Direction de la Diffusion et de l'Action régionale

H 2012/04

Détection des disparités socio-économiques
L'apport de la statistique spatiale

Jean-Michel Floch

A partir de travaux menés au sein de la Division des études territoriales

Document de travail

Institut National de la Statistique et des Études Économiques

Institut National de la Statistique et des Études Économiques

Série des documents de travail
de la Direction de la Diffusion et de l'Action Régionale

H 2012/04

Détection des disparités socio-économiques
L'apport de la statistique spatiale

Jean-Michel Floch (Insee-DAR)

A partir de travaux menés au sein de
la Division des études territoriales(*)

Décembre 2012

(*)Voir l'introduction pour un aperçu rapide de ces travaux, et la liste de celles et ceux qui, autour de Jean-Luc Lipatz, alors chef de la DET y ont contribué.

Table des matières

RÉSUMÉ 7

INTRODUCTION 8

1 - GÉNÉRALITÉS 9

1 - GÉNÉRALITÉS 9

1-1 « Analyse spatiale » et « statistique spatiale » 9

1-2 « Maup » : agrégation et zonage 9

1-3 Modèles temporels et modèles spatiaux 11

1-4 Hétérogénéité et dépendance 12

1-5 - Les trois grandes branches de la statistique spatiale, et leur manière de prendre en compte l'autocorrélation 12

2- AUTOCORRÉLATION, POINTS CHAUDS ET FROIDS : LE CAS DES DONNÉES SURFACIQUES 14

2-1 Règles de contigüité 14

2-2 Mesures d'association globales 15
2-2-1 Indicateur de Moran 16
2-2-2 Indicateur de Geary 18
2-2-3 Indicateur G de Getis et Ord 19

2-3 Indicateurs locaux d'association spatiale 19
2-3-1 Indicateur de Getis et Ord 19
2-3-2 Indicateur de Moran local 20
2-3-3 Indicateur de Geary local 22

3 - CONFIGURATIONS SPATIALES DE POINTS ET RATIOS DE DENSITÉ 23

3-1 Généralités 23

3-2 Propriétés du premier ordre 24

3-3 Propriétés du second ordre 25

3-4 - Densité de probabilité et intensité d'un processus 27

3-5 - L'estimation non-paramétrique de la densité 28

3-5- Des densités aux ratios de densité 34

3-6- Du ratio théorique au ratio estimé 37

3-7 Comparaison avec les résultats des Lisa 38

4- EXTENSION DU CHAMP D'APPLICATION **40**

4-1 - Le principe **40**

4-2 Application à l'analyse des données **40**

4-3 Un exemple de classification **42**

4-3 Utilisation pour le calcul des indicateurs de ségrégation spatiale **42**

5 - LA RÉGRESSION GÉOGRAPHIQUE PONDÉRÉE (RGP) **45**

5-1- Présentation générale **45**

5-2 Théorie élémentaire de la RGP **46**
5-2-1 Les bases 46
5-2-2 La détermination des paramètres optimaux 47
5-2-3 Retour sur l'exemple 48
5-2-4 Test sur la non stationnarité 48

5-3 Une utilisation de type « petits domaines » **49**

5-4 Lien avec l'estimation de densité **49**

6 - DÉTECTION DE CLUSTERS D'ACTIVITÉ OU D'ÉQUIPEMENTS **51**

6-1 Indicateurs de Ripley et prolongements **51**

6-2 Le M de Marcon et Puech **51**

BIBLIOGRAPHIE **54**

Résumé

Ce document présente quelques méthodes, issues de la statistique spatiale, permettant de déterminer des zones de concentration de difficultés sociales.

Après une présentation générale des problèmes liés à l'utilisation des données spatialisées (MAUP, problème de l'échiquier), il présente les indicateurs d'autocorrélation globaux (Moran , Geary) et locaux (Getis & Ord, LISA) relevant des méthodes « surfaciques ».

La partie la plus importante du document traite de l'estimation de la densité des données ponctuelles, grâce à des méthodes non paramétriques et de leur utilisation pour la détermination de zones de surreprésentation grâce aux ratios de densité.

D'autres méthodes relevant des processus ponctuels (Fonctions K de Ripley et dérivées) ou de la régression (Régression géographique pondérée) sont également présentées.

Mots clefs : *statistique spatiale, autocorrélation, estimation non-paramétrique, régression géographique, indicateurs de ségrégation.*

Abstract

This paper presents some methods, issued from spatial statistics, used to detect area with concentration of social problems.

It begins with a general presentation of spatial data (Modifiable areal unit problem, checkerboard problem) and go on with global spatial autocorrelation measures (Moran , Geary) and locals (Getis & Ord, LISA) issued from "areal" methods.

The most important part of this paper is about density estimation of point patterns, with non parametric methods. Ratios of estimated densities are used to detect clusters of social problems..

Others methods (K Ripley's function, geographically weighted regresion) are also presented.

<Keywords : *spatial statistics,autocorrelation,kernel density estimation, geographically weighted regression*

Introduction

Ce document de travail présente quelques unes des méthodes utilisées dans les travaux menés au sein de la Division des études territoriales (DET) de l'Insee. Il est donc le produit d'un travail collectif, initié il y a un peu plus de 10 ans par Michel Hanoun et Jean-Luc Lipatz, alors chef de la division. Ce travail visait à produire, en rapprochant données géolocalisées et techniques de statistique spatiale, des outils utilisables par les directions régionales de l'Insee pour l'analyse urbaine. Responsable plusieurs années du PSAR-Analyses urbaines au sein de la DET, l'auteur de la présente publication est dans une large mesure le porte-plume, si l'on peut encore utiliser ce terme de nos jours, d'une aventure collective dont il a été un des protagonistes.

L'utilisation de méthodes de statistique non paramétrique, déjà préconisées par Philippe Chataignon a débouché sur l'écriture par J .-L. Lipatz d'un kit de macros SAS destiné, dans le cadre de la mise en place des Contrats urbains de cohésion sociale (CUCS) ,à aider les décideurs dans la recherche des zones en situation précaire. Ce kit a été enrichi au cours des années, et a été au cœur de l'investissement « Synthèse urbaine ». Disponible sur ftp\consultation\infracommunal\kde, il est à l'origine d'une bonne partie des cartes de ce document. La DET et le PSAR ont également exploré quelques autres méthodes (Régression géographique pondérée, fonctions de Marcon et Puech..).

Pour aider les chargés d'étude travaillant à partir de ces méthodes, une formation à l'analyse spatiale a été organisée à partir de 2004. Y sont intervenus, outre J-L Lipatz et l'auteur de ce document, Michel Hanoun, Jean-Luc Le Toqueux, Stéphanie Himpens, Stéphanie Mas, Cynthia Gaborieau-Faivre, Marc Branchu et Benoît de Lapasse. Ce document doit beaucoup à leurs contributions respectives, et ils retrouveront sans doute au fil des pages des graphiques issus de leurs interventions lors des formation.

Envisagé d'abord comme un accompagnement à cette formation, ce papier en développe davantage certains aspects méthodologiques et cherche à resituer les méthodes utilisées dans le vaste domaine de la statistique spatiale. Certaines des interprétations sont personnelles, et selon la formule, les erreurs qui apparaîtraient ne sont imputables qu'à l'auteur du document.

1 - Généralités

Ce document présente quelques notions de base de statistique spatiale, mais n'en constitue pas une introduction, dans la mesure où il privilégie un nombre limité de méthodes, celles qui ont été utilisées pour la détection de zones de surreprésentation de difficultés sociales lors de la mise en place des CUCS.

De nombreux manuels ou documents en ligne permettent aux intéressés d'aller plus loin dans le vaste domaine de la statistique spatiale : Bailey et Gattrell (1995), Loyd (2006) pour une première approche, Diggle (2003), Gotway et Schabenberger (2004) ou Cressie (1993) pour une approche systématique, par exemple.

1-1 « Analyse spatiale » et « statistique spatiale »

Il est question dans ce document de statistique spatiale. On s'accordera avec ce que dit le géographe Claude Grasland lorsqu'il distingue l'analyse spatiale de la statistique spatiale. Pour lui, « *l'analyse spatiale est une branche des sciences sociales qui s'attache aux déterminants de la localisation des hommes et des activités, tandis que la statistique spatiale est une branche de la statistique qui étudie les particularités des distributions de population dans un espace bidimensionnel* » Selon Grasland, le concept central en analyse spatiale est celui d'interaction, tandis qu'en statistique spatiale, c'est l'autocorrélation qui est essentielle.

Pour le « statisticien spatial », l'introduction de l'espace correspond à un enrichissement, au prix de difficultés conceptuelles assez considérables, de méthodes élaborées dans un contexte « a-spatial ». Les modèles spatiaux plus riches du statisticien restent néanmoins pour l'analyste des visions simplifiées de la réalité.

Ces différences d'approches exposéees, il reste que ces deux mondes entretiennent des relations étroites. L'introduction d'une forte dose de quantitativisme dans les études géographiques, l'utilisation des méthodes statistiques et économétriques en économie ou en géographie ont contribué à des rapprochements entre ces univers. Les questions posées de façon insistante par les géographes (comme celles du Modifiable areal unit problem) n'ont pas été sans influence sur la statistique spatiale, pour la remise en cause d'hypothèses difficilement tenables.

Dans la littérature, cette proximité se manifeste particulièrement dans la « statistique spatiale exploratoire », dans le sillage des méthodes initiées par J.Tukey. Un livre comme celui de Bayley et Gattrell (1995) en témoigne. Une technique comme la régression géographique pondérée, présentée dans ce document, est due à des géographes. Les travaux de Moran, publiés dans des revues de statistique, alimentent les travaux de géographes.

Ce document expose quelques méthodes de statistique spatiale, mais il est destiné à des utilisateurs qui seront surtout amenés à faire de l'analyse spatiale.

1-2 « Maup » : agrégation et zonage

MAUP est l'acronyme de « modifiable areal unit problem ». Il a été introduit sous ce terme par le géographe Steve Openshaw, qui a reformulé des questions anciennes, en s'aidant des possibilités offertes par l'informatique pour faire des simulations. Questions anciennes, parce que l'on sait de longue date qu'en changeant de découpage, on obtient des représentations cartographiques différentes.

Comment résumer ce terme de Maup? Sans doute en disant qu'il s'agit d'un problème d'agrégation et/ou de zonage.

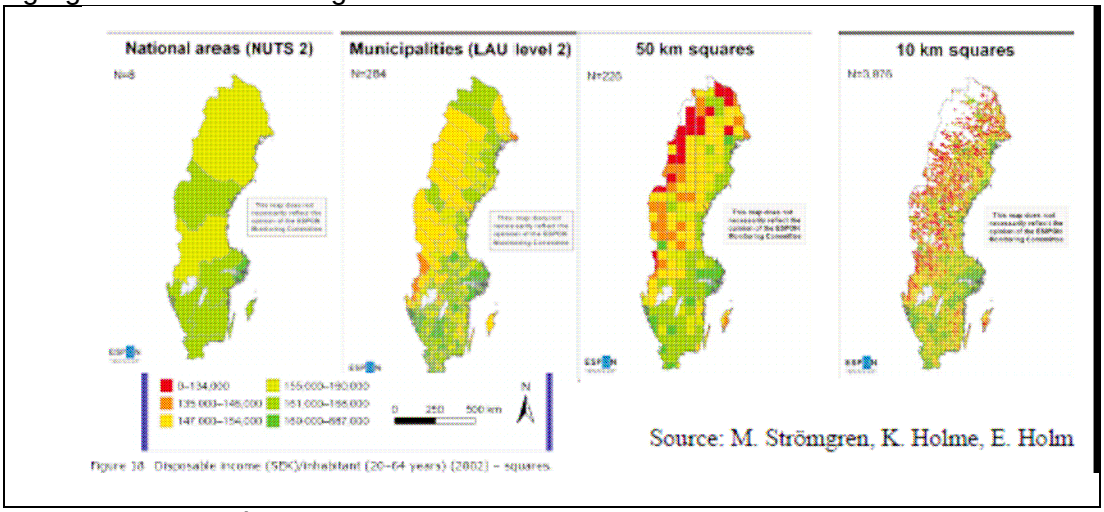


Figure 1.1

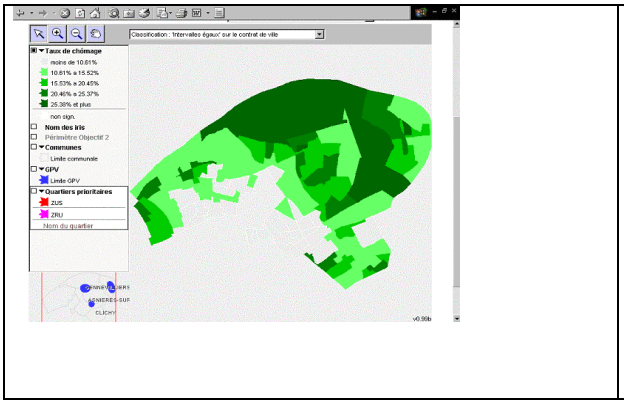
La géographe Léna Sanders, à qui est empruntée la carte ci-dessus en pointe quelques conséquences :

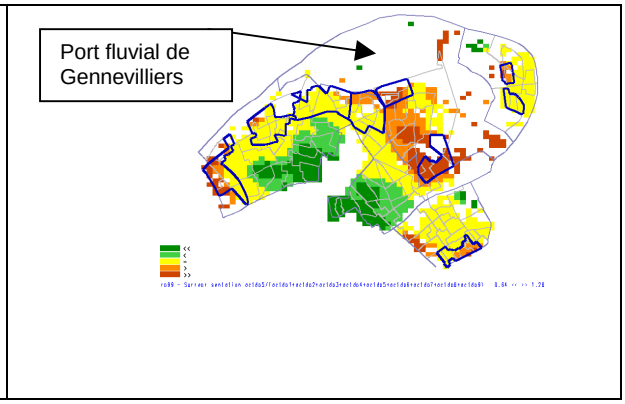
http://wwwinseefr/fr/insee_regions/lor/actionregionale/colloques/Diaporama_Sanders2011pdf

a- Les représentations cartographiques et les analyses statistiques vont donner des résultats différents selon les entités spatiales élémentaires choisies ;
b- Selon le zonage choisi, les évolutions spatiales sont différentes ;
c- Les correspondances entre des limites administratives et des discontinuités spatiales sont possibles, mais sont rarement la règle ;
d- Avec différentes définitions d'entités spatiales, les formes d'évolution apparaissent différemment ;
e- Les relations entre les phénomènes changent en fonction de l'échelle d'observation ;
f- Les relations entre les phénomènes dépendent de l'étendue spatiale de la région étudiée.

On peut remarquer que ces questions se posent dans des termes proches lorsque l'on publie des résultats dans des nomenclatures à différents niveaux. C'est le cas par exemple des travaux sur la concentration industrielle, qui peuvent donner des résultats différents, voire contradictoires, selon que la nomenclature utilisée pour décrire le tissu industriel est plus ou moins fine.

Les représentations cartographiques peuvent renforcer l'effet du Maup, pour des raisons d'ordre visuel. Ainsi une zone de faible effectif, mais de grande superficie, produit effet marqué risquant de donner une image déformée (exemple du chômage à Gennevilliers, cartes 1.1 et 1.2).





Carte 1.1 (site de la DIV) **Carte 1.2 (carte DET)**

Plusieurs solutions ont été proposées pour pallier à ces problèmes de Maup. Toujours en suivant L Sanders, on peut citer :

- l'identification du niveau pertinent ;
- la recherche de partitions optimales;
- le niveau individuel ;
- le carroyage et le lissage ;
- l'analyse multi-scalaire.

Ces propositions ne sont pas antagoniques, mais aucune ne résout la question du Maup. On verra par la suite que les méthodes non paramétriques, qui permettent de s'abstraire des découpages administratifs, donnent des résultats qui sont tributaires de la fenêtre d'estimation, et qu'il est difficile de définir des solutions optimales.

L'analyse multi-échelle constitue en tout état de cause une voie intéressante, car elle a l'avantage de prendre en compte les observations qui peuvent être pertinentes à chacun des niveaux d'observation considéré, et à les combiner pour obtenir une analyse pertinente.

1-3 Modèles temporels et modèles spatiaux

Les modèles spatiaux présentent des analogies avec les modèles utilisés en séries temporelles. Dans ces derniers, le temps, le t de *X(t)* est une information à part entière. Il permet d'étudier la succession des observations, leur périodicité, et d'en construire des modélisations.

Il en va de même pour les observations spatiales. La situation dans l'espace est une information aussi importante que la variable d'intérêt, et le modèle spatial *X(s)* n'est pas la même chose qu'un modèle X, spécifié sans tenir compte de la position des observations.

Sur le plan théorique, les modèles temporels, comme les modèles spatiaux, reposent sur des processus stochastiques. On pourra trouver des analogies entre des outils comme les variogrammes utilisés en géostatistique et les corrélogrammes.

Ceci dit, cette analogie a aussi des limites. En séries temporelles, on a une direction, et des intervalles de temps permettant de faire des modélisations empiriques (moyennes mobiles) ou probabilistes (ARMA et dérivés). Le spatial est plus complexe puisque l'on se situe dans un plan, et en général dans des structures présentant des discontinuités. Les lois

de probabilités qui sont au centre de la modélisation des phénomènes spatiaux et temporels ne sont pas les mêmes.

