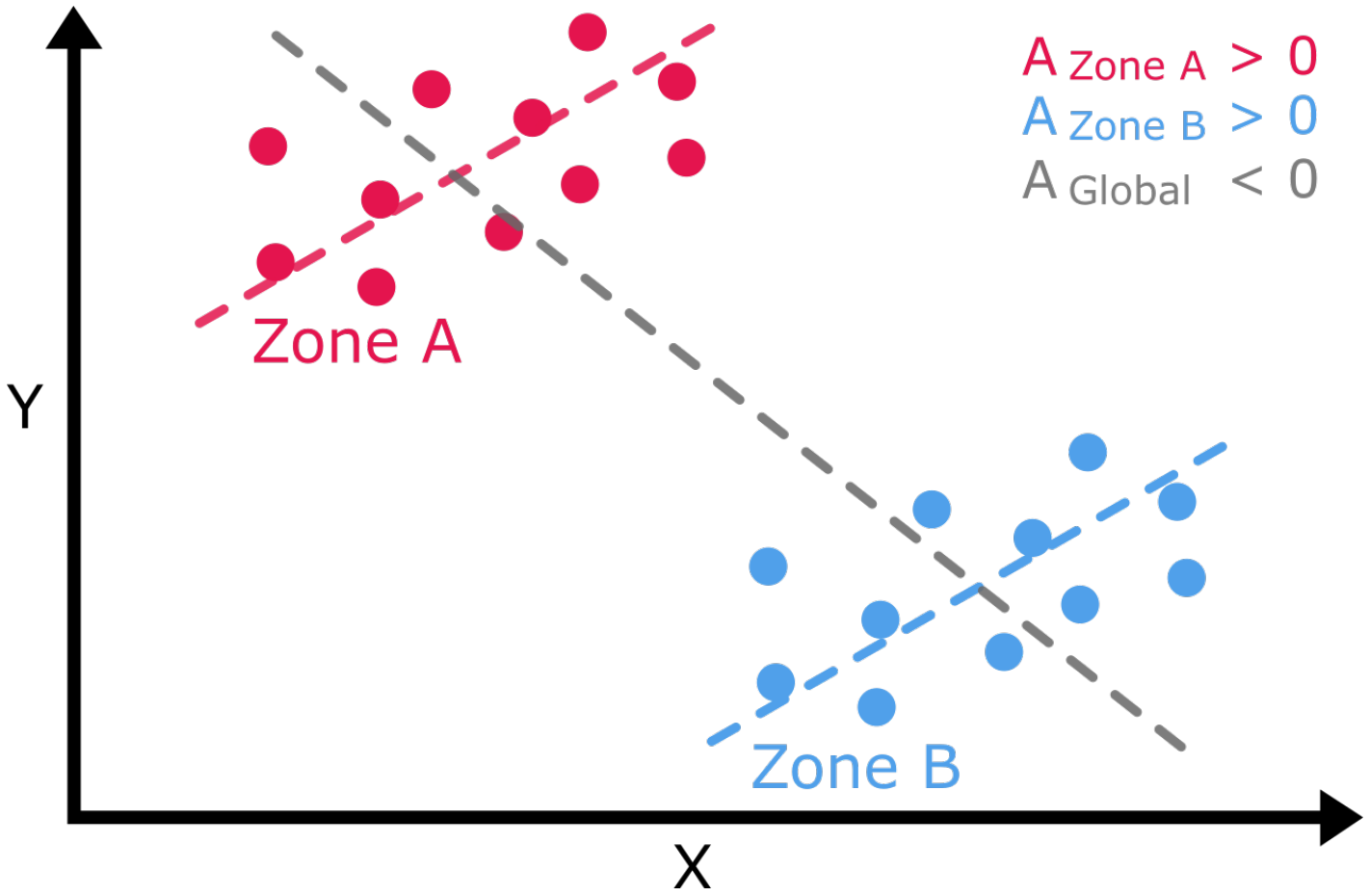


taux de pauvreté avec 2 zonages différents

MAUP : Effet de zonage

Régression spatiale, régression géographiquement pondérée

Les autres écueils des MCO



Régression spatiale, régression géographiquement pondérée

Régression spatiale

Il existe au moins autant de modèles de régression spatiale que de types d'interaction spatiale. On distingue :

- les interactions endogènes, lorsque la valeur d'une variable dans une zone géographique donnée dépend des valeurs de ses voisines (autocorélation spatiale, modèle autoregressif).
- les interactions exogènes, lorsque la variable étudiée dans une zone géographique dépend d'autres caractéristiques observables dans les zones géographiques voisines.
- une corrélation spatiale des erreurs liée à des caractéristiques inobservées, ignorées, négligées.

La prise en compte de ces interactions, complexifie énormément le modèle de régression, c'est le modèle de Mansky :

$$Y = aX + \epsilon$$

$$Y = aX + cW_{ij}X + dW_{ij}Y + eW_{ij}u + \epsilon$$

Régression spatiale, régression géographiquement pondérée

Régression spatiale

Comme il est impossible d'estimer tous les paramètres du modèle général sans qu'ils soient biaisés, puisqu'il est difficile d'identifier ce qui relève de tel ou tel type d'interaction (il y a des effets de pairs), il convient généralement de limiter celui-ci en considérant certains paramètres $= 0$.

Le plus intuitif est de considérer $e = 0$. C'est le modèle de référence, le modèle de Durbin (SDM). Une alternative à ce modèle consiste à négliger les interactions exogènes $c = 0$ pour éviter d'avoir trop de paramètres à calculer, c'est le modèle Spatial Autoregressive Confused (SAC), parfois appelé Kelejian-Prucha.

$$Y = aX + cW_{ij}X + dW_{ij}Y + \epsilon \text{ (SDM)}$$

$$Y = aX + dW_{ij}Y + eW_{ij}u + \epsilon \text{ (SAC)}$$

Régression spatiale, régression géographiquement pondérée

Régression spatiale

Néanmoins les modèles les plus simples à appréhender sont ceux qui ne tiennent compte que d'un seul type d'interaction. d $\neq$ 0 correspond au modèle spatial autoregressif (SAR), e $\neq$ 0 correspond au modèle à erreur autocorrélé spatialement (SEM).

$$Y = aX + dW_{ij}Y + \epsilon \text{ (SAR)}$$

$$Y = aX + eW_{ij}u + \epsilon \text{ (SEM)}$$

Enfin d'autres modèles existent correspondant à d'autres combinaisons de paramètres retenus.

L'interprétation des coefficients est plus complexe que lors d'une simple régression linéaire. Il est possible d'avoir recours à des méthodes pour en simplifier l'interprétation. On utilisera le critère d'information d'Akaike (AIC) pour estimer la précision du modèle (on cherche les valeurs les plus faibles) qui avantage les modèles parcimonieux.

Régression spatiale, régression géographiquement pondérée

Régression spatiale : exemple

Par exemple, une étude sur la ville de Columbus conclura à l'existence d'un lien significatif négatif entre le nombre de cambriolages et de vols de véhicule au sein de ses quartiers (CRIME) et des variables comme la valeur moyenne des logements des quartiers (HOVAL) et le revenu moyen des ménages (INC).

```
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 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
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: 9.341e-09
```

Régression spatiale, régression géographiquement pondérée

Régression spatiale : exemple

Néanmoins, une étude de l'autocorrélation spatiale des résidus remet en question la régression effectuée, la significativité pourrait être surestimée. Des modèles spatiaux doivent être privilégiés.

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 = 0.00734
alternative hypothesis: two.sided
sample estimates:

Observed Moran's I	Expectation	Variance
0.212374153	-0.033268284	0.008394853

Régression spatiale, régression géographiquement pondérée

Régression spatiale : exemple

L'application du modèle SAR, parfois appelé modèle de décalage spatial (LAG), permet de s'affranchir de l'autocorrélation spatiale des résidus. La significativité de Rho (d) confirme qu'il fallait utiliser un modèle spatial. La significativité des deux variables est confirmée.

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

Residuals:
        Min          1Q      Median          3Q         Max
-37.4497093  -5.4565567   0.0016387   6.7159553  24.7107978