1-4 Hétérogénéité et dépendance

Les données spatiales sont caractérisées par des emboîtements d'échelles, les processus pouvant être différents (homogènes à une certaine échelle, hétérogènes à d'autres). Les méthodes statistiques dans le domaine spatial vont être amenées à traiter :

- l'hétérogénéité. Il est bien connu pour les praticiens des données spatialisées que le spatial est le domaine de l'hétérogénéité, et que les transitions peuvent avoir lieu de façon assez abrupte. Le problème se pose alors de voir comment les méthodes conçues dans un cadre standard, pour étudier des phénomènes homogènes vont pouvoir être adaptées à des situations où les processus ne sont pas stationnaires, et où on ne trouve pas a priori des formes de régularité qui servent à modéliser les séries temporelles.
- La dépendance. Elle est exprimée de façon imagée par Tobler, qui en a fait la première « loi de la géographie » « *Everything is related to everything else, but near things are more related than far things* ». Cette situation fait que les hypothèses classiques de la statistique mathématique qui justifient les modèles usuels ne peuvent plus s'appliquer. Les structures de dépendance spatiale sont complexes. Elles ne sont pas toujours faciles à tester et nécessitent de définir un voisinage de façon pertinente.

1-5 - Les trois grandes branches de la statistique spatiale, et leur manière de prendre en compte l'autocorrélation

La statistique spatiale ne s'est pas développée de façon rectiligne et unifiée. De grandes branches, comme la géostatistique se sont constituées de façon assez indépendante. Ce n'est qu'assez tardivement que Cressie (1993) a proposé une unification des différents champs de ce vaste domaine, en reprenant la notion de variable régionalisée introduite par Georges Matheron, père tutélaire de la géostatistique à l'Ecole des Mines de Fontainebleau.

On distingue habituellement trois grands types de données dans la statistique spatiale :

- les données ponctuelles ;
- les données de surface ;
- les données géostatistiques.

L'introduction de ces méthodes a été assez lente à l'INSEE. Elle a été rendue indispensable par la nécessité de manipuler de grosses bases de données géolocalisées. Les problèmes liés à l'autocorrélation spatiale, au MAUP ont fait petit à petit leur chemin.

Il restait à résoudre la question des méthodes pertinentes à utiliser pour traiter les données. Les méthodes géostatistiques ont été testées, avec peu de succès. Les données de surface qui sont en apparence les plus proches de ce qui est pratiqué à l'INSEE n'ont fait l'objet que de peu d'investigations. Ce sont les techniques « ponctuelles » qui se sont avérées les mieux à même de traiter les questions qui se posaient dans le domaine de l'analyse urbaine.

Ces méthodes ponctuelles sont utilisées depuis un certain temps dans plusieurs domaines d'étude. Nous avons surtout utilisé des travaux relatifs à :

- l'épidémiologie ;
- la forêt.

On s'est rendu compte que les spécialistes de ces domaines travaillaient sur des problématiques qui apparaissaient transposables dans le domaine des études urbaines. Les épidémiologistes parlent fréquemment de « risque relatif » dans l'étude de pathologie. Ils essaient de déterminer les zones où le risque est fort, et ensuite de modéliser ce risque en introduisant des variables explicatives. Cela ressemble fortement à ce que l'on recherche quand on s'intéresse au chômage. Les forestiers s'intéressent à la répartition des espèces, à la façon dont les différentes essences végétales s'associent ou au contraire se repoussent.

Les méthodes ponctuelles ont semblé fournir, dans une première étape, les outils les plus porteurs pour commencer à traiter de façon satisfaisante les données géolocalisées dont disposait la statistique publique.

On commencera par présenter ce qui fonde ces méthodes, avant de traiter de façon spécifique le traitement des problèmes de risque relatif et de zones de sur-représentation. Les techniques statistiques présentées seront donc choisies en fonction de leur intérêt pour les analyses spatiales qu'elles permettront d'éclairer.

La prise en compte de l'autocorrélation est un enjeu majeur pour la statistique spatiale, afin de spécifier correctement les modèles. Des techniques spécifiques de mesure de l'autocorrélation ont été élaborées dans chacune des grandes branches de la statistique spatiale :

- variogramme en géostatistique ;
- indicateurs d'association spatiale pour les données surfaciques ;
- fonctions de Ripley et dérivées dans les méthodes ponctuelles.

Ces indicateurs ont en commun de faire intervenir les distances, ou les proximités entre les observations. Les points communs entre les différentes méthodes ne sont pas forcément apparentes. Il a semblé intéressant pour commencer d'étudier l'autocorrélation dans les données surfaciques.

2- Autocorrélation, points chauds et froids : le cas des données surfaciques

La mesure de la corrélation nécessite la prise en compte des données qui sont à proximité des zones d'observation. Cela implique de définir des relations de voisinage, que l'on pourrait définir pompeusement comme topologique, ou comme métriques.

2-1 Règles de contigüité

Prenons comme exemple, à la suite de Cliff et Ord trois types simples de contigüité. Dans le schéma ci-dessous, chaque zone « grise » est contigüe à la zone « jaune » selon les règles du jeu d'échec.

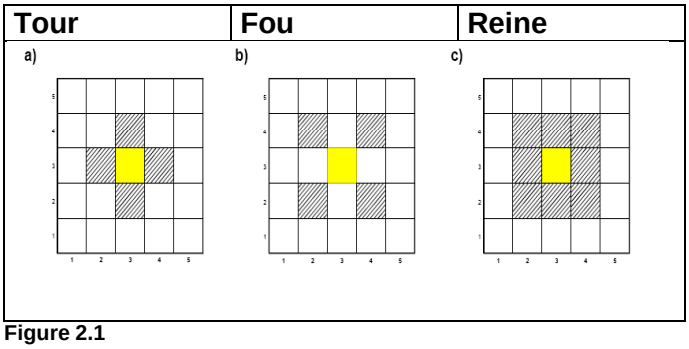


Figure 2.1

Ces règles de contiguité permettent de définir des matrices, dites également de contiguité, qui vont permettre des calculs. La règle la plus simple est de prendre $w_{ij}=1$ si les zones i et j sont contigues, $w_{ij}=0$ dans le cas contraire; en considérant qu'une zone n'est pas contigue à elle-même En reprenant les trois exemples simples du déplacement du fou, de la tour et de la reine sur un échiquier 3x3 les matrices sont les suivantes :

Tour	Fou	Reine
0 1 0 1 0 0 0 0 0	0 0 0 0 1 0 0 0 0	0 1 0 1 1 0 0 0 0
1 0 1 0 1 0 0 0 0	0 0 0 1 0 1 0 0 0	1 0 1 1 1 1 0 0 0
0 1 0 0 0 1 0 0 0	0 0 0 0 1 0 0 0 0	0 1 0 0 1 1 0 0 0
1 0 0 0 0 0 1 0 0	0 1 0 0 0 0 0 1 0	1 1 0 0 1 0 1 1 0
0 1 0 1 0 1 0 1 0	1 0 1 0 0 0 1 0 1	1 1 1 1 0 1 1 1 1
0 0 1 0 0 0 0 0 1	0 1 0 0 0 0 0 1 0	0 1 1 0 1 0 0 1 1
0 0 0 1 0 0 0 1 0	0 0 0 0 1 0 0 0 0	0 0 0 1 1 0 0 1 0
0 0 0 0 0 0 1 0 1	0 0 0 1 0 1 0 0 0	0 0 0 1 1 1 1 0 1
0 0 0 0 0 1 0 1 0	0 0 0 0 1 0 0 0 0	0 0 0 0 1 1 0 1 0

Tableau 2.1

On constate facilement qu'au vu des règles de voisinage qui sont définies :

- les matrices de contigüité sont symétriques ;
- la matrice des déplacements de la reine est la somme de celle de la tour et de celle du fou.

Les exemples de règle de contigüité sont nombreux et plus ou moins complexes. Citons parmi ceux-ci :

a) $w_{ij} = 1$ si les zones *i* et j ont une frontière commune, $w_{ij} = 0$ dans le cas contraire; cette matrice est symétrique ;

b) w_{ij} est le pourcentage du périmètre total de la zone i qu'elle partage en commun avec la zone *j*; cette matrice n'est pas symétrique ;

c) $w_{ij} = 1$ si la distance entre les régions *i* et *j* est inférieure à une distance critique et $w_{ij} = 0$ au delà. Les définitions de cette distance critique sont multiples : distances entre les centres géographiques des zones, entre les capitales administratives etc ;

d) w_{ij} est une fonction décroissante de la distance entre points privilégiés des zones i et j Cliff et Ord (1973) proposent une formulation du type $wij = d_{ij}^{-a} b_{ij}^{b}$, avec d_{ij} distance entre les zones *i* et *j*, b_{ij} pourcentage du périmètre de la zone *i* constitué par la frontière avec la zone *j* et *a* et *b* deux paramètres à estimer ;

e) w_{ij} est une mesure de l'accessibilité entre les zones *i* et *j*, distance ou temps de transport par la route, rail, air, etc ;

e) w_{ij} peut enfin traduire des proximités non géographique de type organisationnelle.

Les règles de contigüité ci-dessus sont selon les cas topologiques, métriques ou mêlent les deux. On remarquera que dans certains cas, la matrice de contiguité est une fonction de la distance, et s'exprime comme $w_{ij}(d)$.

Les matrices de contiguité s'expriment souvent sous une forme normalisée en ligne, c'est à dire de telle façon que la somme en ligne des pondérations soit égale à 1.

Cette normalisation des coefficients permet de calculer des indicateurs locaux et de comparer la valeur observée de la variable d'intérêt en i, et sa valeur dans le voisinage.

2-2 Mesures d'association globales

Une mesure d'association spatiale est définie de façon générale comme la somme pondérée sur les couples de points possibles d'un indicateur de similarité, noté sim_{ij} soit :

$$\frac{\sum_{i=1}^{N}\sum_{i=1}^{N} w_{ij}\, sim_{ij}}{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}}$$

Les indices utilisés classiquement, ceux de Moran et Geary sont des exemples de cette formulation générale.

2-2-1 Indicateur de Moran

Il est défini en utilisant l'indicateur de similarité $sim_{ij} = (z_i - \bar{z})(z_j - \bar{z})$, $\bar{z}$ étant la moyenne arithmétique sur l'ensemble du champ d'observation. Si s^2 désigne la variance des observations, l'indicateur de Moran, noté I pour la variable Z s'écrit de la façon suivante :

$$I = \frac{1}{s^2} \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}(z_i - \bar{z})(z_j - \bar{z})}{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}}$$

L'indicateur de Moran I est une variable aléatoire, dont la distribution est déterminée par celle de la variable d'intérêt et des interactions spatiales.

Les résultats asymptotiques ont été calculés par Cliff et Ord (1973). Ces auteurs ont montré que sous hypothèse d'indépendance :

$E(I) = \frac{-1}{N-1}$, et donc que l'indicateur de Moran a une valeur qui tend vers 0 lorsque le nombre de zones d'observation s'accroit.

On peut remarquer que l'indicateur de Moran peut s'écrire également sous la forme

$$I = \frac{\sum\sum wij \frac{(zi - z)(zj - z)}{\sum\sum wij}}{\sqrt{\sum \frac{(z_i - \bar{z})^2}{N}} \sqrt{\sum \frac{(z_j - \bar{z})^2}{N}}}$$

ce qui donne à l'indicateur de Moran une allure proche (aux pondérations spatiales près) du classique coefficient de corrélation de Pearson. Pour diverses bonnes raisons exposées par exemple dans Gotway et Waller(2004), ce coefficient ne peut cependant pas être interprété comme un coefficient de corrélation. Il n'est pas inférieur ou égal à 1 en valeur absolue comme le coefficient de Pearson et ne peut être considéré comme un cosinus d'angle. Ses bornes peuvent être calculées. On peut trouver des résultats sur les limites des valeurs de I dans Cliff & Ord(1973) ou Gotway et Waller(2004). On a, sauf cas particuliers $|I| < 1$.

Pour comparer la valeur observée de l'indicateur d'association spatiale à l'hypothèse nulle, il faut réaliser des tests Dans le cas où l'on se réfère à des hypothèses (fortes et peu vraisemblables…) de normalité, on va comparer $\frac{I - E(I)}{\sqrt{V(I)}}$ à la distribution d'une loi normale N(0,1).

L'expression de la variance peut être calculée à l'aide de la matrice des poids, et s'exprime de la façon suivante : $V(I) = \frac{N^2 S_1 - N S_2 + 3 S_0^2}{(N-1)(N+1) S_0^2} - \left(\frac{1}{N-1}\right)^2$ où :

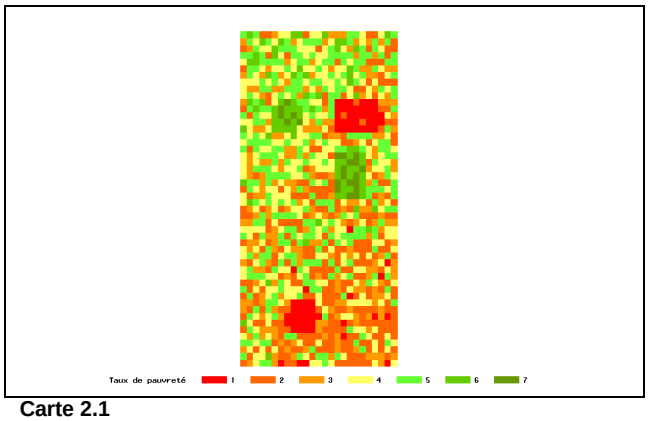
$$S_0 = \sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}$$

$$S_1 = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} (w_{ij} + w_{ji})^2$$

$$S_2 = \sum_{i=1}^{N} \left(w_{i.} + w_{.i} \right)^2$$

Ces formules paraissent un peu lourdes mais ne sont pas trop difficiles à implémanter. Par contre, elles reposent sur des hypothèses assez fortes. C'est pourquoi on préfère souvent calculer l'espérance et la variance de I en effectuant un grand nombre de permutations aléatoires des observations.

On trouvera ci-dessous un exemple de calcul du coefficient de Moran, réalisé sur des données simulées de taux de pauvreté au niveau de carreaux (carte 2.1). La pauvreté est décroissante des couleurs chaudes vers les couleurs froides.



Carte 2.1

L'indicateur de Moran est calculé sur cet exemple en utilisant la matrice de proximité la plus simple (contiguité d'ordre 1 au sens de la Reine). Comme on le fait souvent dans le calcul de ces indicateurs, les pondérations ont été normalisées afin que la somme en ligne soit égale à 1.

Il y a dans la carte 1250 carreaux qui interviennent dans le calcul. Les valeurs obtenues sont les suivantes :

***Indicateur de Moran* :**	0.524
Espérance de l'indicateur	-0.O007
Variance de l'indicateur	0.0196
Valeur à tester	26.69
P-value	0.00

Tableau 2.2

Ces résultats indiquent qu'on ne peut rejeter l'autocorrélation spatiale. La valeur à tester est très supérieure à 0. La probabilité pour qu'avec la valeur observée, on rejette à tort l'hypothèse nulle (absence d'autocorrélation spatiale) est très faible.

Une autre façon de tester la valeur du coefficient de Moran, par rapport à l'hypothèse nulle d'absence d'autocorrélation spatiale, est de réaliser un grand nombre de permutations des résultats observés, et d'étudier la distribution des résultats obstenus. On pourra ainsi voir dans quel pourcentage des cas le résultat que l'on obtient est dépassé par les permutations réalisées.

L'indicateur de Moran peut présenter des limites. Les formules présentées ci-dessus sont des variables quelconques. Si on l'applique à des données de comptage, il faut l'utiliser avec prudence, comme le montrent de nombreux exemples d'épidémiologie. Si on s'intéresse à la mesure de l'autocorrélation des cas d'une pathologie, on peut avoir une

autocorrélation positive des cas constatés qui ne fait que refléter la répartition spatiale de la population de référence.

C'est pourquoi des statisticiens travaillant dans le domaine de la santé ont proposé des formes alternatives de l'indice de Moran, qui s'expriment de la façon suivante :

$$I_{CR} = \frac{\sum_{i=1}^{N}\sum_{i=1}^{N} w_{ij} \frac{Z_i - rn_i}{\sqrt{rn_i}} \frac{Z_j - rn_j}{\sqrt{rn_j}}}{\sum_{i=1}^{N}\sum_{i=1}^{N} w_{ij}}$$

Cette expression fait apparaître l'écart entre la grandeur observée Z_i (le comptage d'une pathologie, par exemple) à la valeur attendue rn_i, obtenue en appliquant un risque constant r à la population totale de la zone, l'expression étant normée par la racine de rn_i .

On peut l'interpréter comme le résultat de l'application d'une régression poissonnienne de la variable explicative z_i sur l'effectif n_i de la zone. Pour arriver à l'expression mentionnée, il faut faire l'hypothèse qu'en moyenne, les résidus studentisés sont nuls. On trouvera dans Gotway & Waller (2004) des exemples tirés d'un même jeu de données et qui conduisent à des résultats très différents entre I et I_{CR}.

2-2-2 Indicateur de Geary

Cet indicateur est basé sur une autre formule de similarité : $sim_{ij} = \left(z_i - z_j \right)^2$

Il a la forme suivante :

$$c = \frac{N-1}{2\sum_{i=1}^{N}\left(Z_i - \bar{Z}\right)^2} \frac{\sum_{i=1}^{N}\sum_{i=1}^{N} w_{ij}\left(Z_i - Z_j\right)^2}{\sum_{i=1}^{N}\sum_{i=1}^{N} w_{ij}}$$

Cet indicateur prend des valeurs comprises entre 0 et 2, la valeur 0 indiquant une autocorrélation spatiale positive, 2 une autocorrélation spatiale négative. Cet indicateur présente des analogies de forme avec celui de Durbin et Watson.

On trouve dans l'article d'Anselin (1995) des résultats théoriques sur les moments de l'indicateur de Geary :

$$E(c) = 1$$

$$Var(c) = \frac{(N-1)(2S_1 + S_2) - 4S_0^2}{2(N+1)S_0^2}$$

avec les conventions d'écriture adoptées pour le calcul des indices de Moran.

On peut tester, avec les mêmes hypothèses fortes et les mêmes limites, l'autocorrélation spatiale au sens de Geary en utilisant la normalité asymptotique de $\frac{I - E(c)}{\sqrt{Var(c)}}$

Cliff et Ord (1973) ont montré le lien qui existe entre l'indicateur de Moran et celui de Geary, qui peut s'écrire :

$$c = \frac{N-1}{S_0} \frac{\sum_{i=1}^{N} w_{i.} \left(z_i - \bar{Z} \right)^2}{\sum_{i=1}^{N} \left(z_i - \bar{Z} \right)^2} - \frac{N-1}{N} I$$

Les indicateurs de Moran et de Geary ont été abondamment utilisés, tout particulièrement le premier. Ils ne sont cependant pas sans inconvénients On peut renvoyer aux limites que souligne d'Aubigny dans Droesbeke et alii (2005), et que l'on peut résumer ainsi :

- ces indices sont construits par analogie avec des indicateurs existants (Pearson, Durbin & Watson) ;
- la principale faiblesse réside dans les poids égaux attribués à chaque entité spatiale, ce qui apparaît peu réaliste ;
- les bornes de variation ne sont pas bien connues ;
- on a deux indices concurrents, sans lien exploitable.

2-2-3 Indicateur G de Getis et Ord

Une autre statistique d'association spatiale a été proposé par Getis et Ord (1996), qui consiste à définir des poids $w_{ij}(d)$ définis pour une distance d déterminée. G peut s'écrire :

$$G(d) = \frac{\sum_{i=1}^{N} \sum_{i=1}^{N} w_{ij}(d) z_i z_j}{\sum_{i=1}^{N} \sum_{i=1}^{N} z_i z_j}$$

Cet indicateur global est donné ici pour mémoire, l'indicateur de Getis et Ord étant plutôt utilisé comme un indicateur local, permettant de détecter des « points chauds »

2-3 Indicateurs locaux d'association spatiale

Ces indicateurs sont connus dans la littérature sous le nom de LISA (Local indicators of spatial association). Ils ont été introduits par Luc Anselin 1995) dans un article du même nom.

Ces indicateurs peuvent être définis dans la zone *i* comme une fonction $f(zi, zj)$ mettant en relation la valeur observée dans la zone et celle qui est observée dans son voisinage. Ils ont deux propriétés :

- ils indiquent l'importance de la propension à former des grappes autour d'une zone considérée ;
- leur somme est proportionnelle à un indicateur global d'association spatiale. On aura donc des Moran locaux et des Geary locaux.

2-3-1 Indicateur de Getis et Ord

Getis et Ord (1996) ont suggéré deux indicateurs permettant de déterminer des points « chauds » et « froids » . La différence entre les deux formulations tient à la prise en compte (ou non) dans le calcul de l'indicateur de la valeur au point d'observation.

Intuitivement, ces indicateurs sont simples à comprendre puisqu'ils mettent en rapport une valeur calculée au voisinage du point d'observation, et une valeur calculée sur l'ensemble du territoire considéré, auquel on a enlevé le point d'observation.

La première forme de l'indicateur est la suivante : $G_i(d) = \dfrac{\sum_{j=1, j\neq i}^{n} w_{ij} x_j}{\sum_{j=1, j\neq i}^{n} x_j}$

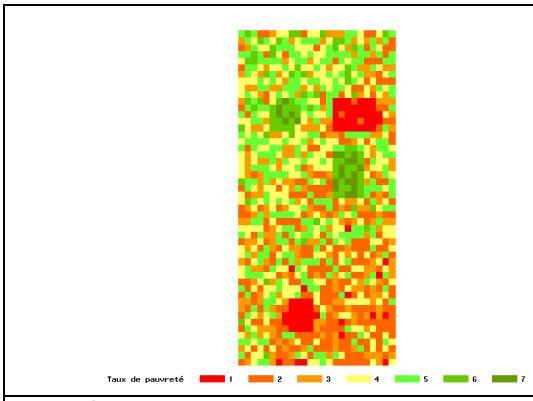
On peut calculer l'espérance et la variance de ces indicateurs, leurs expressions étant les suivantes :

$E(G_i) = \dfrac{W_i}{n-1}$ et $Var(G_i) = \dfrac{W_i(n-1-w_i)\,S^{2}_{\neq i}}{(n-1)^2(n-2)(\bar{x}_{\neq i})^2}$

On construit comme pour les autres indicateurs, un test asymptotique, permettant de repérer les points chauds et froids.

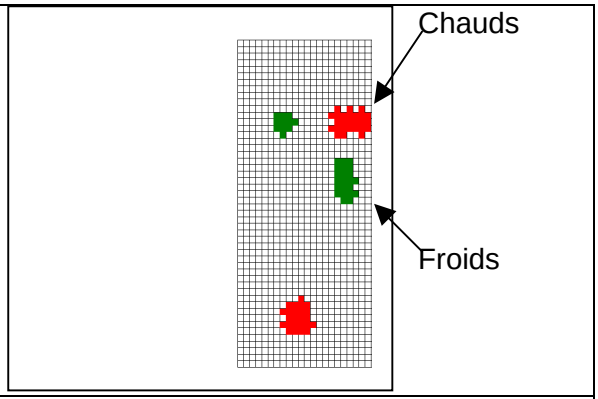
Getis et Ord (1996) suggèrent (classiquement) d'utiliser un test basé sur la normalité (hypothétique) de : $Z(G_i) = \dfrac{G_i - E(G_i)}{\sqrt{Var(G_i)}}$.

Dans l'exemple du taux de pauvreté présenté ci-dessus, on obtient le résultat suivant :



Données brutes

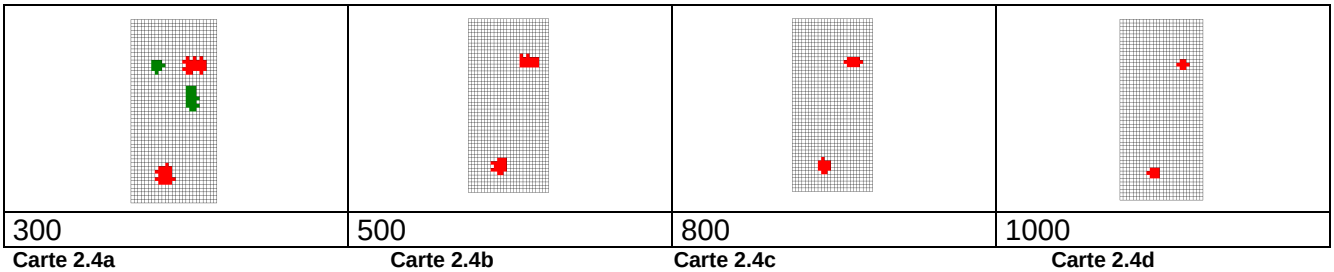
Carte 2.2



Points chauds et froids (300 mètres)

Carte 2.3

En faisant varier la distance d, on obtient les cartes suivantes :



300	500	800	1000

Carte 2.4a **Carte 2.4b** **Carte 2.4c** **Carte 2.4d**

2-3-2 Indicateur de Moran local

C'est le plus utilisé. Il a été proposé par Anselin (1995), et se présente sous différentes formes, dont la plus fréquente est la suivante :

$I_i = \dfrac{z_i}{m_2} \sum_{j=1}^{n} w_{ij} z_j$ où m_2 désigne le moment d'ordre 2 de la variable,soit $m_2 = \dfrac{1}{n}\sum_{i=1}^{n} z_i^{2}$

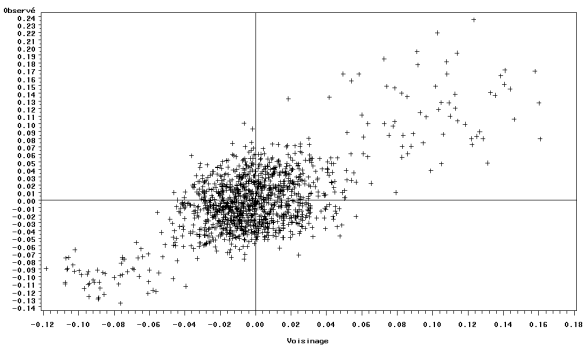
On vérifie facilement que $\sum_{i=1}^{n} I_i = \frac{1}{m_2} \sum_{i=1}^{n} \sum_{j=I}^{n} w_{ij} z_i z_j = \frac{1}{S_0} I$

L'expression $\sum_{i=1}^{n} w_{ij} z_j$ représente la valeur moyenne de la variable d'intérêt dans le voisinage de le zone *i* (plus exactement de l'écart de la variable d'intérêt à la valeur moyenne sur le périmètre d'étude).

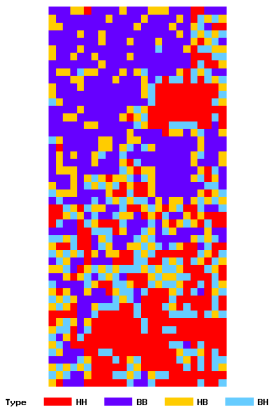
Le graphique de Moran (graphique 2.1) est tout simplement la représentation des couples de points : valeur dans la zone, valeur dans le voisinage. Ce graphique de Moran permet de constituer quatre secteurs, appelés habituellement :

- HH valeurs élevées dans un environnement élevé ;
- HB valeurs élevées dans un environnement bas ;
- BH valeurs basses dans un environnement élevé ;
- BB valeurs basses dans un environnement bas.

Le graphique ci-dessous représente le diagramme de Moran réalisé à partir de l'exemple ci-dessus de la carte 2.1. Si on effectue une régression de la valeur du voisinage sur la valeur de la zone, on obtient pour paramètre estimé 0,07334, ce qui, corrigé par S_0/N redonne l'indicateur global de Moran.



Graphique 2.1



Carte 2.5

Les calculs théoriques permettent d'arriver à des expressions de l'espérance et de la variance des indicateurs Avec les conventions d'écriture utilisées habituellement dans la littérature, n obtient les expressions suivantes :

$$E(I_i) = -\frac{\sum_{j=1}^{n} w_{ij}}{n-1}$$

Dans cette expression, le numérateur devient 1 lorsque la somme en ligne des pondérations vaut 1.

L'expression de la variance n'a pas cette belle simplicité :

$$Var(I_i) = \frac{w_{i(2)}(n-b_2)}{(n-1)} + \frac{2w_{i(kh)}(2b_2-n)}{(n-1)(n-2)} - E(I_i)^2$$

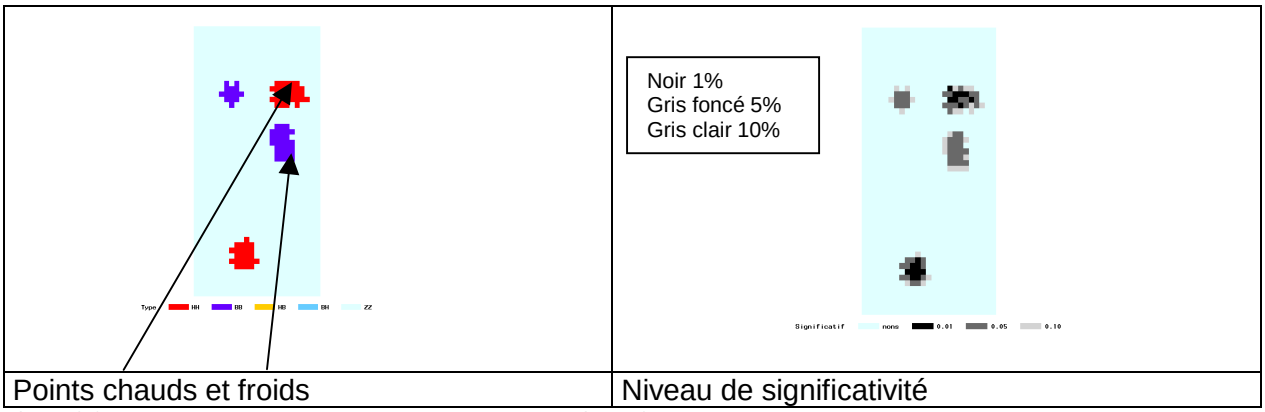
avec $b_2 = \frac{m_4}{m_2^2}$, sachant que $m_r = \sum_{i=1}^{n} \frac{z_i^r}{n}$

$$w_{i(2)} = \sum_{j=1}^{n} w_{ij}^{2}$$

$$w_{i(kh)} = \sum_{k=}^{n} \sum_{h=1}^{n} w_{ik}\, w_{ih}$$

Getis et Ord suggèrent (classiquement) d'utiliser un test basé sur la normalité (hypothétique) de : $Z(I_i) = \dfrac{I_i - E(I_i)}{\sqrt{Var(I_i)}}$.

Les tests permettent de faire apparaître les zones qui se distinguent de façon significative de leur environnement. On trouvera ci-dessous les zones les plus significatives, ainsi que la carte faisant apparaître les points « chauds » et « froids ».



Points chauds et froids — **Carte 2.6**

Niveau de significativité — **Carte 2.7**

2-3-3 Indicateur de Geary local

Des LISA calculés dans une optique « Geary » ont été également proposés par Anselin (1995). Ils sont définis de la façon suivante :

$$c_i = \frac{n}{\sum_{i=1}^{n} z_i^{2}} \sum_{j=1}^{n} w_{ij} \left(z_i - z_j \right)^{2}$$

On a donc : $\sum_{i=1}^{n} c_i = \dfrac{2nS_0}{n-1} c$.

L'espérance et la variance peuvent être approchés comme suit :

$$E(c_i) = \frac{2n\sum_{j=1}^{n} w_{ij}}{n-1} \text{ et } Var(c_i) = \frac{n}{n-1}\left(w_{i\bullet}^{2} + w_{i(2)}\right)(3 + b_2) - E(c_i)^{2}$$

Dans la littérature, on trouve de nombreux exemples d'indicateurs de Moran locaux, mais pratiquement jamais d'exemples d'indicateurs de Geary locaux. On n'en donnera pas d'illustration numérique dans ce document.

3 - Configurations spatiales de points et ratios de densité

Ce paragraphe vise à donner une petite « teinture » sur les configurations de point, et introduire quelques notions utiles pour la compréhension de ce qui va suivre.

3-1 Généralités

Un processus spatial est défini de façon générale comme :

$$\{Z(s); s \in D \subset R^2\}$$

D désigne le domaine d'étude qui est une portion du plan.

Z(.) désigne une variable aléatoire, le processus d'ensemble étant la réalisation de chacune de ces variables aléatoires en un certain nombre de points.

Si le processus est un processus de nature géostatistique, comme la température, une réalisation du processus pourra être l'ensemble des températures mesurées en $s_1,\ldots,s_n$. De façon standard, s désigne en statistique spatiale le couple (x,y) de coordonnées. Un changement du dispositif expérimental produirait une autre réalisation du processus. Les données géostatistiques pourraient être mesurées en tout point du domaine.

Dans le cas de données « de surface », la grandeur mesurée est associée à une région définie d'un domaine d'étude (la commune dans la région par exemple).

Dans le cas des processus ponctuels, la variable Z est en quelque sorte « dégénérée », puisqu'elle ne traduit que la localisation de l'unité statistique au point s. La réalisation d'un processus ponctuel consiste en un nombre fini de localisations (localisation des arbres dans une forêt, par exemple). Cette réalisation peut être appelée semis de points.