Type: lag
Coefficients: (asymptotic standard errors)
             Estimate Std. Error z value  Pr(>|z|)
(Intercept) 46.851431   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
```

Régression spatiale, régression géographiquement pondérée

Régression spatiale : exemple

On peut comparer les résultats de différents modèles, ici ceux obtenus avec le modèle SEM qui se révèlent moins bons (voir AIC), mais confirment le modèle SAR.

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

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

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

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

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

Régression spatiale, régression géographiquement pondérée

Régression géographiquement pondérée

Les relations entre les variables étudiées peuvent varier dans l'espace. Or, les régressions présentées jusqu'à maintenant donnent une valeur globale pour l'ensemble du territoire étudié et il serait intéressant d'avoir des mesures de cette variation dans l'espace à l'image de l'autocorrélation spatiale locale.

Pour cela, la régression géographiquement pondérée propose de calculer un modèle de régression propre à chaque entité géographique en se fondant sur les valeurs voisines (proches) et non sur l'ensemble des données.

On aura autant de coefficients de corrélation et de régression que d'entités géographiques.

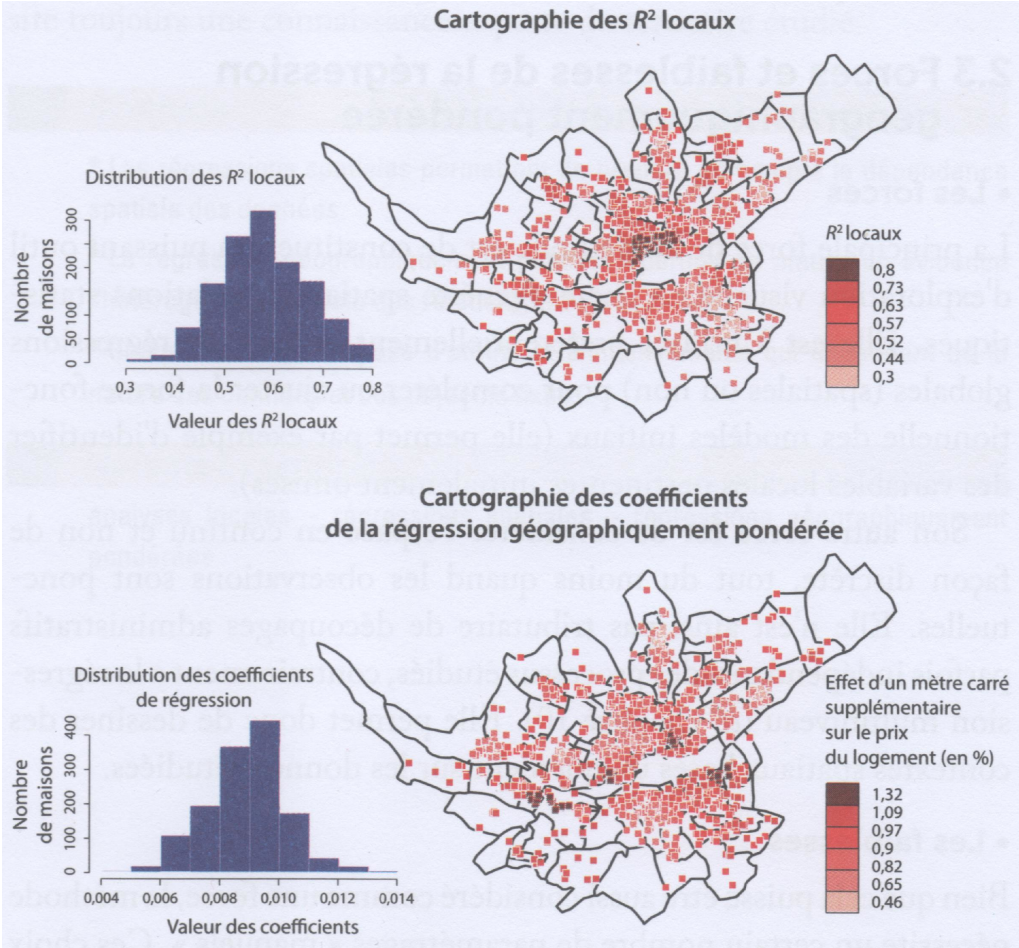
De nouveau, il faudra définir le voisinage, souvent par une fenêtre limite (une bande passante), et la matrice de poids correspondante : Gaussienne, Boxcar, Exponentielle, Bicarrée. On parle de noyau (kernel).

Le noyau peut être fixe, mais on préférera souvent un noyau adaptatif.

Régression spatiale, régression géographiquement pondérée

Régression géographiquement pondérée

Une régression classique conclura que le logarithme des prix des maisons augmente globalement de 0,96% par mètre carré à Nantes. En fait, cette valeur varie dans l'espace.



Analyse de semis de points

Présentation

Cette partie traite de la répartition d'ensemble de lieux qui correspondent aux différentes localisations d'un phénomène.

Ces lieux peuvent être des habitations, des commerces, des personnes, des clients...

Ces lieux peuvent être traités comme des points à un certain degré de généralisation. On parlera donc de semis de points.

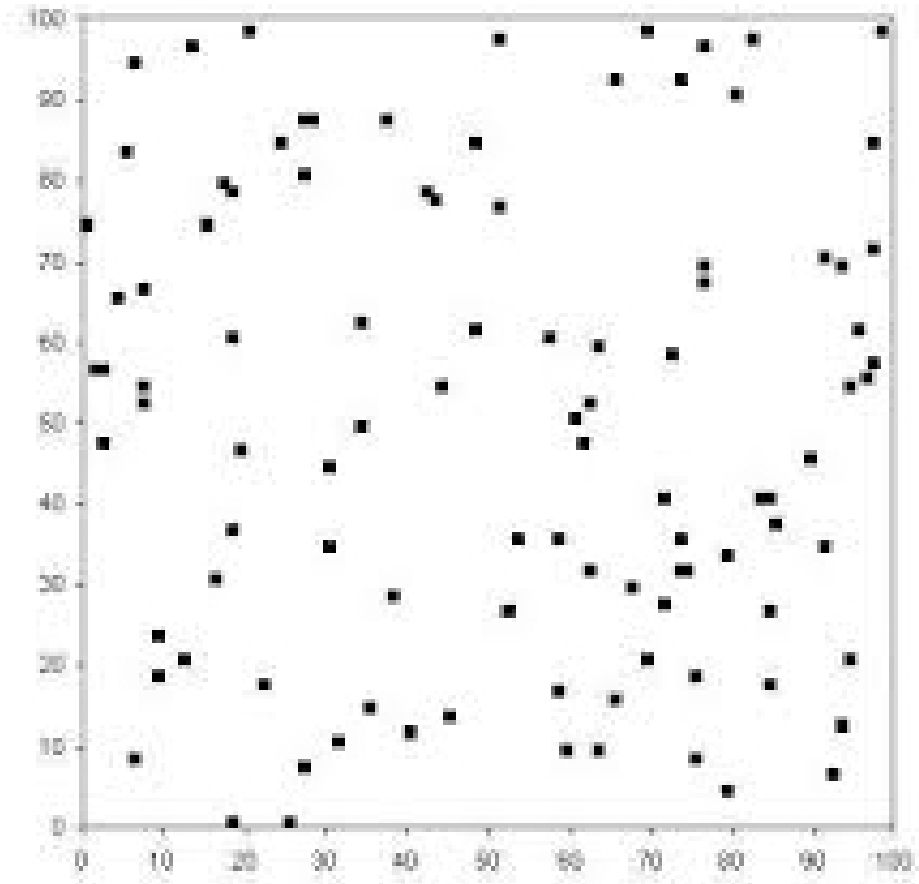
Pour comparer des semis de points ou pour mettre en exergue certaines de leurs spécificités, on va être amené à étudier leur forme.

Dans ces analyses, l'espace est souvent considéré comme homogène.

L'avantage de travailler avec des méthodes adaptées à des données ponctuelles (individuelles), c'est qu'elles ne sont en théorie pas sensibles aux problèmes d'échelle.

Analyse de semis de points

Présentation



Analyse de semis de points

Identifier le centre d'un semis de points : la position moyenne

Le point moyen :

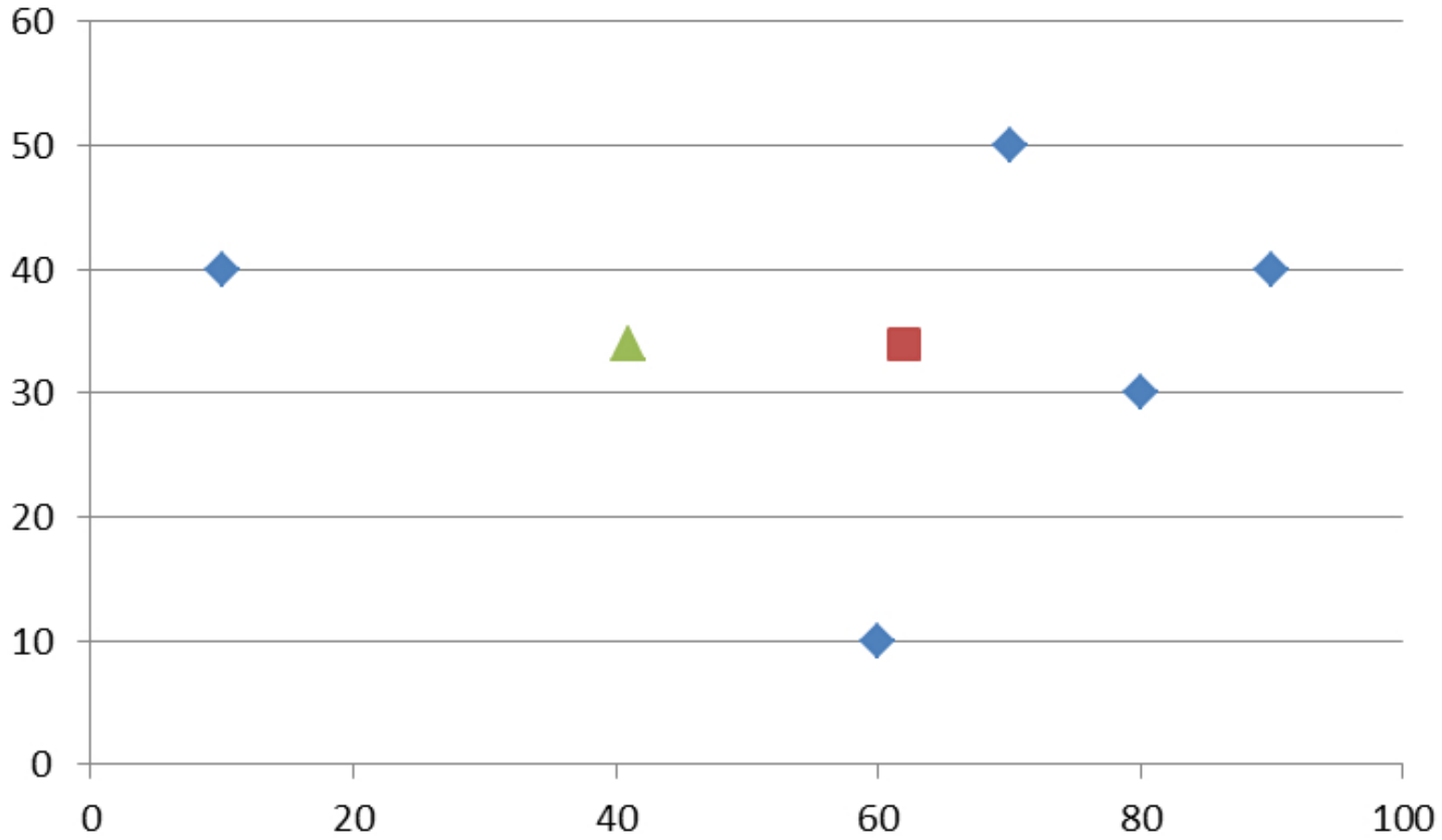
$$\bar{X} = \frac{1}{N} \times \sum_{i=1}^{N} X_i \text{ et } \bar{Y} = \frac{1}{N} \times \sum_{i=1}^{N} Y_i$$

Le point moyen pondéré :

$$\bar{X_p} = \frac{\sum_{i=1}^{N} (P_i \times X_i)}{\sum_{i=1}^{N} P_i} \text{ et } \bar{Y_p} = \frac{\sum_{i=1}^{N} (P_i \times Y_i)}{\sum_{i=1}^{N} P_i}$$