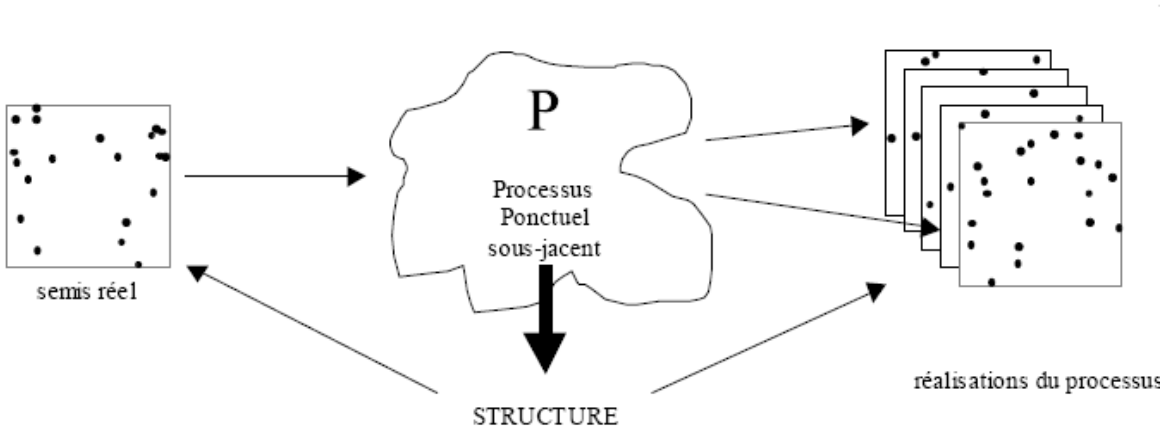
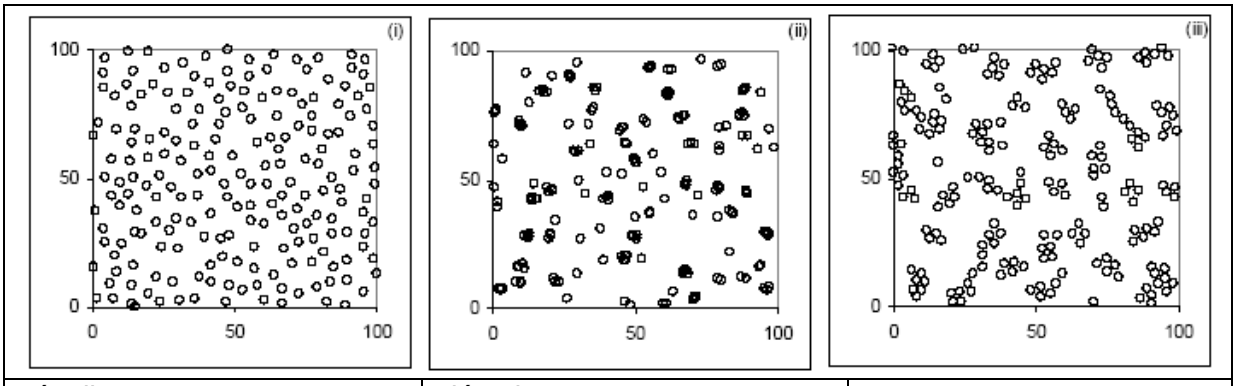


Figure 3. 1 Source : GOREAUD(1998)

Ces semis de points peuvent être de nature diverse. Il est fréquent (et pédagogiquement pertinent) de distinguer trois formes typiques de semis de points :



Régulier	Aléatoire	Amas

Figure 3.2 ***Source : GOREAUD (1998))***

La répartition « aléatoire » des points s'oppose à la répartition régulière (cas des plantations en forêt) où la présence d'un point a un effet « répulsif », et aux amas de points. La présence de ces amas peut être elle-même induite par un autre processus (présence de champignons autour de certains arbres pour rester dans le domaine forestier).

Ces processus spatiaux sont caractérisés par leurs propriétés du premier et second ordre, concepts que l'on va essayer d'introduire par rapport à la statistique inférentielle classique, ou aux processus temporels, mieux connus à l'INSEE. En statistique classique, on privilégie dans l'étude des distributions une propriété du premier ordre (moyenne) et une propriété du second ordre (variance). Ces indicateurs caractérisent le comportement moyen (précisément) et la variation autour de ce comportement moyen. On se rapproche de ce qu'on fait en statistique spatiale lorsqu'on étudie des processus temporel. Là encore, on étudie le comportement « moyen » (E(X(t)), et la corrélation temporelle (quelle relation entre ce qui se passe au temps t et au temps t+τ). On définit ainsi des processus stationnaires au second ordre.

Les processus spatiaux ont un air de famille avec ces processus temporels, mais vont s'en distinguer par une complexité plus grande (on se situe dans R^2 et non dans R). De ce fait la référence pour la stationnarité du processus ne sera plus le bruit blanc, mais une distribution spatiale particulière, appelée distribution de Poisson homogène.

3-2 Propriétés du premier ordre

La caractéristique spatiale du premier ordre est appelée intensité du processus, et caractérise un comportement moyen. Elle se définit de la façon suivante :

$$\lambda(s) = \lim_{ds \to 0} \left\{ \frac{E(N(ds))}{ds} \right\}$$

Cette expression s'interprète facilement. On se place dans un disque de rayon *ds*, situé autour d'un point du plan *s. N(ds)* désigne le nombre de points qui se trouvent dans ce disque. Comme on se place dans le cadre d'un processus aléatoire, on va s'intéresser au nombre de points attendus *E((N(ds)).* En faisant tendre *ds* vers 0, on obtient la valeur locale de l'intensité du processus. L'intensité λ(s) a donc la dimension d'un effectif par unité de surface. Elle traduit la plus ou moins grande propension du phénomène à se réaliser au point *s.*

Lorsque le processus est homogène, la valeur de l'intensité est constante sur tout l'espace étudié. On a, pour tout s, $\lambda(s)=\lambda$. Un processus homogène particulier, appelé processus homogène de Poisson va caractériser les processus stationnaires dans l'espace.

$$P(N(S)=n)=e^{-\lambda S}\frac{(\lambda S)^n}{n!}$$

Dans cette expression *N(S)* désigne le nombre de points dans le domaine S, λS le produit da l'intensité par la surface qui correspond au nombre « moyen » de points.

Dans tous les processus socio-démographiques, l'homogénéité sera l'exception et les processus seront mieux décrit par des processus de Poisson inhomogènes, ou des dérivés de ces processus :

$$P(N(S)=n)=e^{-\nu(S)}\frac{\nu(S)^n}{n!}$$

Dans cette formule, $\nu(S)$ désigne le nombre attendu de réalisations sur la zone S. Cette expression peut s'écrire, si on se réfère à ce qu'on a dit précédemment sur l'intensité des processus comme l'intégrale de l'intensité du processus sur la zone S :

$$\nu(S)=\int_S \lambda(s)ds$$

Un processus de ce type peut tendre à réaliser des amas (clusters dans la littérature anglo-saxonne). On ne s'étendra pas plus sur la théorie des processus spatiaux. On en retiendra l'intérêt des propriétés du premier ordre, de l'intensité du processus qui va permettre dans le paragraphe suivant de faire le pont avec les méthodes non-paramétriques d'estimation de la densité.

3-3 Propriétés du second ordre

Elles sont définies de façon générale de la façon suivante :

$$\lambda_2(s_i,s_j))=\lim_{ds_i,ds_j\to 0}\left\{\frac{E(N(ds_i)N(ds_j)}{ds_i\,ds_j}\right\}$$

Cette définition est sans doute moins immédiatement lisible que la propriété du premier ordre. On peut mieux le comprendre si on se ramène à l'autocorrélation spatiale. Si au point s_i et au point s_j, les valeurs sont simultanément fortes, la valeur sera élevée. On peut s'intéresser à ce qui se passe à une distance d de s_i. On dit que le processus est stationnaire lorsque la valeur du coefficient λ_2 est telle que $\lambda_2(s_i,s_j)= \lambda_2(s_i-s_j)$. La relation entre les réalisations en deux points de l'espace ne dépend que de leur distance.

On va pouvoir s'intéresser de cette façon à des problèmes proches de ceux que l'on étudie en séries temporelles, ou dans d'autres branches de la statistique spatiale.

La formule qui est présentée ci-dessus n'est pas maniable, et l'outillage de base pour les propriétés du second ordre est fourni par la fonction K, introduite par Ripley (1977)et ses nombreux dérivés. Les bons ouvrages, Diggle(2003), Gotway et Schabebberger(2005) donnent les justifications du passage de la fonction λ_2 à la fonction K.

Cette fonction K est définie de la façon suivante :

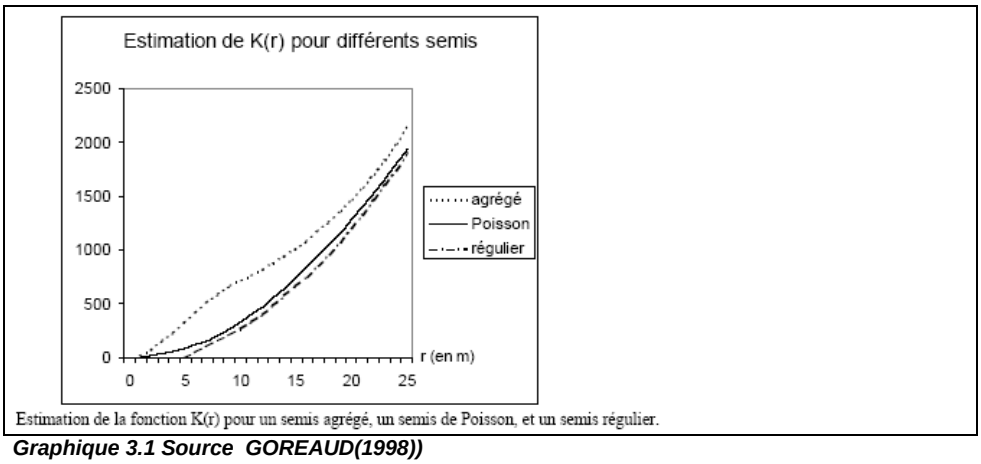
λ K(d)= E(nombre de voisins à une distance $\leq$ d de Ai),où λ désigne l'intensité du processus.

Cette fonction peut-être estimée sur un semis de points par :

$$\hat{K}(d) = \frac{1}{\hat{\lambda}} \frac{1}{n} \sum_{i=1}^{n} \sum_{j \neq i} 1_{ij}$$ où $\mathbf{1}_{ij}$ vaut 1 si distance(i,j)≤d, 0 sinon

Elle mesure, pour une distance d, le nombre de couples de points séparés par une distance inférieure à d. S'il y a des amas, par exemple, le nombre de couples sera très important à petite distance.

Dans le cas du processus de Poisson homogène, qui joue le rôle d'hypothèse nulle dans le cas des processus spatiaux, l'espérance du nombre de voisins à une distance d s'écrit simplement $\lambda\pi d^2$ (proportionnalité à la surface) et K(d)= $\pi d^{2.}$. Le graphique 3.1 permet d'illustrer ceci. Il renvoie aux trois exemples classiques de processus spatiaux présentés dans les graphiques du 2.1.



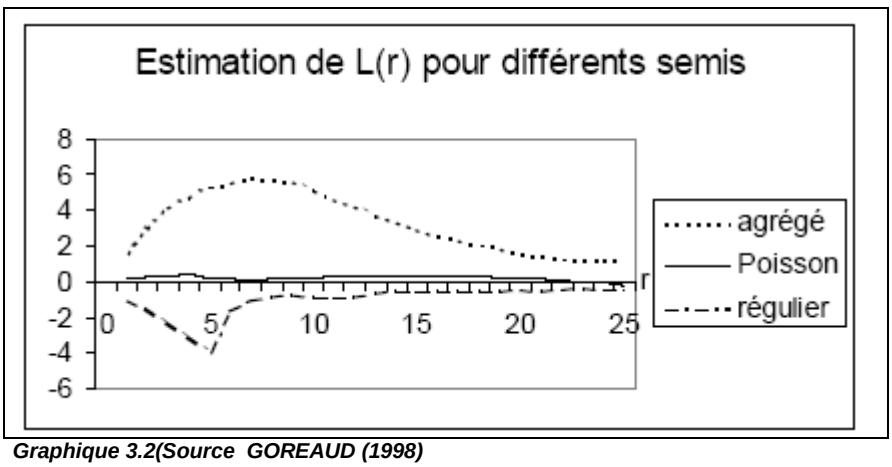
Graphique 3.1 Source GOREAUD(1998))

Lorsqu'il y a tendance à former des amas, la valeur de la fonction K est élevée aux petites distances. Dans un semis régulier, il y a au contraire des tendances répulsives, et la valeur de K est faible aux petites distances.

Un dérivé de la fonction K de Ripley, la fonction L de Besag fournit une lecture plus facile. Cette fonction s'écrit simplement :

$$\hat{L}(d) = \sqrt{\frac{\hat{K}(d)}{\pi}} - d$$

Les représentations des trois processus précédents sont les suivantes :



Graphique 3.2(Source GOREAUD (1998)

L'avantage de cette représentation est d'avoir une droite pour le « bruit blanc spatial » et d'avoir une lecture plus simple : au dessus de 0 pour les clusters, en dessous pour les processus réguliers.

Cette approche a été très féconde. Elle se prolonge aux processus bivariés et donne lieu à des indicateurs « intertypes » qui permettent d'étudier l'association spatiale entre deux processus (exemple : deux essences d'arbres dans une fôret).

Des utilisation pour des phénomènes économiques ont été proposée par Duranton et Overman 2004), ainsi que par Marcon et Puech 2004). Cette dernière sera présentée rapidement dans le chapitre 6.

3-4 - Densité de probabilité et intensité d'un processus

Les programmes développés au sein de la Division « études territoriales » de l'Insee font avant tout référence aux méthodes de statistique non paramétrique. Dans ce paragraphe elles sont introduites de façon « naturelle » dans le cadre des processus ponctuels, en présentant le lien entre l'intensité d'un processus et la fonction de densité de probabilité d'une distribution spatiale bivariée.

Les spécialistes de la statistique spatiale ont montré l'équivalence entre estimation de la densité et estimation de l'intensité (voir par exemple Diggle et Marron (1989) .

« Kernel smoothing is an attractive method for the nonparametric estimation of either a probability density fonction or the intensity function of a nonstationary Poisson process(..).Another benefit is that this duality between problemes makes it clear how to apply the well-developped asymptotic methods for understanding density estimation in the intensity setting".

On reviendra plus avant sur les méthodes non paramétriques dans le paragraphe 3.5, en ne présentant ici que quelques justifications intuitives du lien entre intensité et densité. Il suffit de retenir ceci : en pratique, l'estimateur de l'intensité spatiale est obtenue par la méthode de l'estimation de densité

Pour voir la parenté des méthodes, on considérera que l'on peut voir le processus spatial comme n réalisations d'un processus aléatoire ayant une distribution de probabilité f.

L'estimateur non paramétrique est le suivant, où f désigne la densité de la distribution de probabilité associée au processus :

$$\hat{f}(s) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{s - s_i}{h}\right)$$

Si on revient à la définition des propriétés du premier ordre :

$$\lambda(s) = \lim_{ds \to 0}\left\{\frac{E(N(ds)}{ds}\right\}$$

On exprime le nombre attendu de réalisation de la distribution f dans la surface A par :

$$E(N(A)) = \sum_{i=1}^{n} E(1_i(A) = \sum_{i=1}^{n} \int_D 1_i(A)\, f(s)ds = n \int_A f(s)ds$$

où $1_i(A)$ est une indicatrice d'appartenance à A

On en déduit, d'une façon sans doute hâtive du point de vue mathématique, que : $\lambda(s)=n\, f(s)$. On retrouve bien le fait que l'intensité renvoie bien à un nombre de réalisations.

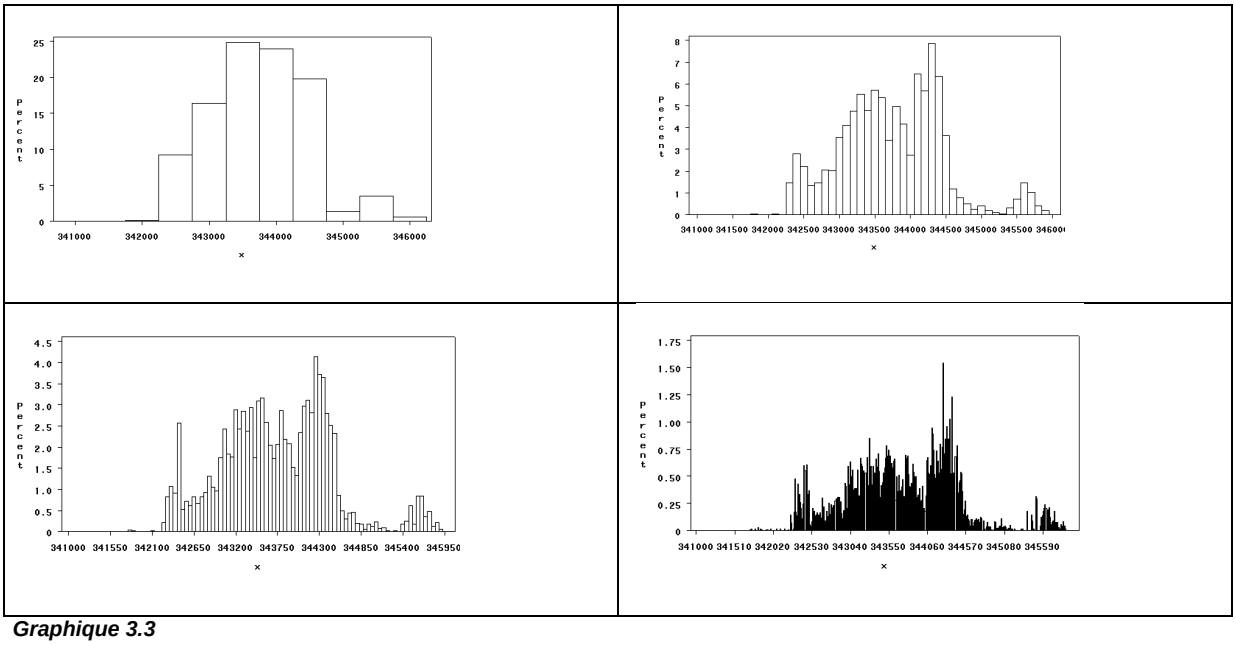
La différence entre les estimateurs tient au facteur 1/n qui apparaît dans l'estimation de la densité et non dans celle de l'intensité

Densité $\hat{f}(s) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{s - s_i}{h}\right)$ et intensité $\hat{\lambda}(s) = \frac{1}{h} \sum_{i=1}^{n} K\left(\frac{s - s_i}{h}\right)$

3-5 - L'estimation non-paramétrique de la densité

Il existe nombre de bons ouvrages comme Härdle(1998), Delecroix (1997), Silverman(1986), pour n'en nommer que quelques uns, qui permettent d'approfondir ces méthodes, en particulier pour tout ce qui concerne les problèmes de précision des estimateurs. Rathelot et Sillard (2010) en présentent les grandes lignes de façon rigoureuse mais plus concise.

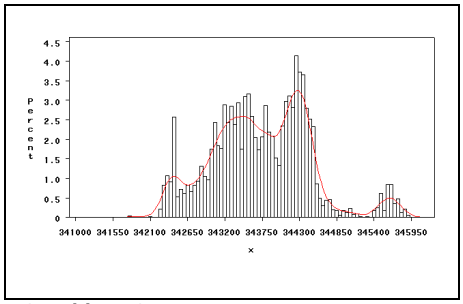
Une présentation rapide et plus intuitive de ces méthodes, peut être faite en partant du cas unidimensionnel et du « lissage de l'histogramme », suivant ainsi Härdle(1998).



Graphique 3.3

De nombreuses représentations sont possibles, certaines étant « meilleures » que d'autres.

On cherche à obtenir, au lieu de l'histogramme, une représentation continue, plus lisse de la distribution utilisée :



Graphique 3.4

Härdle (1998), donne une construction assez intuitive des estimateurs à noyaux. Il commence par donner une formule de l'histogramme, correspondant à une représentation en escalier.

L'histogramme peut s'écrire, sachant que le domaine d'étude a été divisé de la façon suivante, en considérant une origine x_0, ,comme la suite des B_j, chacun des intervalles étant de taille fixe, de la forme :

$B_j = [x_0 + (j-1)h; x_0 + jh]$ j étant un entier relatif .

Ce découpage en intervalle permet une écriture formelle de l'histogramme comme :

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n}\sum_{j} 1(X_i \in B_j)1(x \in B_j)$$

Les 1 désignent des variables indicatrices Les Xi désignent les observations, le point x le point auquel s'effectue l'estimation.

On peut ainsi commencer à introduire l'estimateur à noyau :

L'histogramme permet d'estimer f(x) comme :

$$\frac{1}{nh} Card\{\text{observations qui tombent dans un petit intervalle contenant x}\}$$

Card{.} désigne la taille de l'ensemble

C'est ce qui se passe avec l'histogramme de fenêtres fixe h.

On a un meilleur estimateur si on se plaçait en x et si l'on estimait f(x) comme

$$\frac{1}{nh} Card\{\text{observations qui tombent dans un petit intervalle autour de x}\}$$

L'intervalle est cette fois-ci de taille 2h, et on peut écrire un nouvel estimateur :

$$f_h(x) = Card\{X_i \in [x-h; x+h]\}$$

On a, sans le dire, introduit une fonction « noyau », (de l'allemand kernel), ce qui justifiera la notation K pour ces fonctions d'une manière générale, on pourra écrire successivement :

$$\hat{f}_h(x) = \frac{1}{2nh}\sum_{i=1}^{n} 1\left(\left|\frac{x-Xi}{h}\right| \le 1\right)$$

soit

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} \frac{1}{2} 1\left(\left|\frac{x-Xi}{h}\right| \le 1\right)$$

et avec la notation que l'on généralisera ensuite $K(u) = \frac{1}{2} 1(|u| \le 1)$

$$\widehat{f_h}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - Xi}{h}\right)$$

On a encore un estimateur acccordant le même poids pour l'estimation à toutes les observations qui se situent entre x-h et x+h (noyau uniforme).

Il est assez « naturel » de rechercher des noyaux qui vont accorder un poids différent aux observations selon qu'elles se situent plus ou moins loin du point d'estimation.

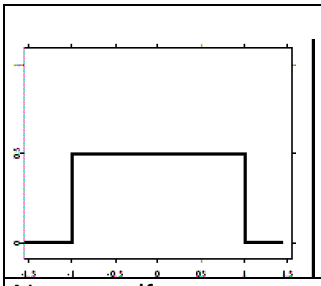
La fonction noyau peut être vue comme une fonction générant localement des poids. Les fonctions noyau doivent vérifier : $\int_R K(u)du = 1$. Dans le cas du noyau uniforme, les poids sont constants.

Les plus couramment utilisées sont :

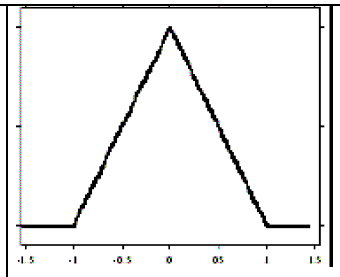
Noyau	**Forme fonctionnelle**
Uniforme	$\frac{1}{2} 1(\|u\| \le 1)$
Triangle	$(1 - \|u\|) 1(\|u\| \le 1)$
Epanechnikov	$\frac{3}{4}(1 - u^2) 1(\|u\| \le 1)$
Biweight	$\frac{15}{16}(1 - u^2)^2 1(\|u\| \le 1)$
Gaussien	$\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} u^2\right)$

Tableau 3 .1

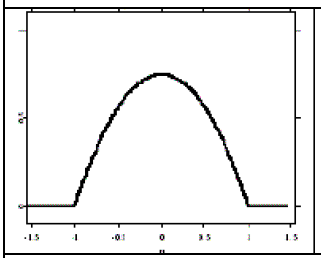
Quelques représentations graphiques



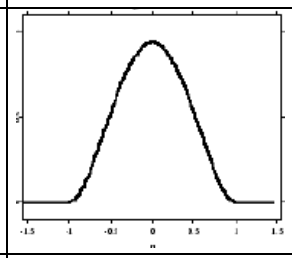
Noyau uniforme



Noyau triangulaire



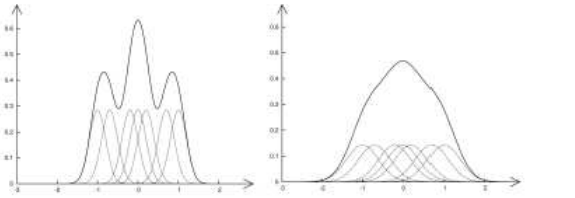
Noyau d'Epanechnikov



Noyau biweight

Graphique 3.5 Extrait de Härdle(1998)

Ces noyaux sont tous symétriques et sont à valeur sur un intervalle [-1 ;+1] à la notable exception (entre autre parce qu'il est utilisé par SAS) du noyau gaussien.



Représentation comme une somme de bosses

Figure 3.3

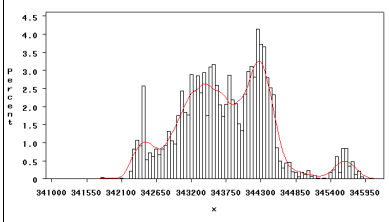
Une façon de considérer les choses et qui renvoie à la figure ci-dessus, désignée dans la littérature comme « sum of bumps », est de voir le noyau comme un opérateur qui étale la mesure ponctuelle sur un intervalle centré autour du point d'observation. Elle remplace un pic par une bosse, la fonction de densité estimée étant la somme de ces bosses.

On peut voir, à partir de $\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x - Xi}{h})$ que l'estimation produite à partir des points d'observation dépendait de deux choses, les X_i et le nombre d'observations étant fixés :

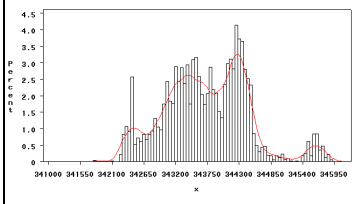
- la nature de noyau ;
- la taille de la fenêtre.

Les influences respectives de ces deux paramètres sont très différentes. Il existe un large consensus dans le monde de la statistique non paramétrique pour considérer que la forme fonctionnelle, à partir du moment où elle a une allure raisonnable, n'a pas une grande influence sur les résultats produits.

Le graphique 3.6 montre les résultats obtenus respectivement à l'aide d'un noyau assez sommaire (noyau triangulaire) et par un noyau plus élaboré (noyau biweight) : on est ici dans l'épaisseur du trait.



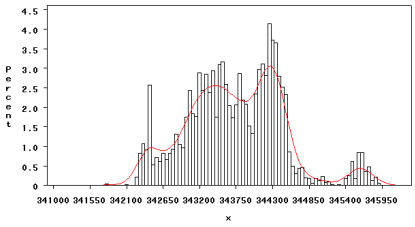
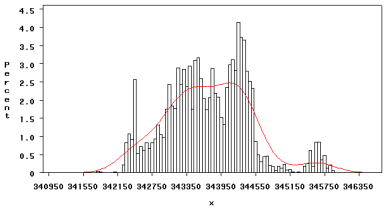
Triangulaire



Biweight

Graphique 3.6

Il n'en va pas de même pour l'influence de la fenêtre, qui constitue le problème en matière d'estimation non-paramétrique.

	
Lissage par défaut dans SAS	Bande passante deux fois supérieure à la précédente

Graphique 3.7

La littérature sur la fenêtre optimale de lissage est une des plus abondante de la statistique non paramétrique, mais avec au final des résultats limités. En général, les critères d'optimalité ne peuvent être produits que si l'on suppose aux données utilisées de bonnes propriétés, ce qui est rarement le cas dans les exemples réels.

De ce fait, on en est réduit à l'utilisation conjointe de quelques résultats théoriques de portée restreinte, par le caractère contraignant des hypothèses introduites, et d'une bonne dose d'empirisme.

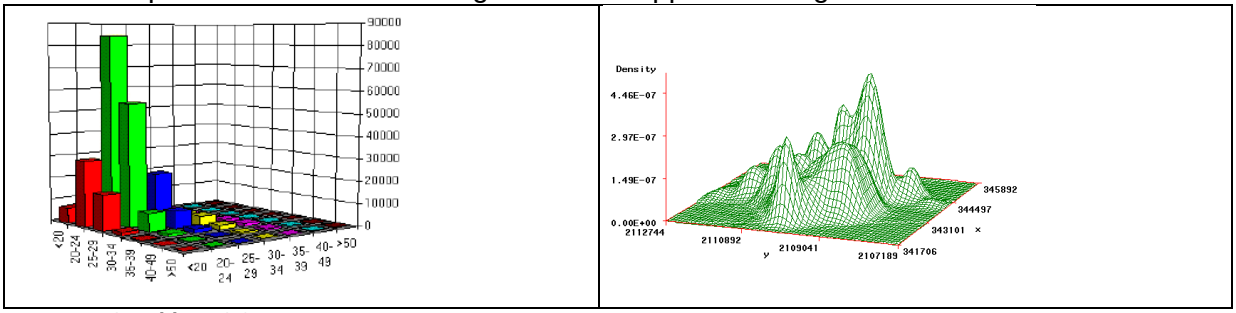
Le lecteur intéressé pourra trouver dans les ouvrages cités en référence les résultats relatifs aux propriétés statistiques des estimateurs à noyau de la densité (biais, mean squared error (MSE), mean integrated squared error (MISE)…). Les calculs sont assez lourds (développements de Taylor, intervention dans les formules des dérivées première et seconde de la fonction de densité et donc nécessité d'approcher ces dernières…). Ils ne sont pas utilisés dans tout ce qui suit.

Ce sont ces calculs qui permettent dans un cadre gaussien de calculer dans le cadre unidimensionnel la taille optimale de la fenêtre, la « rule-of-thumb » de Silverman(1986) et qui conduit à $\hat{h}_0 = 1.06\hat{\sigma} n^{-\frac{1}{5}}$.

Dans cette formule $\hat{\sigma}$ désigne l'écart-type de la distribution de X_i.

Les données unidimensionnelles permettent d'exposer les résultats de façon assez lisible et pédagogique. L'estimation de densité s'étend aux données bivariées. Cette transposition à deux dimensions ne va pas sans poser quelques problèmes théoriques nouveaux, qui ne seront qu'effleurés ici.

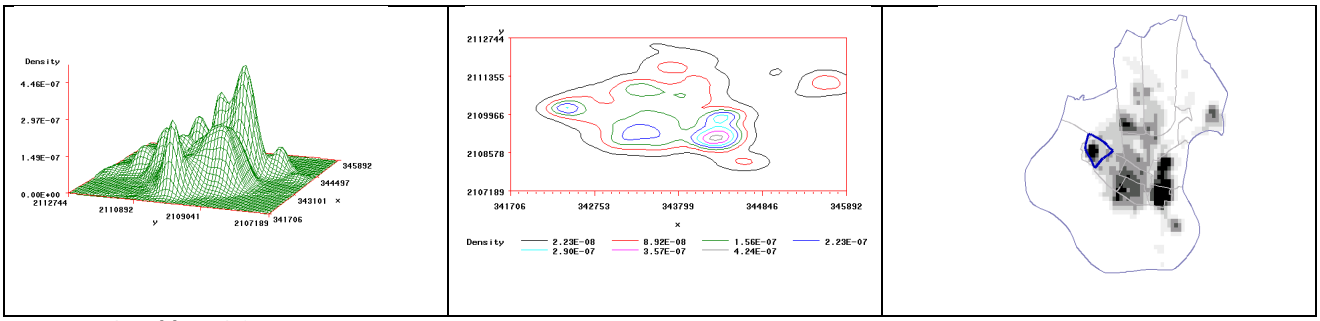
Le pendant bivarié de l'histogramme est appelé stéréogramme



Graphique 3.8

Au lieu d'estimer une courbe, comme dans le cas de l'histogramme on estime une surface. Cette surface peut être représentée par une image à trois dimensions,

par des représentations sous forme de courbes de niveau, ou par des dégradés de couleur, ce qui sera en général la règle dans les productions de l'Insee.



Graphique 3.9

La mise en œuvre effective des calculs présentés dans ce document est effectuée à l'aide du logiciel SAS®, l'outil central étant la procédure KDE, qui offre une puissance de calcul assez impressionnante associée à une syntaxe des plus simples.

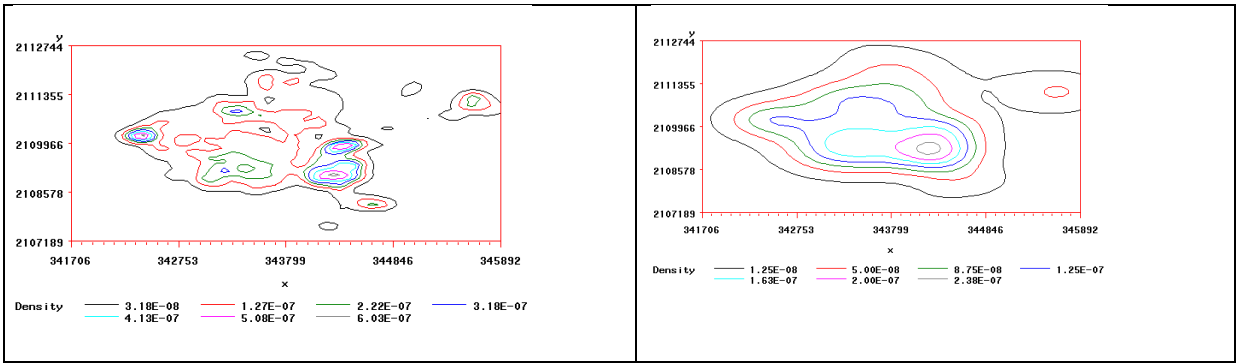
La procédure KDE utilise un noyau gaussien, qui prend donc en compte toutes les observations, la pondération devenant rapidement infinitésimale.

$$\frac{1}{2\pi h_x h_y}\ \exp\left(-\frac{\left(\frac{x}{h_x}\right)^2+\left(\frac{y}{h_y}\right)^2}{2}\right)$$

Il ne s'agit ici que d'une pseudo-fenêtre, puisque toutes les valeurs sont prises en compte.

Avec tout une batterie de bonnes hypothèses, les pseudo-fenêtres optimales sont les suivantes : $h_x = \hat{\sigma}_x n^{-\frac{1}{6}}$. (respectivement h_y).