Analyse de semis de points

Identifier le centre d'un semis de points : la position moyenne



Analyse de semis de points

Mesurer la dispersion d'un semis de points

Ayant déterminé le point moyen, on peut chercher à mesurer la dispersion des lieux autour de ce point central. On parle de distance-type :

$$\sigma_D = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2 + (Y_i - \bar{Y})^2} = \sqrt{\sigma_X^2 + \sigma_Y^2}$$

Analyse de semis de points

Mesurer la concentration d'un semis de points

Une distribution est aléatoire si :

1. Tous les emplacements de l'espace ont la même probabilité d'accueillir un point.
2. La position d'un nouveau point est indépendante de la position des points précédents.

Une distribution aura tendance à être concentrée si :

1. Certains emplacements de l'espace ont plus de chances d'accueillir un point.
2. La localisation d'un premier point favorise l'apparition d'autres points à proximité.

Une distribution aura tendance à être régulière si :

1. Tous les emplacements de l'espace ont la même probabilité d'accueillir un point
2. La localisation d'un premier point défavorise l'apparition d'autres points à proximité.

Analyse de semis de points

Mesurer la concentration d'un semis de points

La méthode des quadrats permet de mesurer des concentrations (des densités) dans un semis de points :

1. Soit un semis de N points distribués sur un espace E.
2. On recouvre l'espace E d'un ensemble de K mailles d'une forme régulière (carré, rectangle, cercle).
3. Le nombre moyen de points théorique par maille est égale à D=N/K.
4. On associe à chaque maille i le nombre de points qu'elle contient, puis on calcule la variance du nombre de points par maille V(D) et on en déduit un indice de concentration(Ic). IC=V(D)/D.

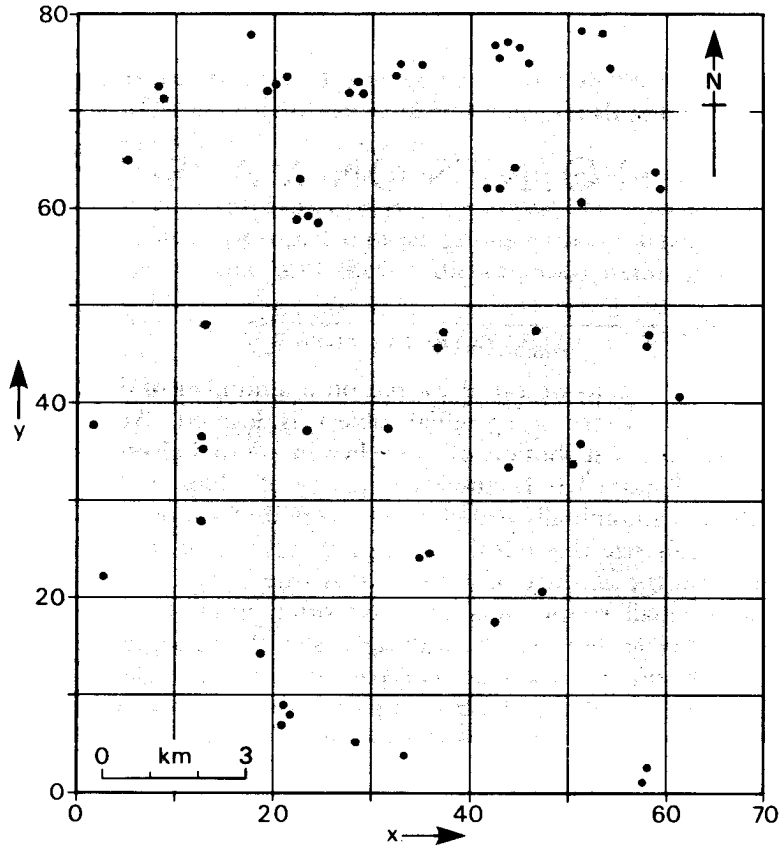
Si IC $= 1$, la distribution est aléatoire.

IC > 1, la distribution est plutôt concentrée.

IC < 1, la distribution est plutôt régulière.

Analyse de semis de points

Mesurer la concentration d'un semis de points



Attention

Cette méthode est sensible au MAUP !

Analyse de semis de points

Mesurer la concentration d'un semis de points

Nombre de points **n**	Nombre de quadrats **K**	Nombre de points **n.K**
0	25	0
1	15	15
2	8	16
3	5	15
4	1	4
5	2	10
Total	56	60

Analyse de semis de points

Mesurer la concentration d'un semis de points

Nombre de points **n**	Nombre de quadrats **K**	Nombre de points **n.K**	Ecarts à la moyenne **(n-D)**	**$K(n - D)^2$**
0	25	0	-1.071	28.676
1	15	15	-0.071	0.076
2	8	16	0.929	6.904
3	5	15	1.929	18.605
4	1	4	2.929	8.579
5	2	10	3.929	30.874
Total	56	60		93.714

Densité moyenne **D** = nb. de points / nb. de quadrats = 60/56 **= 1.071**
Variance **V(D)** = 93.714 / 55 = **1.704**
Indice de concentration IC = V(D)/D= **1.590**

Analyse de semis de points

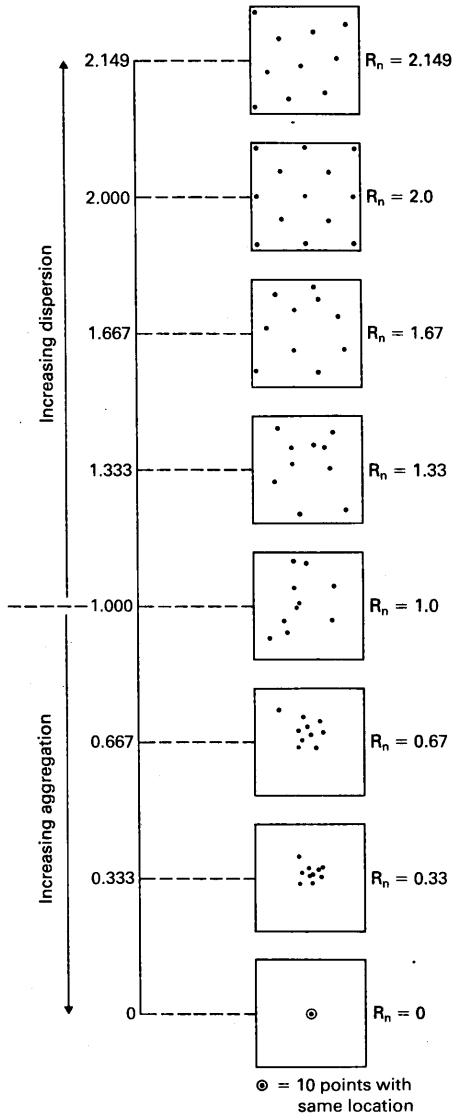
Mesurer la forme d'un semis de points

La méthode du plus proche voisin permet aussi d'étudier la dispersion, donc la forme d'un semis de points.

1. Soit un semis de N points distribués sur un espace de surface S. On note D la densité moyenne de points par unité de surface (D=N/S).
2. On calcule pour chaque point i la distance Dmin(i) qui le sépare de son voisin le plus proche.