Le caractère plus ou moins lisse de l'estimateur est géré par un paramètre appelé multiplicateur de bande passante (BWM en anglais et en SAS)



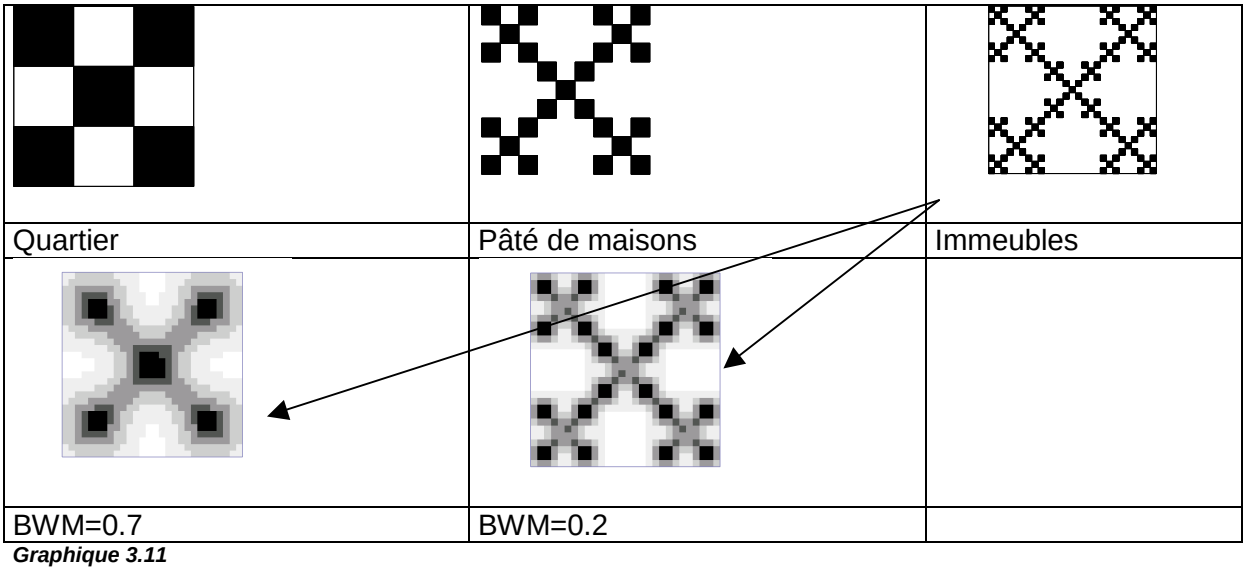
Bande passante standard divisée par 2	**Bande passante standard multipliée par 2**

Graphique 3.10

Le support carroyé permet d'avoir une représentation simplifiée de la répartition spatiale de la population étudiée. Les programmes du kit %KDE, conçu à la DET de l'Insee, et relatifs à l'estimation simple de densité associent à la procédure SAS la gestion des fonds de carte et des carroyages. On se référera à Lipatz (2009).

Il est apparu assez rapidement, en examinant les sorties cartographiques sur des villes bien connues, que le paramètre par défaut de SAS (BWM=1) ne convenait pas. Empiriquement, on peut dire que la valeur proposée par SAS masque les hiérarchies internes en cherchant une distribution à un seul mode.

Une bande passante de 0.5 est mieux adaptée à une configuration spatiale qui a l'allure (très simplifiée) suivante : elle fait apparaître une autosimilarité dans les structures, modélisée par un tapis de Sierpinski (figure fractale utilisée en modélisation urbaine).On peut en trouver une présentation rapide dans Tannier et Pumain (2005).



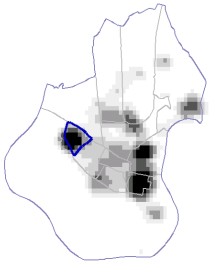
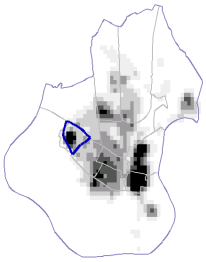
Graphique 3.11

Les communes françaises ont plusieurs modes : la dimension fractale est voisine de 1,6 correspondant à une hiérarchie marquée de type alternance de zones peuplées et inoccupées.

Dans le kit de programme de l'Insee, c'est la macro %kde qui réalise tout ceci. Elle permet de paramétrer un certain nombre de grandeurs, et surtout de gérer les bases de données cartographiques (IRIS, ZUS…) pour pouvoir fournir de façon rapide des représentations cartographiques.

3-5- Des densités aux ratios de densité

Les méthodes présentées ci-dessus permettent de simplifier la représentation d'une population dans l'espace. Cela présente une utilité propre, mais ce n'est pas suffisant pour déterminer des zones de surreprésentation. En effet tous les phénomènes que l'on étudie sont liés à la population et, de ce fait, la comparaison des densités estimées fait apparaître une ressemblance globale, et quelques différences peu lisibles. La répartition des bas revenus en fournit un exemple. Les bas revenus sont importants - c'est trivial - là où la population est importante. Ce qu'on cherche à savoir, ce sont les lieux où lls sont relativement plus importants.

	
Bas revenus	Population

Carte 3.1

Pour traiter cette question en restant dans le cadre défini précédemment, l'idée qui vient est de faire le ratio des densités. C'est cette opération qu'il s'agit de justifier. En effet, ce n'est pas parce que l'opération semble aller de soi qu'elle est mathématiquement justifiée.

Cette question n'est pas nouvelle et il y a déjà nombre d'années que les statisticiens spatiaux travaillant dans le domaine de l'épidémiologie y sont confrontés. Les épidémiologistes rapprochent les populations présentant une pathologie des populations à risque, ceci dans le but de faire apparaître des différences spatiales dans la propension à la développer.

Cette approche peut être de l'ordre d'un simple constat (on développe plus la maladie dans telle zone géographique) ou faire l'objet d'une modélisation plus poussée (explication faisant intervenir des facteurs explicatifs comme l'émission de polluants). Les épidémiologistes parlent de « relative risk », et ils utilisent la modélisation suivante :

$$\lambda(s) = \vartheta(s) * \lambda_0(s)$$

où λ et λ_o désignent les intensités des pathologies et des populations à risque, tandis que θ désigne le risque relatif. Cette modélisation fait apparaître de façon naturelle le rapport des intensités **:**

$$\theta(s) = \frac{\lambda(s)}{\lambda_0(s)}$$

Ce rapport est un nombre sans dimension, contrairement au numérateur et au dénominateur.

Cette modélisation conduit à s'intéresser aux rapports des densités, puisqu'on a vu en 3.4 le lien entre intensité et densité. Est-il fondé d'utiliser le ratio de deux densités ?(ce qui est différent du rapport de deux variables aléatoires).

Que cherche-t-on en effet ? Si on reprend l'exemple précédent, on cherche les zones dans lesquelles les ménages à bas revenus sont plus représentés que la population totale. S'il n'y avait aucune variation spatiale, la densité des ménages à bas revenus serait partout la même que celle des ménages, tous niveaux de revenus confondus. Ou pour le dire d'une autre façon, relativement, conditionnellement à la répartition des ménages, celle des ménages à bas revenus serait uniforme.

Dans les problèmes que l'on traite, la sous-population est incluse dans la population de référence, ce qui facilite le traitement statistique. Dans ce cadre, on peut proposer le traitement probabiliste présenté ci-dessous.

Un peu plus de probabilités

Si on formalise un peu autrement, on définit dans un espace $D \subset R^2$ les mesures m_A et m_B, dont les densités par rapport à la mesure de Lebesgue dans le plan sont f_A et f_B, où A et B désignent population et sous-population Pour un sous-ensemble F de D, on aura :

$$m_A(F) = \int_F f_A(s)\, d\mu(s)$$

$$m_B(F) = \int_F f_B(s)\, d\mu(s)$$

Si actifs et chômeurs se répartissaient de la même façon, on aurait $f_A = f_B$, et pour tout F, $m_A(F) = m_B(F)$. La densité des chômeurs est calculée par rapport à la mesure de Lebesgue. On pourrait essayer de l'exprimer par rapport à la mesure des actifs m_A, et essayer d'obtenir une expression de la forme :

$$m_B(F) = \int_F g(s)\, dm_A(s)$$

Dans le cas de la répartition uniforme, celle ou g(s)=1, on a $m_B(F) = \int_F dm_A(s) = m_A(F)$

Sous de bonnes conditions, et en particulier que $m_A(F)=0 \Rightarrow m_B(F)=0$, le théorème de Radon-Nikodym permet de justifier l'existence de la densité g, par rapport à la mesure m_A Comme on a $dm_A(s)=f_A(s)d\mu(s)$:

$$m_B(F) = \int_F f_B(s)\, d\mu(s) = \int_F g(s) dm_A(s) = \int_F g(s)\, f_A(s) d\mu(s)$$

et donc $f_B(s) = f_A(s) g(s)$ *et* $g(s) = \frac{f_B(s)}{f_A(s)}$

g(s), rapport de deux densités est aussi une densité, mais par rapport à une mesure différente.

On peut aussi proposer une formalisation basée sur le conditionnement. Cette démarche est plus intuitive, mais elle est difficile à expliciter en terme de densité. On l'approchera par l'espérance conditionnelle.

A les notations précédentes, on calcule le nombre d'individus de la population B attendus dans la zone F. Si on appelle n_A la taille de la population totale et n_B celle de la sous-population, 1_F l'indicatrice d'appartenance à la région F, $\mathbf{1}_B$ l'indicatrice d'appartenance d'un individu de la population A à la population B.

Le nombre attendu d'individus de la population B (PopB) dans la zone F peut être écrit de deux façons :

- à partir de la loi f_B

$$E_B(Pop_B(F)) = E(\sum_1^{nB} 1_F(s)) = n_B \int_{R2} 1_F(s)\, f_B(s) ds = n_B \int_F f_B(s) ds$$

- à partir de la loi de f_A

$$E_A(Pop_B(F)) = E(\sum_1^{nB} 1_F(s) 1_B(s)) = n_A \int_{R2} 1_F(s) 1_B(s)\, f_A(s) ds = n_A \int_F 1_B(s)\, f_A(s) ds = n_B \int_F \frac{1_B(s) n_A}{n_B} f_A(s) ds$$

En identifiant les deux expressions du nombre d'individus attendus dans la zone F, on voit que le rapport des densités $\frac{f_B}{f_A} = \frac{n_A}{n_B} 1_B(s)$ traduit bien la présence relative de la population B conditionnellement à sa position s.

L'important, après ces considérations mathématiques un peu longues est de savoir que le ratio des densités a un sens et qu'il traduit la sur-représentation locale de la sous-population étudiée.

Remarque : l'indicateur de divergence de Kullback-Leibler fait aussi intervenir un rapport de densité, plus exactement le logarithme de ce ratio de densité :

$$KL(f,g) = \int_D f(s) Log \frac{f(s)}{g(s)} ds$$

Lorsque les deux variables sont distribuées de la même façon, le rapport f/g est égal à 1 et la divergence est nulle. Cet indicateur mesure une distance entre les distributions, comme le chi-2 ou la distance de Hellinger.

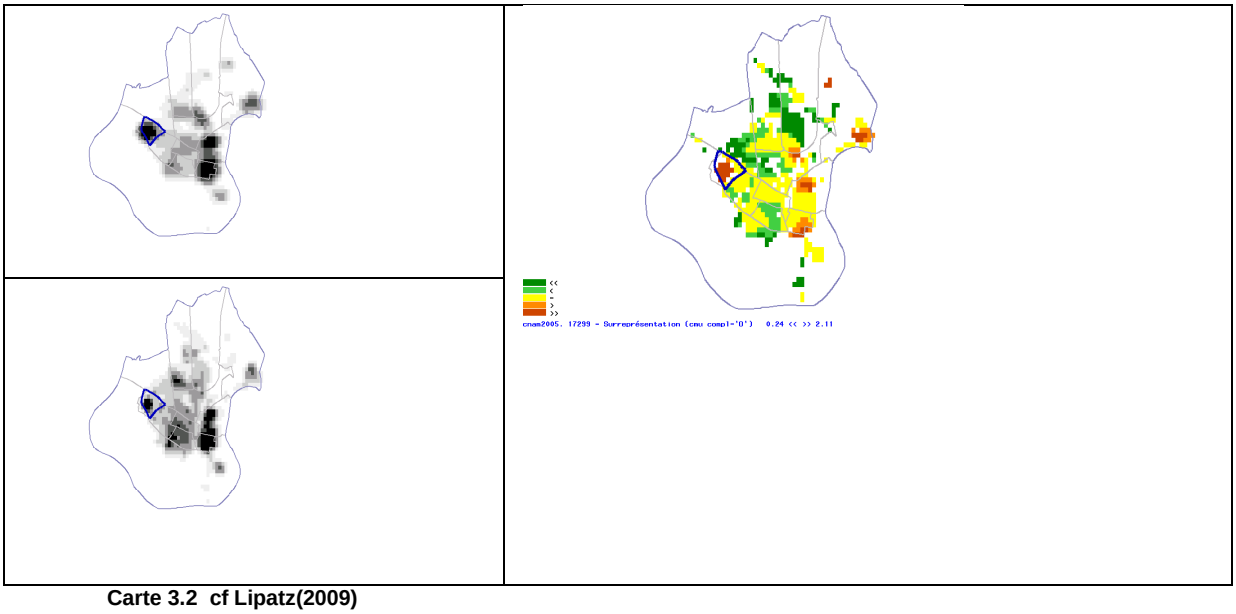
3-6- Du ratio théorique au ratio estimé

On retrouve bien entendu tous les problèmes d'estimation qui ont été rencontrés dans l'utilisation des estimateurs à noyaux, avec quelques complications supplémentaires.

Le choix du noyau n'est pas décisif. Par contre, le choix de la fenêtre est toujours décisif. La fenêtre optimale n'est pas la même au numérateur et au dénominateur, puisqu'il n'y a pas le même nombre d'observations qui rentrent en jeu et que les variances ne sont pas les mêmes. Bailey et Gattrell (1998), par exemple, préconisent l'utilisation de la même fenêtre pour les estimations du numérateur et du dénominateur. C'est ce qui a été fait dans les procédures SAS de la DET (Lipatz (2009)).

Des problèmes numériques sont rencontrés au cours des estimations, le rapport des estimations pouvant conduire à des valeurs extrêmes. Le problème a été résolu en éliminant les points d'estimations pour lesquelles les valeurs sont trop faibles.

On trouvera ci-dessous une illustration de ce résultat, sur la commune de Niort. Le résultat n'est pas une valeur lissée du taux de chômage, mais une représentation de l'écart à une situation uniforme (répartition des chômeurs semblable à celle des actifs)

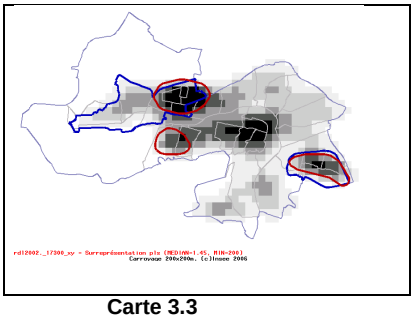


Carte 3.2 cf Lipatz(2009)

Cette carte constitue une représentation simplifiée de la ville, avec une situation « équilibrée » dans le centre-ville, un chômage plus faible dans les quartiers de peuplement proches de la ville-centre et une sur-représentation des chômeurs essentiellement dans les ZUS (contours bleus).

Les programmes mis au point à l'Insee permettent de fournir très rapidement des cartes d'étude de ce type, le temps de traitement statistique étant très inférieur au temps de constitution des fichiers de données. Les programmes ont été utilisés sur les résultats du recensement de 1999, mais ils sont utilisables sur tout fichier de données géoréférencées. Ce sont les macros %kde2 et %kde2f (selon que les données sont individuelles ou déjà agrégées) qui permettent de réaliser ces cartes.

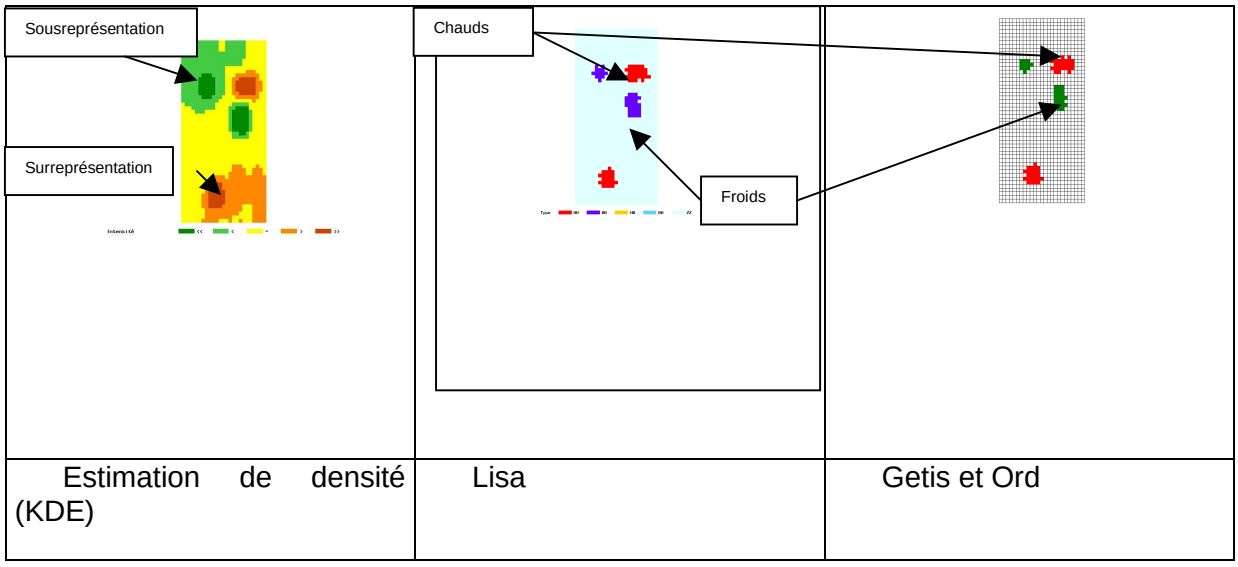
La macro %aspic spécialement conçue pour la préparation de la réponse aux questions sur le dispositif CUCS fournit une carte de synthèse. Elle fait apparaître de façon lissée les zones de sur-représentation, sous la forme de patatoïdes rouges, et la densité de la population, sous la forme de dégradés de gris (exemple de La Rochelle en carte 3.3). le descriptif de cette macro se trouve dans Lipatz (2009).



Carte 3.3

3-7 Comparaison avec les résultats des Lisa

On reprend ici l'exemple présenté en 2.2, qui a été utilisé pour la détection des points chauds et froids à partir des indicateurs de Getis et Ord et des Lisa de Moran.



Estimation de densité (KDE)	Lisa	Getis et Ord

Carte 3.3

La carte issue de l'estimation par les méthodes à noyaux identifie bien les points chauds et froids, mais dans le cas présent avec une tendance à étaler les zones de sur ou sous-représentation.

La méthode des ratios de densité a de gros avantages en terme de mise en œuvre, puisqu'elle ne nécessite que la connaissance des coordonnées des points et qu'il n'y a pas à gérer de voisinages comme dans le cas des indicateurs surfaciques. On est ici dans lun cas d'école où les méthodes surfaciques fonctionnent particulièrement bien (répartition plutôt régulière de la population sans « trous », régularité de la maille de référence). A l'inverse l'estimation de densité est faite à partir de données déjà agrégées.

La comparaison a aussi été introduite pour montrer une des limites des méthodes non paramétriques, à savoir la difficulté à mettre en évidence des seuils et à les justifier par des tests. Ls indicateurs surfaciques permettent la construction de tests (formulations analytiques ou replication d'échantillons) tandis que les ratios de densité sont calculés à partir d'information déjà agrégée.

Pour montrer la difficulté, on rappellera que les estimateurs de la densité présennt des biais. On peut en calculer la variance. Celle-ci s'exprime de façon assez complexe On en trouve des résultats détaillés dans Hardle (1998). Dans le cas le plus simple, on a $\mathrm{var}\left(\hat{f}(s)\right)=\frac{1}{n}\|K\|^2{}_2\, f(s)$ avec : $\|K\|^2{}_2=\int K^2(s)ds$

Le fait de pouvoir calculer des estimateurs de la variance de la densité au numérateur et au dénominateur ne donne pas la variance du ratio et encore moins un test. Les travaux théoriques sur la comparaison des distributions dans un espace à deux dimensions (analogues des tests de type Cramer-Von Mises ou Kolmogorov) sont embryonnaires.

Quand ils donneraient une information sur une différence significative entre les fonctions de densité, il faudrait encore qu'ils puissent informer sur les régions qui contribuent à cette différence.

Kelsall et Diggle(1995) suggèrent un test, basé sur des replications aléatoires dans le cas unidimensionnel. On pourrait concevoir des tests conçus de la façon suivante : si P est la proportion globale de notre population à risque, on simule (1000 fois par exemple) un échantillon ou l'on conserve la population totale Pop(s), mais ou la population à risque R(s) est la réalisation d'une loi binomiale $\mathcal{B}(Pop(s),P)$. A partir de ces similations, on peut construire un intervalle de confiance, et voir dans quelles parties du plan la sur-représentation est significativement différente de 1.

Cette méthode reste complètement empirique, et n'a pas de fondements mathématiques. De plus, il faudrait pouvoir tester des seuils de sur-représentation différents de 1.

4- Extension du champ d'application

4-1 - Le principe

La facilité des traitements permise par les méthodes non-paramétriques a conduit a utiliser l'estimation de densité dans le cadre d'une partition de la population, en s'inspirant d'un article de Diggle, Zheng et Durr (2005).

Si la population P est partitionnée en K sous-populations $P_1,\ldots,P_K$, on peut calculer en tout point autant de ratios de densité (ou d'intensité) qu'il y a de sous-population.

On effectue donc K+1 estimations de densité, une pour chacune des sous-populations, $\hat{f}_k(s)$ et une pour la population d'ensemble $\hat{f}(s)$. On peut en déduire des ratios de sur-représentation, ainsi que des parts locales. Les effectifs estimés pour chacune des sous-populations sont $\hat{n}_k = n_k * \hat{f}_k(s)$, où n_k est la population de la k^{ème} sous-population sur l'ensemble du périmètre d'étude.

Cette part s'exprime de la façon suivante : $\hat{p}_k(s) = \frac{\hat{n}_k}{\hat{n}} = \frac{n_k}{n} * \frac{\hat{f}_k(s)}{\hat{f}(s)} = \frac{n_k}{n}\hat{s}_k(s)$.

4-2 Application à l'analyse des données

La prise en compte des proximités est une question que L. Lebart avait déjà abordé en 1978 dans le cadre de la Classification ascendante hiérarchique (CAH). On se réfère surtout ici aux travaux de Brigitte Escofier (1990), qui a abordé ces questions dans plusieurs articles. Elle y introduisait les notions d'analyse lissée et d'analyse des différences locales, les problèmes étant traités dans le cadre général des analyses en composantes multiples (ACM).

On se limitera ici à l'analyse d'un tableau de contingence particulier, croisant en ligne une zone géographique (Iris, carreau) et en colonne une variable qualitative (PCS, tranche de revenu). Au lieu d'analyser le tableau brut, on analyse le tableau des données estimées par les méthodes non paramétriques. On remplace n_{ik} par $\hat{n}_{ik}$, i désignant le territoire et k la sous-population. Dans l'estimation de la i^{ème} ligne interviennent tous ses voisins. Lorsqu'on utilise les estimateurs non-paramétriques à noyaux, le nombre des voisins est variable. Le tableau de contingence des valeurs lissées a quelques caractéristiques :

- les sommes en colonne restent identiques, puisque l'on répartit autrement une sous-population entre les territoires ;
- les sommes en ligne sont modifiées, et donc le poids qu'aura dans l'analyse factorielle des correspondances (AFC) chacun des territoires.

Si on veut privilégier le poids des individus, en conservant le poids initial de l'individu, on fait une simple règle de trois, qui ne modifie pas le profil. Par contre, on perd dans ce cas l'égalité des poids des colonnes.

On peut comparer les résultats obtenus par une AFC sur les revenus fiscaux localisés, sur données brutes d'abord puis sur données estimées. Les carreaux de 200 m de côté sont en ligne, les quintiles de revenu (calculés au niveau national) en colonne, le champ de l'analyse étant la ville de Rennes dans l'exemple ci-dessous. Dans l'analyse

lissée, la quasi-totalité de l'inertie est conservée sur le premier plan factoriel, alors qu'elle est de seulement 85% dans l'analyse brute. Quid des commentaires sur le troisième axe ?

Inertia and Chi-Square Decomposition

Singular Value	Principal Inertia	Chi-Square	Percent	Cumulative Percent	13 26 39 52 65
049369	024373	639955	6628	6628	**********
027369	007491	196679	2037	**8665**	********
016601	002756	72357	749	9415	***
014670	002152	56503	585	10000	**
Total	036772	965494	10000		

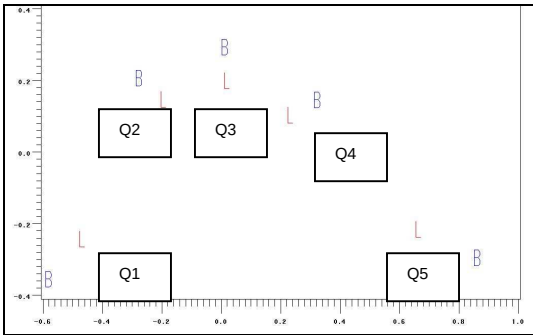
Degrees of Freedom = 7392

Inertia and Chi-Square Decomposition

Singular Value	Principal Inertia	Chi-Square	Percent	Cumulative Percent	15 30 45 60 75
038493	014817	3386E7	7712	7712	**********
019113	003653	8347860	1901	**9613**	******
006712	000450	1029446	234	9847	*
005419	000294	671131	153	10000	*
Total	019214	4391E7	10000		

Tableau 4.1

Les projections des variables sur les deux premiers axes de l'analyse sont les suivants :

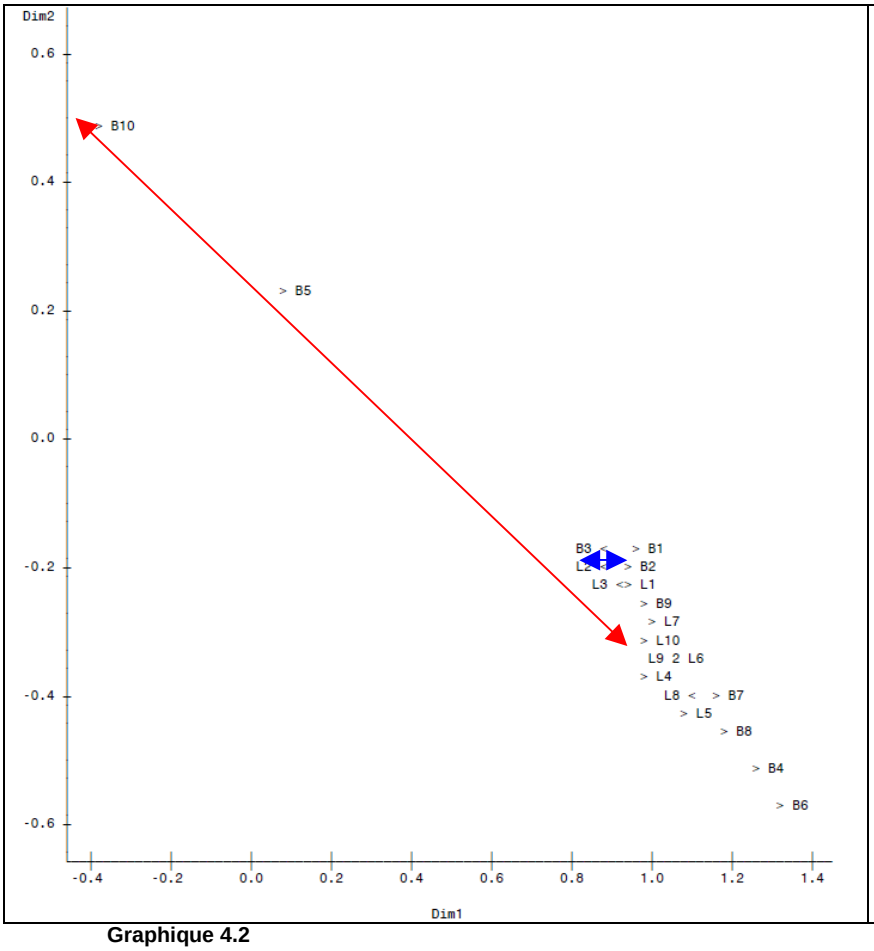


Variables

Quintile 1 → Quintile 5

En bleu les résultats bruts, en rouge les résultats lissés

Graphique 4.1



Projection des carreaux les mieux représentés dans l'analyse des données lissées et différences avec les projections des données brutes.

Dans l'analyse, les données brutes sont en observations supplémentaires.

La représentation graphique permet de montrer les points où les observations estimées diffèrent le plus des observations brutes.

Graphique 4.2

4-3 Un exemple de classification

Les profils estimés permettent de construire une typologie des carreaux, au niveau national, à partir des effectifs ventilés par quintile de revenu.

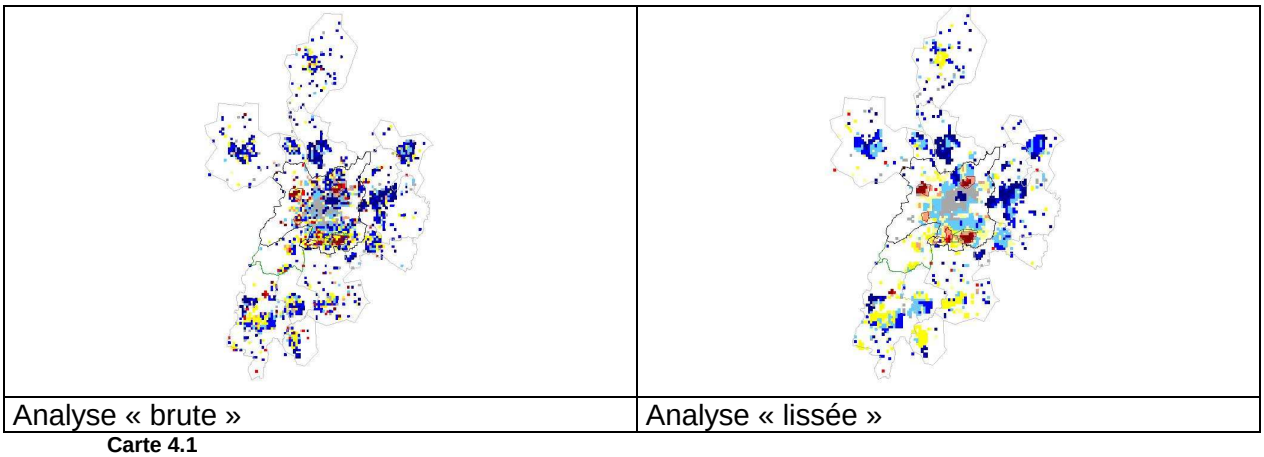
Dans cette analyse, les méthodes factorielles ont été utilisées à titre exploratoire pour repérer quelques uns des profils les plus fréquents, et déterminer un nombre indicatif de classes. A partir de ces informations, on a constitué des profils-type, plus interprétables que ceux qui sortent directement de la classification ascendante hiérarchique (CAH).

La typologie adoptée est la suivante :

P1 P2 MP R1 R2 MR MM ME MA	Ségrégé riche	**R1**	Dominante du 5°quintile
		R2	Dominante des 4°et 5°quintiles
	Ségrégé pauvre	**P1**	Dominante du 1°quintile
		P2	Dominante des 1°et 2°quintiles
	Mixte	**MR**	Proche de l'équilibre, dominante riche
		MP	Proche de l'équilibre, dominante pauvre
		ME	Equilibré
		MM	Dominante des quintiles 2,3 et 4
		MA	Atypique : dominante des deux extrêmes

Tableau 4.2

La cartes 4.1 présente les rsultats de l'analyse sur l'unité urbaine de Rennes. Elles montrent le gain en lisibilité obtenu par l'analyse « lissée ».



Analyse « brute »	Analyse « lissée »

Carte 4.1

4-3 Utilisation pour le calcul des indicateurs de ségrégation spatiale

Les indicateurs de ségrégation spatiale sont classiquement utilisés dans les études urbaines, en particulier par les sociologues et les économistes. Ces indicateurs sont très nombreux, et une présentation systématique en a été faite dans Massey & Denton (1998).

Le problème de l'utilisation de ces indices vient de ce qu'ils sont « aspatiaux », et qu'ils sont sensibles au problème de l'échiquier et à celui du MAUP. Prenons un exemple très simple, exposé dans Feitosa (2004) et qui décrit un territoire formé de 144 carreaux, regroupés en 4 grands ensembles, et peuplés de quatre sous-populations de même taille.

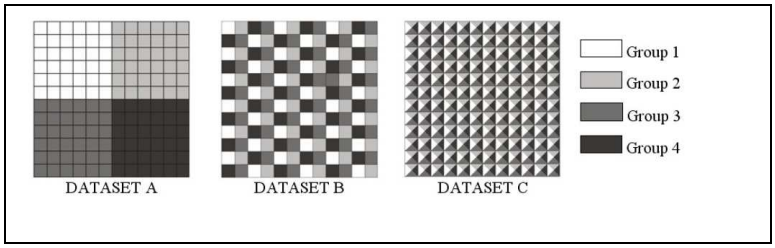


Figure 4.1

Si l'on calcule le classique indicateur d'entropie de Theil pour chacune des quatre configurations, on trouve la même valeur (Log4), malgré les évidentes différences morphologiques. La configuration A montre une forte ségrégation à grande comme à petite échelle. La C, l'absence complète de ségrégation. La B est intermédiaire puisqu'il y a ségrégation à petite échelle, celle-ci ne se manifestant pas à l'échelle de l'ensemble du territoire. Mais les indices d'entropie calculés au niveau du carreau, et qui prennent la valeur 0 ne permettent pas de distinguer les situations A et B.

On rencontre d'autres difficultés liées à l'invariance des résultats par permutation des groupes, traduisant le fait qu'un indice élevé n'informe pas sur la situation qualitative, mais on laissera ce sujet de côté.

Pour pallier à ces difficultés, des solutions ont été proposées, en particulier dans l'article séminal de Reardon et O'Sullivan (2004). On en trouve une présentation dans Le Toqueux (2004) ou Rathelot et Sillard (2010).

Reardon et O'Sullivan(2004) définissent la densité $\tilde{\tau}_p$ au voisinage d'un point et la densité $\tilde{\tau}_{pm}$ e la sous-poulation m au voisinage de ce point, puis une valeur estimée de la proportion notée : $\tilde{\pi}_{pm} = \frac{\tilde{\tau}_{pm}}{\tilde{\tau}_p}$ (notations de Reardon et 0'Sullivan).