3. On calcule ensuite la moyenne des distances observées au plus proche voisin D0.
4. On détermine la distance théorique moyenne au plus proche voisin DT dans le cas d'une distribution aléatoire ($DT=0.5/\sqrt{D}$).
5. On calcule l'indice de dispersion qui est le rapport : R=D0/DT.

Analyse de semis de points

Mesurer la forme d'un semis de points



Analyse de semis de points

Mesurer la forme d'un semis de points

i	Xi	Yi
1	1,5	7
2	1	7
3	1,5	6,8
4	0,5	5,8
5	2,2	7,5
6	0,3	7
7	0,6	4,8
8	1,8	4,1
9	2,1	5,2
10	4,3	5,8
11	1,6	7,2
12	3,1	6,4
13	0,7	2,9
14	0,1	2,6
15	1,5	4,4
16	3,1	5,3
17	5,2	6,2
18	5,1	7,9
19	1,7	1
20	2,4	1,8
21	4,2	5
22	7	6,1
23	6,8	3,8
24	7,2	0,3

Analyse de semis de points

Mesurer la forme d'un semis de points

i	dmin
1	0,2
2	0,5
3	0,2
4	1,0
5	0,7
6	0,7
7	1,0
8	0,4
9	1,0
10	0,8
11	0,2
12	1,1
13	0,7
14	0,7
15	0,4
16	1,0
17	1,0
18	1,7
19	1,1
20	1,1
21	0,8
22	1,8
23	2,3
24	3,5

Analyse de semis de points

Mesurer la forme d'un semis de points

$D0 = 0.99$

Comme la surface est égale 64 et l'effectif est égal à 24, on obtient une densité de 0,375 et par conséquent $DT = 0.816$

$R = D0/DT = 1{,}22$

Analyse de semis de points

Analyser la configuration d'un semis de points : la fonction K de Ripley

Les valeurs synthétiques précédentes peuvent masquer des phénomènes plus complexes.

C'est pourquoi on préfère généralement utiliser la fonction K de Ripley qui permet des analyses plus qualitatives, sans biais d'échelle et relativement exhaustives.

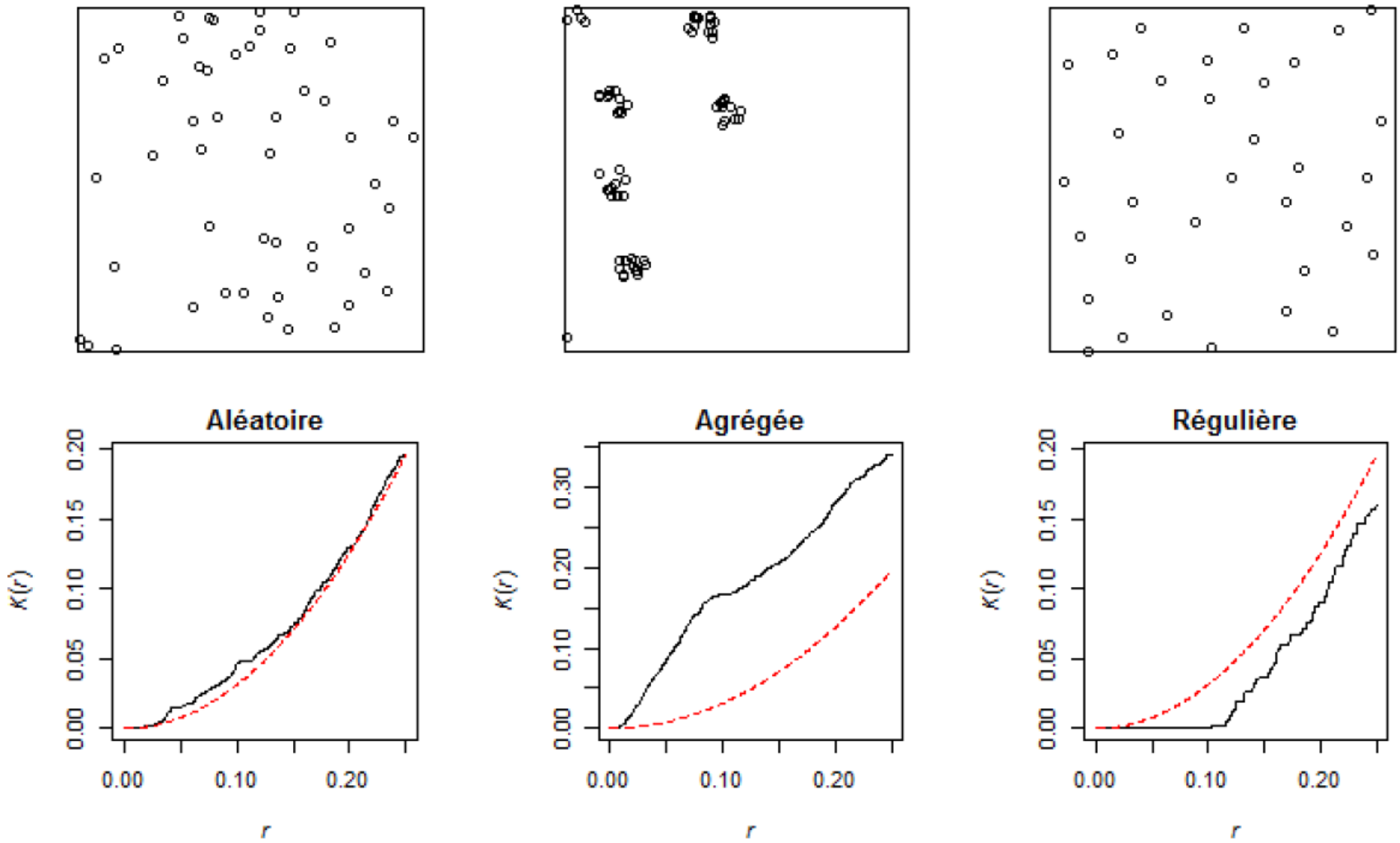
La fonction K(r) de Ripley est une fonction cumulative qui calcule simplement le nombre moyen de voisins de chaque point situés à une distance inférieure à r.

Ce nombre moyen est standardisé par l'intensité du processus (la densité n / W où W est l'aire étudiée).

Il convient ensuite de comparer cette fonction standardisée avec celle d'une distribution aléatoire (un processus de Poisson) homogène. Cette fonction est simplement égale à πr^2.

Analyse de semis de points

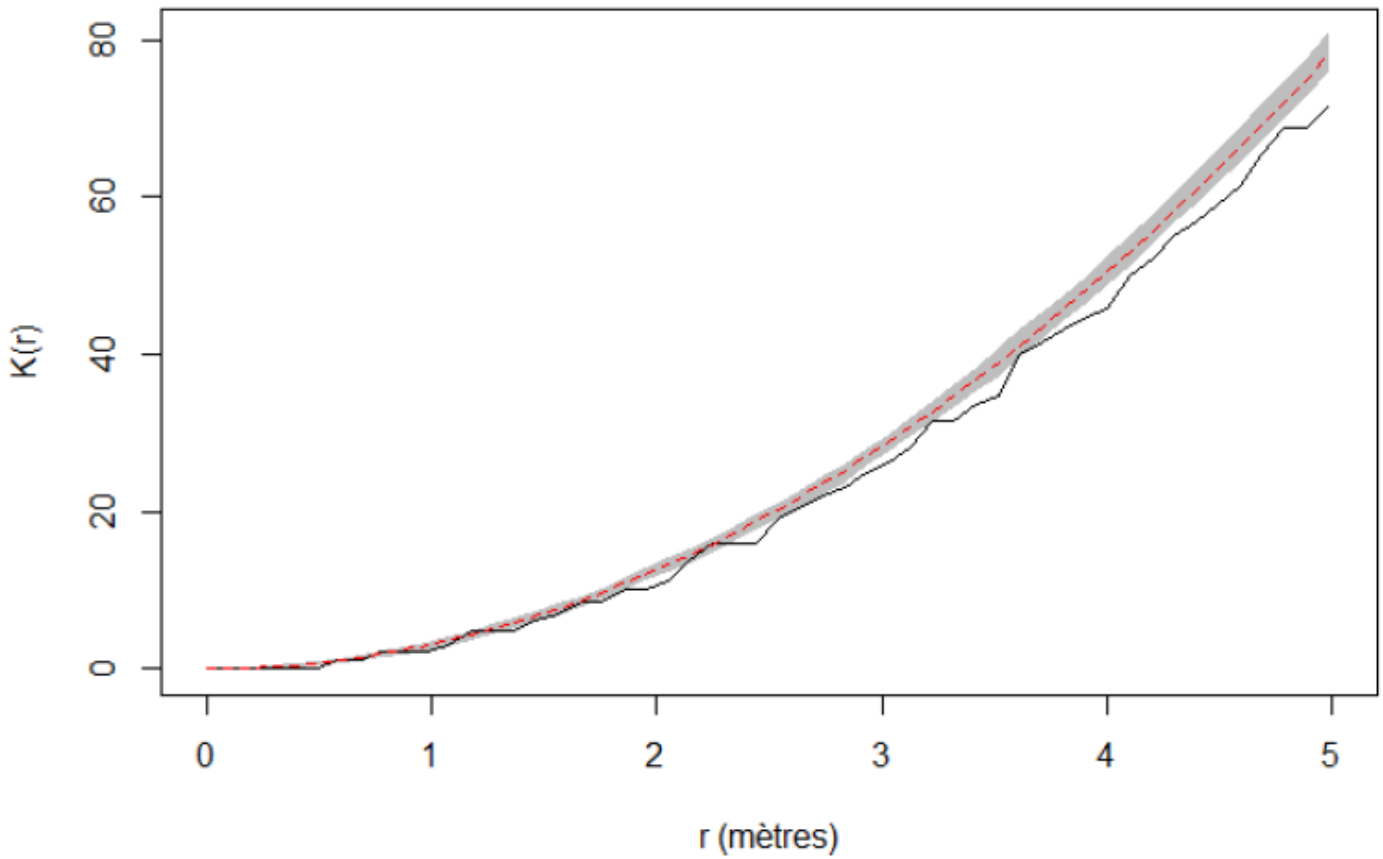
Analyser la configuration d'un semis de points : la fonction K de Ripley



Analyse de semis de points

Analyser la configuration d'un semis de points : la fonction K de Ripley

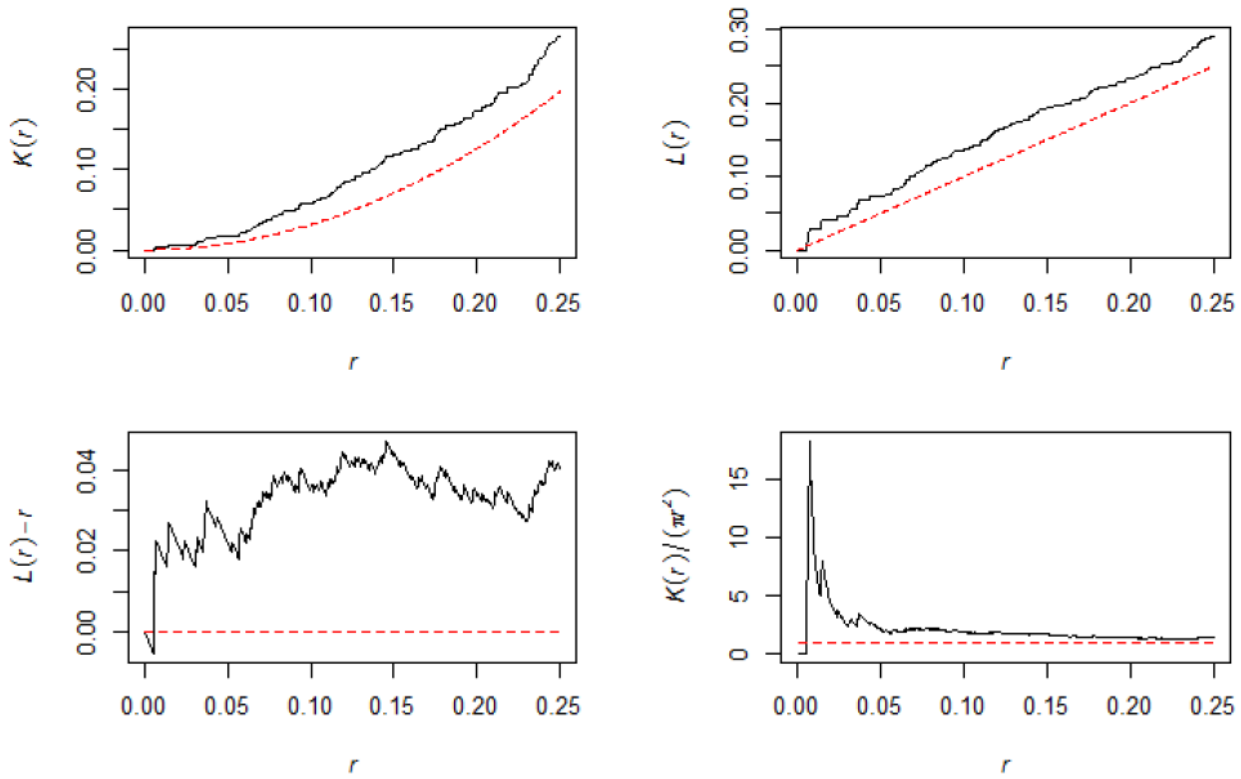
La technique la plus courante pour tester la significativité des valeurs obtenues est le recours à la simulation d'un intervalle de confiance par la méthode de Monte Carlo.



Analyse de semis de points

Analyser la configuration d'un semis de points : la fonction K de Ripley

Il existe de nombreuses variantes à la fonction de Ripley comme par exemple la fonction L de Besag (qui permet de comparer les différentes valeurs de la fonction en rapportant notamment K par π) et la fonction D de Diggle notamment (qui compare la fonction K à des processus non-homogènes).



Analyse de semis de points

M de Marcon et Puech et fonction intertype

L'indicateur M de Marcon et Puech est un indicateur cumulatif relatif qui va comparer la proportion de points d'intérêt dans un voisinage à celle que l'on observe sur l'ensemble du territoire analysé.

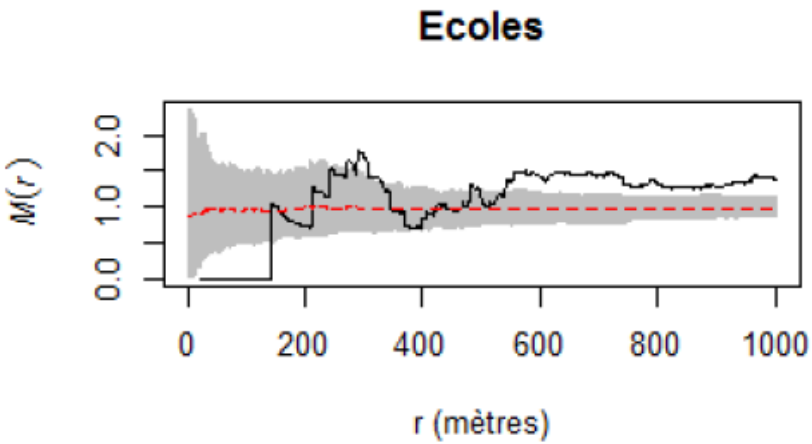
L'intérêt d'une approche relative, c'est qu'elle peut facilement être appliqué à deux types de configuration de points en comparant une proportion locale à une proportion globale mais où le type de points voisins d'intérêt n'est pas le même type que celui des points centre.

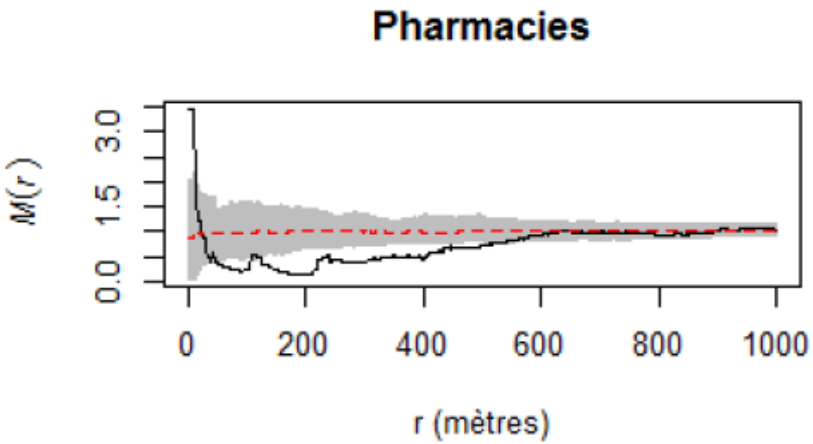
Par exemple, si nous suspectons une attraction des points de type T par ceux de type S, nous allons comparer la proportion locale de voisins du type T autour de points du type S à la proportion globale observée sur tout le territoire considéré.

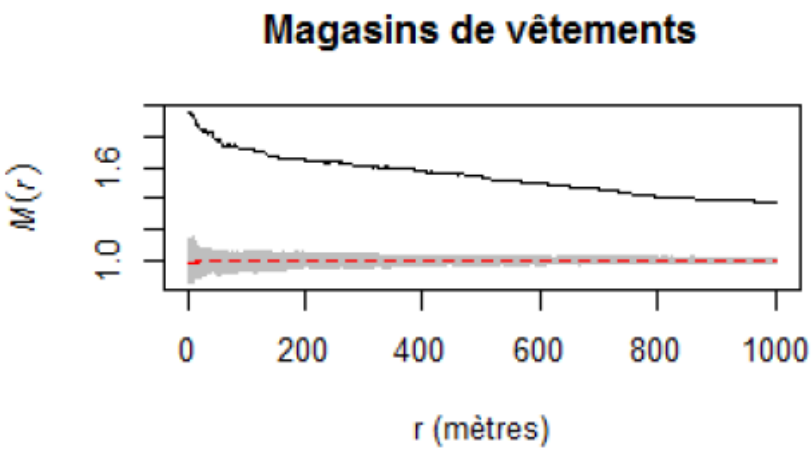
Si l'attraction entre les points de type T autour de type S est réelle, la proportion de points de type T autour de ceux du type S devrait être localement plus importante que celle observée sur toute l'aire d'étude.

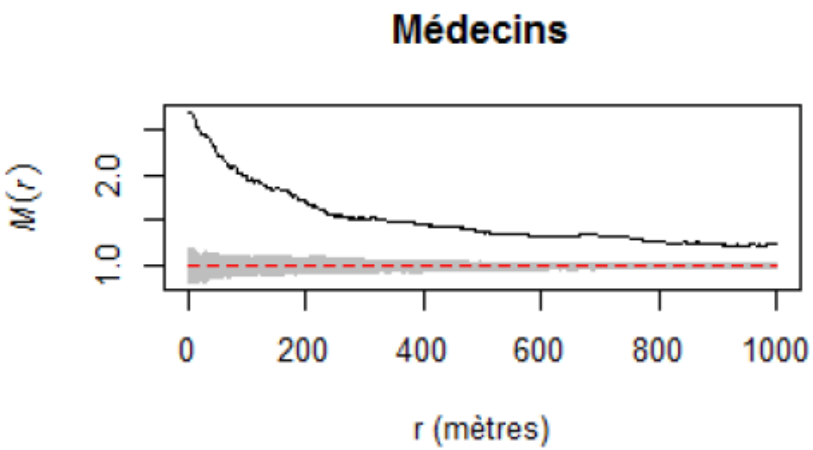
Analyse de semis de points

M de Marcon et Puech et fonction intertype



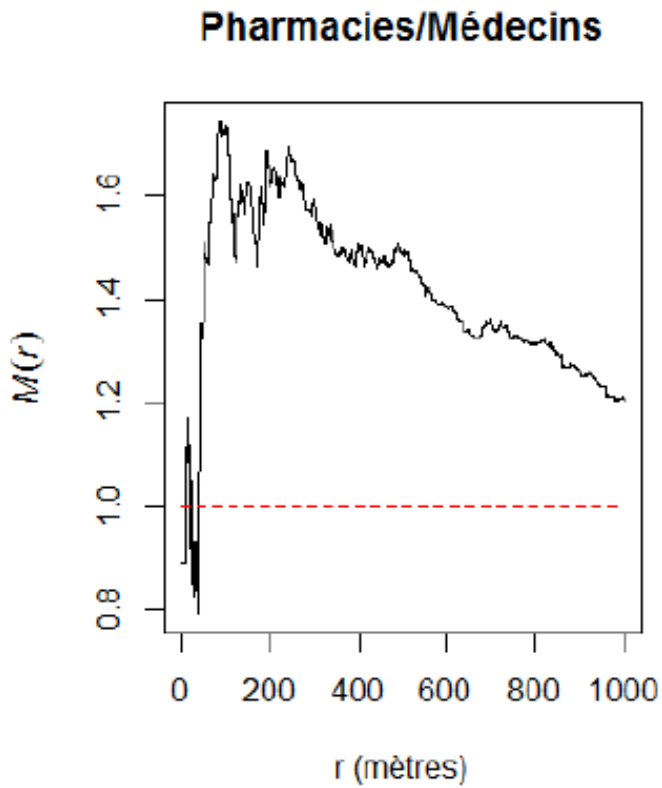


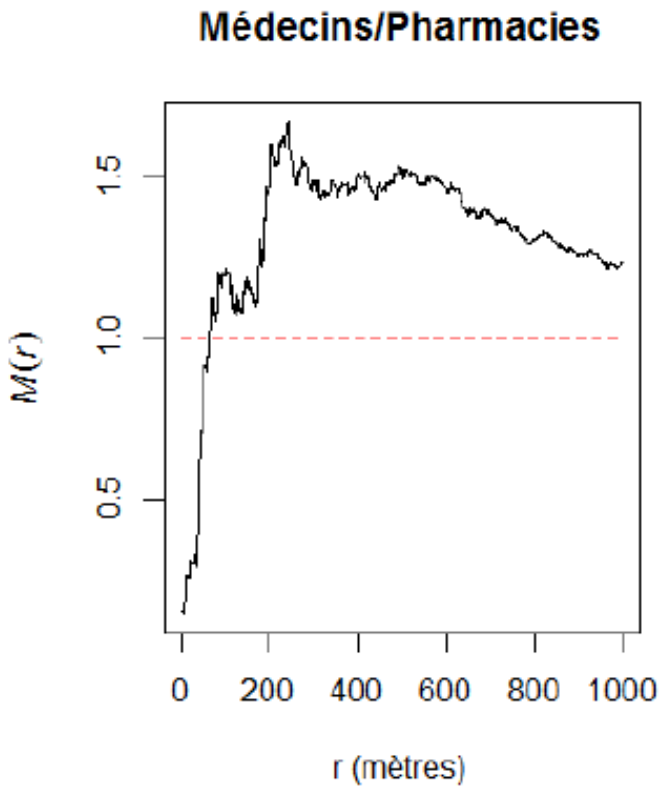




Analyse de semis de points

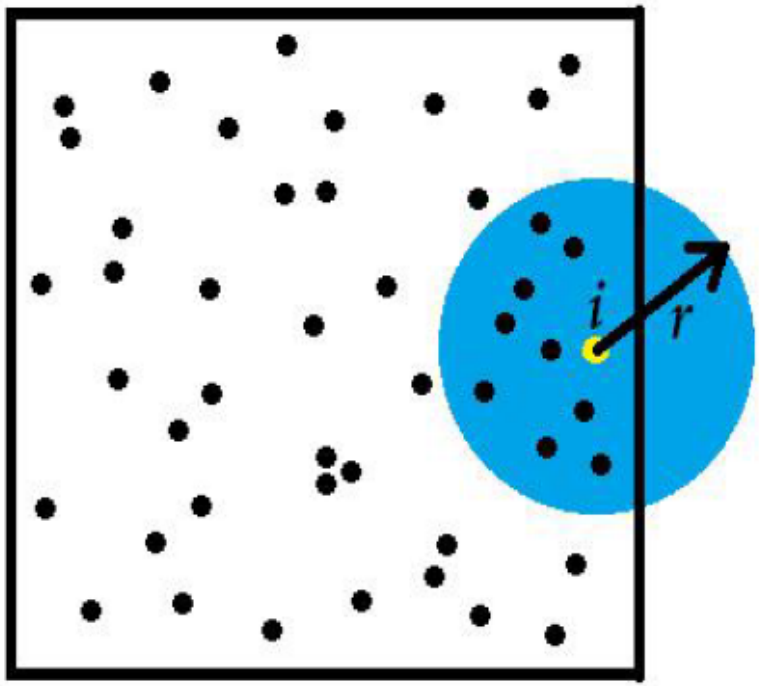
M de Marcon et Puech et fonction intertype





Analyse de semis de points

Attention aux effets de bord de la fenêtre d'observation



Généralement, quel que soit le domaine d'application, ce biais potentiel est jugé suffisamment sévère pour que l'on recoure à une technique correctrice prenant en compte les "effets de bord". On a pour cela souvent recours à du lissage spatial...

Interpolation spatiale et lissage spatial

Présentation

L'interpolation spatiale consiste à estimer des valeurs en différents points de l'espace à partir de valeurs connues en un nombre limité de points. Cette technique est essentiellement utilisée pour modéliser des phénomènes physiques continus (en climatologie, géophysique...).

Elle peut néanmoins être utilisée dans le domaine du géomarketing dans le but de compléter des données manquantes, voire pour simplifier un phénomène et ainsi dégager des tendances nettes. Ce dernier point la rapproche du lissage spatial.

Le lissage spatial consiste précisément à filtrer l'information pour révéler des structures spatiales sous-jacentes et régionaliser l'information. Mathématiquement et historiquement, c'est une fonction d'intensité d'un phénomène.

Combiner interpolation spatiale et lissage spatial permet souvent de s'affranchir des problèmes de maillage en ramenant les phénomènes à des grilles régulières d'une certaine résolution qui pourront être par la suite facilement combinées.

Interpolation spatiale et lissage spatial

Lissage spatial

Le principe du lissage spatial est de représenter non pas la valeur observée en un point, mais une moyenne pondérée des valeurs observées au voisinage de ce point dans un rayon prédéfini.

Une fois encore se posera les questions de bande passante, de fonction du noyau pour produire la matrice de poids. En effet, dans cette moyenne pondérée on tiendra davantage compte des lieux les plus proches.

Ici le choix du noyau importe peu, en revanche la taille de la bande passante est primordiale.

Un rayon élevé conduit à une densité très lissée, avec un biais élevé. Un petit rayon génère une densité peu lissée avec une forte variance.

Plusieurs méthodes proposent de calculer une bande passante "optimale" selon différents critères. L'objectif est généralement de minimiser une mesure d'erreur.

Interpolation spatiale et lissage spatial

Lissage spatial

Les limites du lissage spatial

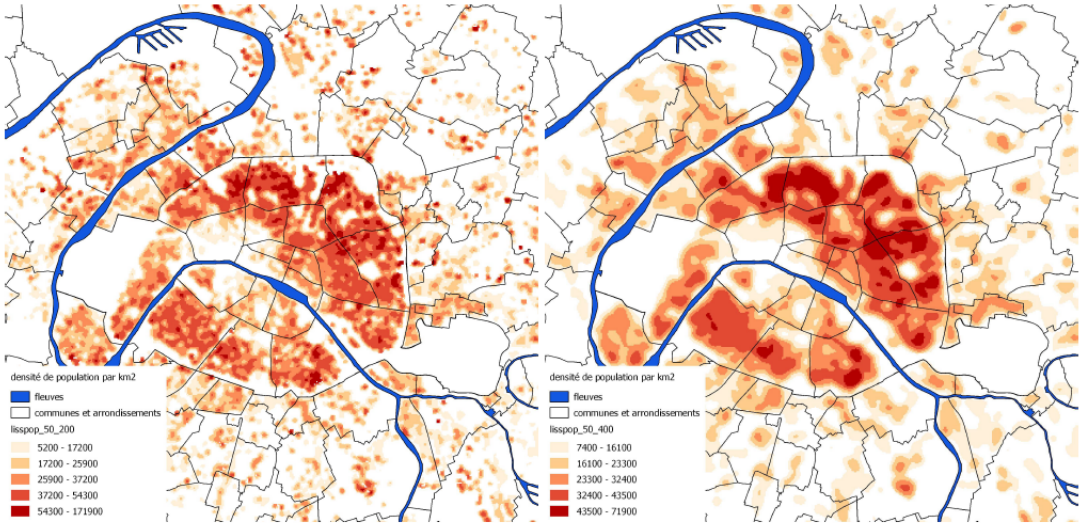
Derrière la qualité esthétique des cartes lissées se cache néanmoins un piège majeur. Par construction, les méthodes de lissage atténuent les ruptures et les frontières et induisent des représentations continues des phénomènes géographiques.

Les cartes lissées font donc apparaître localement de l'autocorrélation spatiale. Deux points proches par rapport au rayon de lissage ont mécaniquement des caractéristiques comparables dans ce type d'analyse.

De ce fait, commenter à partir d'une carte lissée des phénomènes géographiques dont l'ampleur spatiale est de l'ordre du rayon de lissage n'a guère de sens.

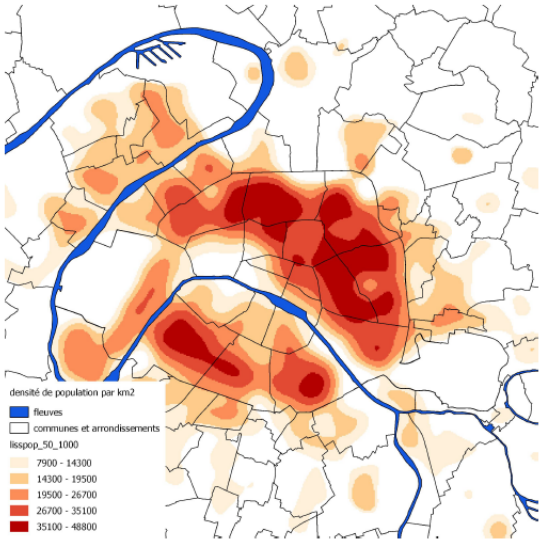
Interpolation spatiale et lissage spatial

Lissage spatial



(a) Rayon de 200 mètres

(b) Rayon de 400 mètres



(c) Rayon de 1000 mètres

Interpolation spatiale et lissage spatial

Interpolation spatiale : le krigeage

La frontière entre lissage spatial et interpolation spatiale est parfois ténue. Elles peuvent servir par exemple à produire ce que l'on appelle des cartes de chaleur, des cartes de densité (heatmap). La méthode de lissage présentée correspond à la méthode barycentrique en matière d'interpolation.

Néanmoins, en théorie, l'interpolation spatiale doit être réservée à des phénomènes continus et constitue la branche principale de la géostatitique. Le résultat d'une interpolation sera toujours une grille régulière (raster).

La méthode phare de l'interpolation spatiale est le krigeage. Le terme de krigeage est dû à Georges Matheron, et fait référence aux travaux pionniers de Danie Krige, ingénieur sud-africain.

Le krigeage réalise l'interpolation spatiale d'une variable régionalisée par calcul de l'espérance mathématique d'une variable aléatoire, utilisant l'interprétation et la modélisation d'un variogramme expérimental.

Interpolation spatiale et lissage spatial

Interpolation spatiale : le krigeage

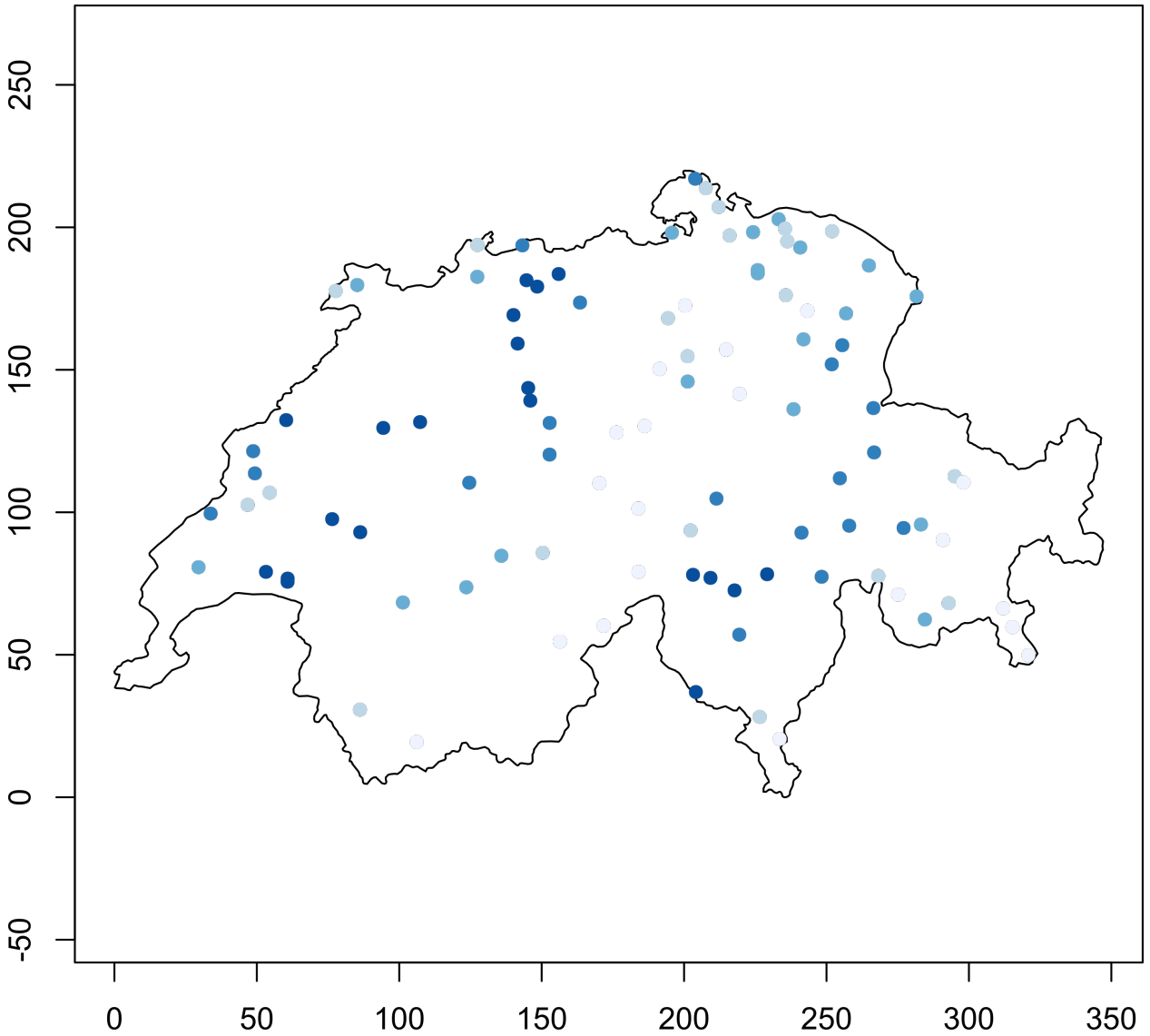
C'est le meilleur estimateur linéaire non-biaisé. Il tient compte non seulement de la distance entre les données et le point d'estimation, mais également des distances entre les données deux-à-deux.

L'idée de base du krigeage est de prévoir la valeur de la variable régionalisée étudiée en un site non échantillonné s0 par une combinaison linéaire de données ponctuelles adjacentes.

Le modèle de base du krigeage a la même forme que les modèles de régression classique ou locale, mais les erreurs sont supposées dépendantes spatialement. Le modèle requiert donc connaitre la dépendance spatiale du phénomène.

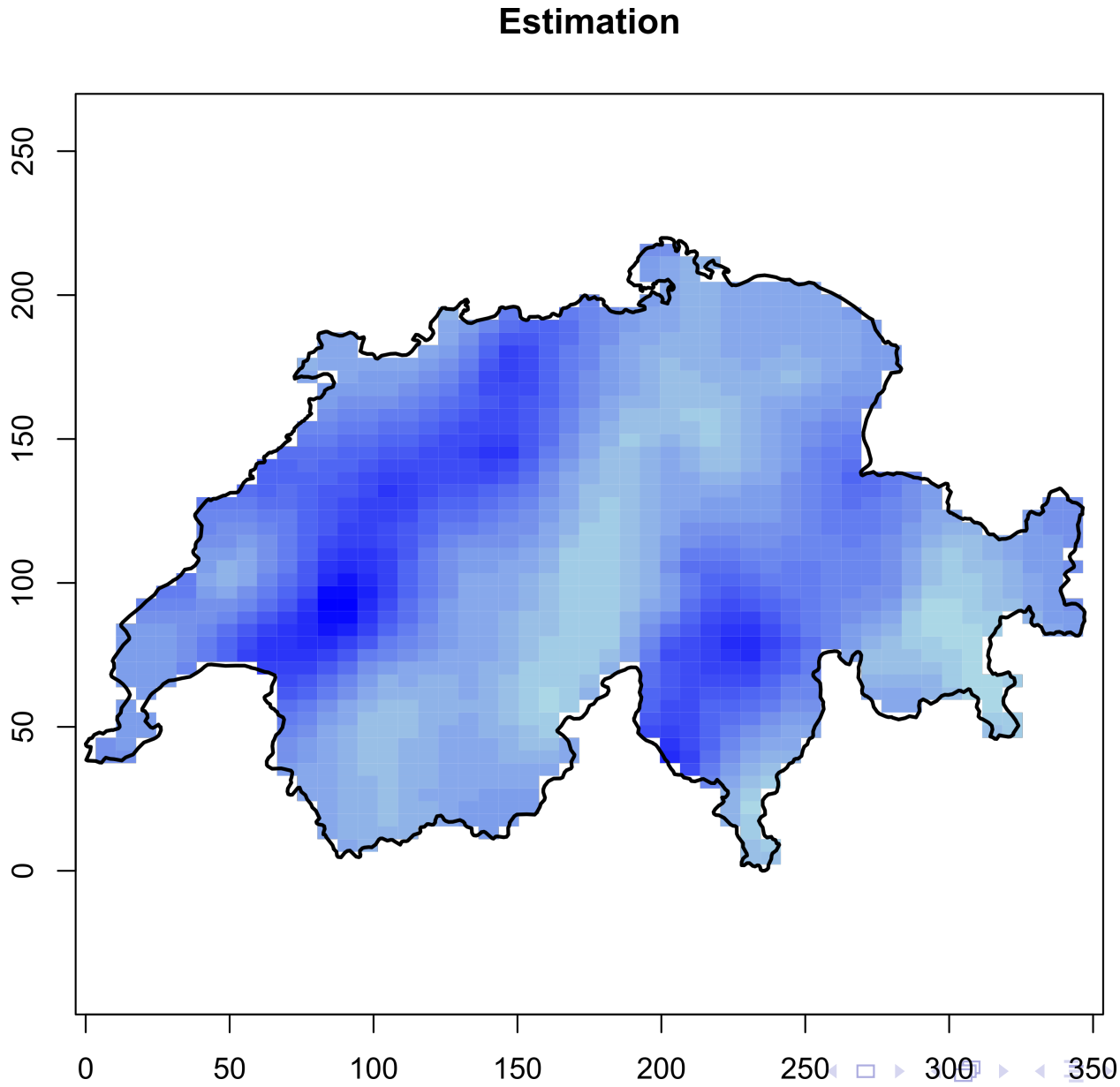
Interpolation spatiale et lissage spatial

Interpolation spatiale : le krigeage



Interpolation spatiale et lissage spatial

Interpolation spatiale : le krigeage



Quelques notions de R

Affectation et calcul

R fonctionne un peu comme une calculatrice. Si vous tapez $2 + 3$, le logiciel vous retournera la valeur 5. Néanmoins, on utilisera R davantage comme un langage de programmation en suivant les principes de l'affectation informatique.

Exemple d'affectation avec R

```
a <- 2
b <- 3
c <- a + b
```

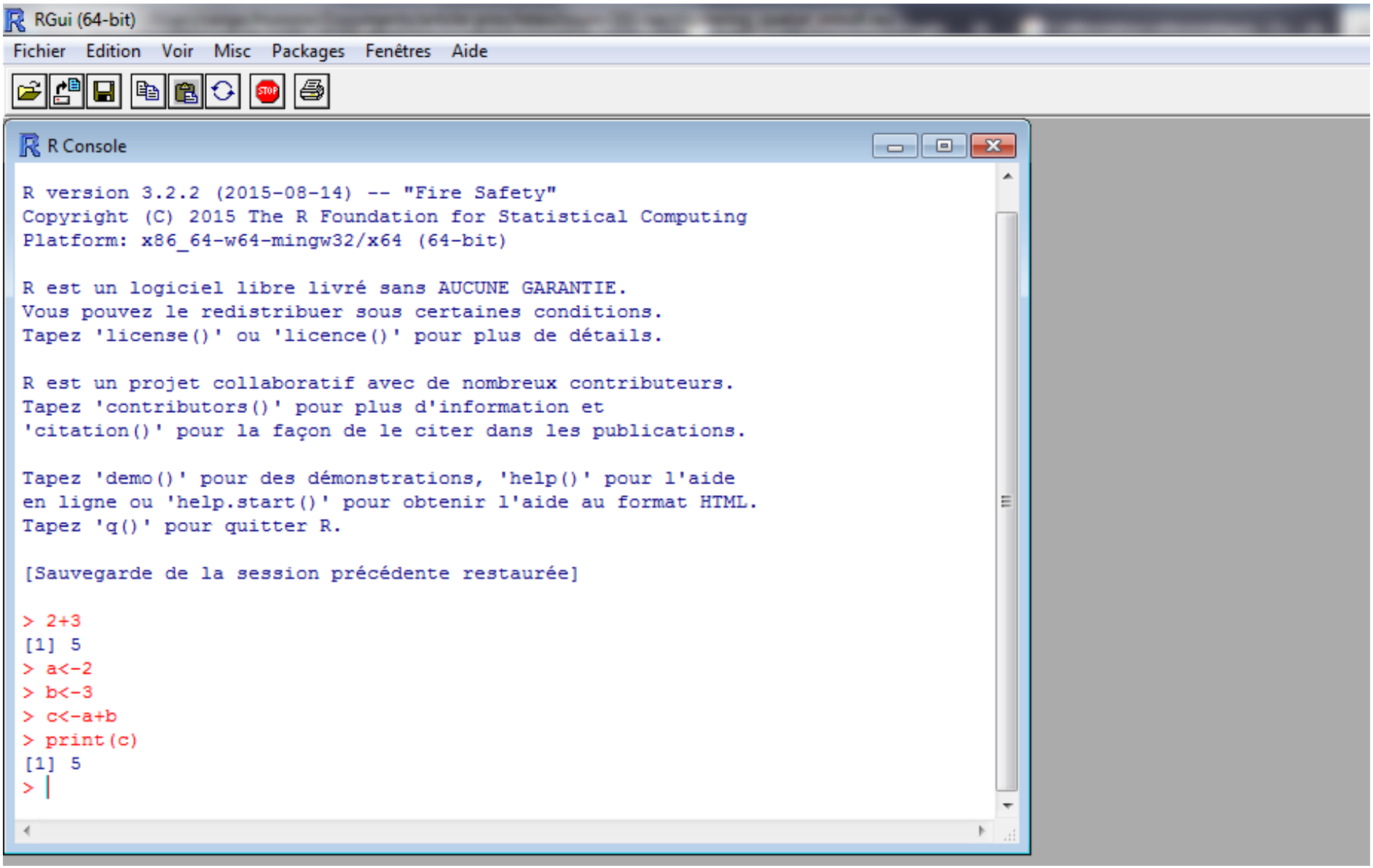
L'affichage des résultats se fera alors en utilisant une fonction : « print() ».

Affichage d'une variable avec R

```
print(c)
```

Quelques notions de R

Affectation et calcul



Quelques notions de R

Les types de données

Il existe de nombreux types de variables dans R.

Les variables de type texte

```
a <- "Texte"
```

Ces variables peuvent être ordonnées dans une liste (un vecteur) ou dans plusieurs listes pour former une matrice (un tableau de valeurs).

Les vecteurs et les matrices

```
b <- c(18, 182, 1.5, 15, 200, 5)
c <- matrix(c (18, 182, 1.5, 15, 200, 5), nrow = 2)
d <- matrix(c (18, 182, 1.5, 15, 200, 5), ncol = 2)
```

Quelques notions de R

Les types de données

Pour accéder à une valeur ou à un ensemble de valeurs, il faut utiliser les index des vecteurs ou des matrices.

Accès aux valeurs des vecteurs et des matrices

```
e <- b[2] + b[3]
f <- c[1,2] + c[2,3]
col <- c[,1]
ligne <- c[1,]
```

Accès avancé aux valeurs des vecteurs et des matrices

```
e <- b[c(2,4)]
f <- c[(c<15)]
g <- b[2 :5]
```

Quelques notions de R

Les types de données

Les data frames permettent de manipuler des tableaux bien structurés. Ce type de données est particulièrement bien adapté aux importations de fichiers textes.

Les Data Frames

```
articles <- c( "un", "le", "la", "les")
sujets <- c( "mot", "terme", "chose", "images")
dfmots <- data.frame(articles, sujets)
dfmots2 <- data.frame(col1 = articles, col2 = sujets)
```

Appel des valeurs des Data Frames

```
print(dfmots$sujets)
print(dfmots[,1])
```

Quelques notions de R

L'import de données et premières fonctions

Importation de fichiers textes

```
MyTexte <- read.table(file="c :/TheData.csv", header=TRUE, sep=",")
MyData <- read.csv(file="c :/TheData.csv", header=TRUE, sep=",")
adresse <- file.choose()
MyData <- read.csv(file=adresse, header=TRUE, sep=",")
```

Fonctions de base

```
res <- summary(b)
plot(d[,1],d[,2])
hist(b)
reg <- lm(d[,1] ~d[,2])
res3 <- summary(reg)
t.test(d[,1], d[,2])
```

Quelques notions de R

Les bibliothèques

Ce qui constitue la puissance de R, ce sont ses nombreuses bibliothèques qu'il faut télécharger.

Les librairies cartographiques

```
library(rgdal)
nuts3 <- readOGR(dsn = adresse, layer = "nuts3", verbose = TRUE)
library(sp)
class(nuts3)
nuts3@proj4string
head(nuts3@data)
plot(nuts3[1, ], col = "#5C99AD", border = " #2A5F70", lwd = 4)
library(rgeos)
europeBuffer <- gBuffer(spgeom = europe, width = 50000)
```

Introduction à R

Les boucles et la programmation

Enfin, comme tout langage de programmation, R permet de répéter les mêmes instructions plusieurs fois en changeant seulement quelques paramètres. Ce sont les boucles. Ces boucles peuvent alors permettre d'effectuer des tests. Ce sont par exemple les Si.

Les boucles

```
for (i in 1 :10) {
        print(i)
        }
for (i in 1 :10) {
        if (i > 5 & i < 8) {
                print(i)
                }
        }
```