On retrouve nos proportions estimées, qui peuvent servir à calculer des estimateurs locaux de l'entropie, de la forme :

$\hat{E}(s) = -\sum_{i=1}^{K} \hat{p}_k(s) * \log(\hat{p}_k(s))$, et à construire des cartographies de la ségrégation résidentielle.

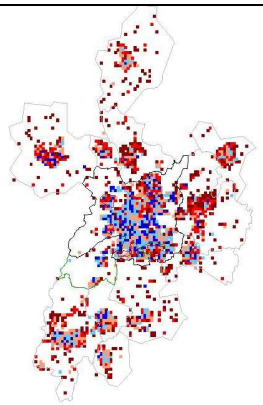
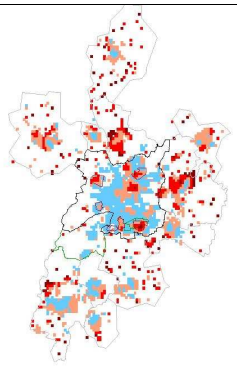
On peut aussi calculer un indice global, dans la perspective proposée par Reardon et O'Sullivan, en calculant :

$H = 1 - \frac{1}{TE} \sum_{j=1}^{T} \tau_j * \hat{E}(s_j)$, T étant le nombre de points d'estimation, τj la part de la population totale au point d'estimation s_j , E l'entropie sur la totalité du territoire.

Il faut cependant rappeler toutes les limites d'un indice global de ce type, une même valeur de l'indice global pouvant renvoyer à des morphologies très différentes de la ségrégation spatiale.

Ces proportions estimées peuvent servir pour tous les indicateurs faisant intervenir des proportions, et mesurer les différentes dimension de la ségrégation spatiale définies par Massey et Denton.

On trouve ci-dessous une illustration sur le cas de Rennes (population partagée en quintiles de revenu), l'unité d'observation étant un carreau de 200 m.

	
Entropie locale : données brutes H=0,12	Entropie locale : données estimées H (Optique Reardon O'Sullivan)=0,06

Carte 4.2

5 - La régression géographique pondérée (RGP)

Il s'agit de la francisation de la « Geographically weighted regression » introduite par Brunsdon, Charlton et Fotheringham. Elle est exposée dans l'article initial de ces trois auteurs, et dans un livre du même nom qui en donne plus de développements. On va en donner ici une présentation rapide. Cette méthode a été utilisée dans divers travaux de l'Iinsee (zones mixtes, Synthèse urbaine).

5-1- Présentation générale

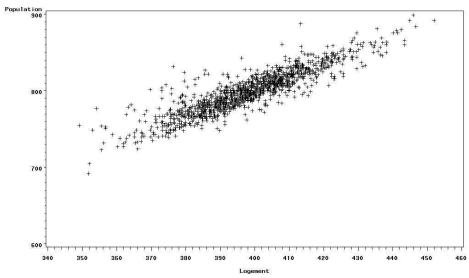
Cette méthode a été proposée par des géographes pour résoudre des problèmes liés à la non-stationnarité des processus (en milieu urbain en particulier). L'exemple donné dans l'ouvrage de référence de Brunsdon, Charlton et Fotheringham (2002) est celui du prix des logements. Quand même des facteurs explicatifs peuvent être identifiés, la façon dont ils jouent peut varier dans l'espace. L'idée finalement assez simple de la régression géographique pondérée est de faire varier les coefficients. Au lieu d'avoir des coefficients globaux, valides en tout point de référence, on a un ensemble de coefficients variant dans l'espace et traduisant la non stationnarité du phénomène que l'on étudie.

C'est en ce sens que la régression est géographique. Elle est estimée localement, à partir des variables observées dans le voisinage du point d'estimation. La pondération que l'on donne est d'autant plus importante que les observations sont proches du point d'estimation.

Appliquée à des phénomènes spatiaux, la régression linéaire classique suppose des hypothèses particulièrement fortes de stationnarité (que les mêmes stimuli provoquent les mêmes réponses en tous les points de l'espace) ; en gros, s'il y a le même nombre d'actifs, il y aura le même nombre de chômeurs, quelle que soit la localisation. La régression globale est une méthode aspatiale.

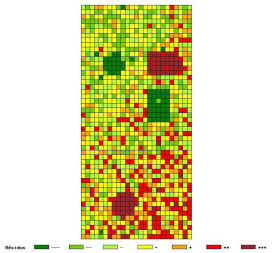
Prenons un exemple particulièrement simple. Le nombre d'habitants est lié au nombre de logements. Une régression du nombre d'habitants sur le nombre de logements traduira en moyenne le nombre d'habitants par logement sur le territoire considéré.

Un petit exemple nous servira à illustrer ceci. On part toujours de données mesurées sur un carroyage. Le lien entre le nombre de logement et la population est linéaire, comme le montre le nuage de points du graphique 5.1. Si on cartographie les résidus (carte 5.1), on peut voir que ces résidus ont une structure spatiale marquée, avec des ensembles de carreaux contigus sur lesquels on a de forts résidus positifs ou négatifs.



Nuage des points

Graphique 5.1



Cartographie des résidus

Carte 5.1

Ce résultat visuel est confirmé par une valeur très élévée de l'indicateur d'autocorrélation de Moran (0,52). Les hypothèses des moindres carrés ordinaires (MCO), qui sembleraient pouvoir être acceptées au vu du nuage de point ne sont pas satisfaites.

5-2 Théorie élémentaire de la RGP

5-2-1 Les bases

Au lieu du modèle de régression classique (aspatial) :

$$y_i = \beta_0 + \sum_{k=1}^{p} \beta_k \, x_{ik} + \varepsilon_i$$

on écrira simplement, pour chaque point d'estimation i, de coordonnées géographiques $\mathbf{s_i}$ (résumé des coordonnées x_i et y_i, et permettant de ne pas intoduire de confusion dans la notation) :

$$y_i = \beta_0(s_i) + \sum_{k=1}^{p} \beta_k(s_i) \, x_{ik} + \varepsilon_i$$

La fonction $\beta(s)$ est une fonction continue sur l'espace, qui va être calculée en un certain nombre de points. Elle va traduire la variablité locale des coefficients.

Dans le cas des moindres carrés ordinaires, l'estimateur est constant sur l'espace et s'écrit, avec les notations matricielles habituelles $\hat{\beta}(s) = \left(X'X \right)^{-1} X'y \quad \forall s$

Dans ce cas, l'estimateur de la variable d'intérêt s'écrit :
$\hat{y} = X(X^T X)^{-1} X^T y = S_0 y \quad \forall s$

L'opérateur S_0 est souvent appelé « matrice chapeau » dans la mesure où il transforme y en $\hat{y}$. Il ne dépend pas des valeurs de y.

Qu'en est-il dans la régression géographique pondérée ? Au lieu d'estimer une seule valeur du vecteur des paramètre, on va en estimer n. Chacune des n estimations est effectuée à l'aide de la méthode des moindres carrés ordinaires, pondérés cette fois en attribuant à chaque observation un poids qui décroit avec la distance au point d'observation. L'estimation au point s_i s'écrit de la façon suivante :

$$\hat{\beta}(s_i) = \left[X^T W(s_i) X \right]^{-1} X^T W(s_i) Y$$

où $W(s_i)$ désigne la matrice des pondérations au point s_i .

Il y a donc autant de matrices de pondérations qu'il y a de points d'estimation. Comme pour les « estimateurs à noyaux », on peut utiliser différentes méthodes pour le calcul des poids.

La matrice s'écrit de la façon suivante :

$$W(si) = \begin{bmatrix} w_{i1} & 0 & 0 & 0 & 0 \\ 0 & w_{i2} & 0 & 0 & 0 \\ 0 & 0 & . & 0 & 0 \\ 0 & 0 & 0 & . & 0 \\ 0 & 0 & 0 & 0 & w_{in} \end{bmatrix}$$

Les pondérations les plus simples sont de type « plus proches voisins » :

$$w_{ij} = 1 \quad si \quad d_{ij} < d$$
$$w_{ij} = 0 \; dans \, les \; autres \; cas$$

Les formes les plus couramment utilisées sont des exponentielles, utilisées dans les articles initiaux de Brunsdon (2004),

$$w_{ij} = \exp\left[-\frac{1}{2} d_{ij}^{\,2}\right]$$

ou la fonction bicarrée de Tukey

$$w_{ij} = (1 - d_{ij}^{\,2})^2 \quad si \quad d_{ij} < d$$
$$w_{ij} = 0 \; dans \, les \; autres \; cas$$

Chaque estimateur $\hat{y}_i$ de y_i peut être estimé de la façon suivante

$$\hat{y}_i = X_i \left[X^T W(s_i) X\right]^{-1} X^T W(s_i) Y$$

On peut définir pour la RGP une forme particulière de matrice chapeau S_1, telle que l'on ait :

$\hat{Y} = S_1 Y$, chaque ligne (notée r_i) de la matrice chapeau r_i s'exprimant comme :

$$r_i = X_i \left[X^T W(s_i) X\right]^{-1} X^T W(s_i)$$

Comme dans le cas des MCO, on peut définir un vecteur des résidus, par différence entre valeur observée et valeur prédite : $\hat{\varepsilon} = Y - \hat{Y} = (I - S_1) Y$. On calcule un estimateur de la variance des résidus, faisant intervenir la somme des carrés des résidus et un paramètre que Brunsdon et alii qualifient de « nombre effectif de degrés de liberté » des résidus.

$$RSS_{GWR} = Y^T (I - S_1)^T (I - S_1) Y$$

On en déduit : $\hat{\sigma}^2 = \dfrac{Y^T (I - S_1)^T (I - S_1) Y}{n - 2\,Tr(S_1) + Tr(S_1^T S_1)}$.

On peut également estimer la variance des paramètres.

5-2-2 La détermination des paramètres optimaux

La valeur des paramètres estimés dépend de la forme fonctionnelle choisie pour le calcul des pondérations. Mais comme pour les estimateurs à noyaux utilisés en estimation nonparamétrique de la densité, c'est la fenêtre utilisée pour l'estimation qui va constituer l'enjeu de la modélisation.

Une des méthodes présentées par Brunsdon est celle de la validation croisée Le critère en est le suivant :

$CV = \sum_{i=1}^{n} \left[y_i - y_{\neq i}(h)\right]^2$ où $y_{\neq i}(h)$ désigne la valeur estimée lorsqu'on a enlevé la i° observation, *h* étant la bande passante.

Le critère le plus souvent préconisé est le critère d'Akaiké, noté AIC, qui s'écrit de la façon suivante : $AIC = 2nLog(\hat{\sigma}) + nLog(2\pi) + n\left\{\dfrac{n + Tr(S_1)}{n - 2 - Tr(S_1)}\right\}$, Tr désignant la trace de la matrice.

Comme les paramètres dépendent du système de poids utilisé, la valeur de l'indicateur d'Akaiké varie en fonction de la distance. L'examen de l'indicateur d'Akaiké permet de déterminer une valeur optimale pour la régression géographique pondérée.

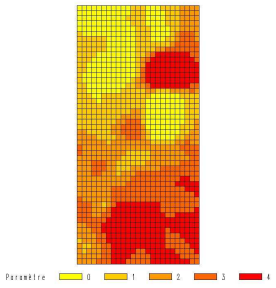
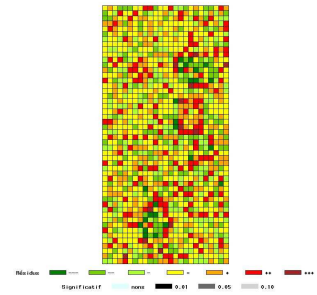
5-2-3 Retour sur l'exemple

L'utilisation de la régression géographique pondérée va permettre de faire apparaître la variabilité spatiale du nombre de personnes par logement. Elle est effectuée à l'aide d'une macro SAS, %rgp.

On commence par déterminer la valeur optimale du paramètre, à l'aide du critère d'Akaiké, ce qui suppose bien sûr de faire tourner plusieurs fois le modèle, en faisant varier nombre de voisins.

Voisins(nombre)	10	20	30	40	50	60	70	80	90
Akaiké	9951	9732	9513	9482	9477	9472	9508	9505	9529

Le calcul effectué en utilisant 60 voisins, correspondant ici au minimum du critère d'Akaiké, permet d'avoir une vision de la non-stationnarité (variation spatiale des valeurs du paramètre de régression) et de voir comment se comportent les résidus.

	
Non stationnarité (variation locale du paramètre)	Résidus

Carte 5.2

La comparaison avec a carte 5.1 permet de voir que la structure spatiale des résidus n'est plus aussi marquée. L' indicateur d'autocorrélation spatiale, le I de Moran a une valeur de 0,07, à comparer avec le 0.52 de la régression par les MCO. Compte tenu du grand nombre d'observations, le test conclut cependant au rejet de H0 (non corrélation des residus).

Dans l'estimation à partir d'un voisinage de 40 carreaux on obtient un I de Moran de 0.01, et on ne rejette pas H0, bien que la valeur de l'indicateur d'Akaiké soit plus élevée.

5-2-4 Test sur la non stationnarité

Dans un article de 1999, Brunsdon propose de tester l'hypothèse MCO versus RGP, en construisant un test de type Fisher. Si on reprend les notations du paragraphe précédent, où S_0 et S_1 désignent respectivement les matrices-chapeau dans le cas des MCO et de la RGP, on définit R_0 et R_1 comme :

$R_0 = (I - S_0)^T (I - S_0)$ (resp S_1 et R_1)

Le test proposé par les auteurs est le suivant :

$$F = \left[\frac{y^t R_0 y - y^t R_1 y}{\upsilon} \right] \left[\frac{y^t R_1 y}{\delta} \right]^{-1}$$

Dans cette expression $\upsilon = Trace(R_0 - R_1) \;\; et \;\; \delta + Trace(R_1)$

Si l'on suit les auteurs, cette statistique suivrait sous de bonnes hypothèses de normalité une distribution de Fischer de paramètres $(\frac{\upsilon^2}{\upsilon'}, \frac{\delta^2}{\delta'})$, avec $\upsilon' = Trace(R_0 - R_1)^2 \; et \; \delta + Trace(R_1)^2$

Si on l'applique à l'exemple précedent, cette statistique prend une valeur de 145 la valeur des paramètres étant de 114 et 1185. La lecture des tables montre que cette valeur est très supérieure à la valeur seuil (12), et on rejette le test d'équivalence entre les MCO et la RGP.

5-3 Une utilisation de type « petits domaines »

Parmi les méthodes relevant des petits domaines, présentées dans Ardilly (2006), se trouve l'estimation par la prédiction. L'utilisation en a été faite pour calculer des estimateurs mixtes, utilisant des données de recensement et des sources administratives. Sur de petits territoires, la variance des estimateurs issus des enquêtes de recensement est très élevée, et l'utilisation de sources administratives comme données auxiliaires peut contribuer à en améliorer la précision.

Ainsi, les sources fiscales permettent de fournir des données proches de l'exhaustivité, entâchées de certains biais (population étudiante, par exemple). Dans le cas du RP, on aura deux types d'adreses :

- les adresses enquêtées pour lsquelles on a l'ensemble de l'information, soit les variables issues du recensement, les variables isues de la source fiscale, et le nombre de logement ;
- les adresses non enquêtées, pour lesquelles on ne disposera que des informations issues de la source fiscale et du nombre de logement.

A partir des adresses pour lesquelles on dispose de l'ensemble de l'ensemble de l'information, on peut construire un modèle expliquant la population du RP par le nombre de logements, et la population isue de la source administrative.

Si a désigne le domaine estimé, et sa l'échantillon sur le domaine, on construit un estimateur qui ne prend pas en compte les poids de sondage, de la forme suivante ;

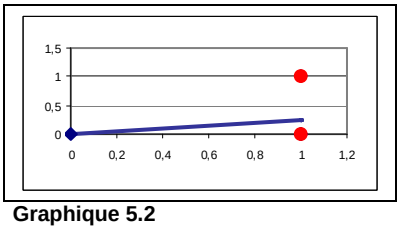
$$\tilde{Y}_a = \sum_{i \in sa} Y_i + \sum_{i \in a, i \notin sa} \hat{Y}_i$$

Dans cette expression, les $\tilde{Y}_i$ sont en général estimés à partir d'un modèle $\hat{Yi} = Xi\hat{\beta}$. C'est là qu'il est intéressant d'utiliser la RGP, pour tenir compte de la non stationnarité des phénomènes et produire des coefficients qui varient spatialement.

5-4 Lien avec l'estimation de densité

Si on applique ces résultats généraux dans un cas très simple, celui d'une régression simple, sans termes constants, sur des couples de variables (x(s),y(s)) qui ne prennent comme valeurs que (0,1) ou (1,1) selon que la personne est employée ou au chômage.

On a donc un modèle particulièrement simple $y = \theta x$, et un nuage de points particulièrement simple :



Graphique 5.2

La droite d'ajustement va passer par un point qui correspondra à la moyenne pondérée locale des x et des y, les pondérations étant obtenues par des estimateurs à noyaux. Le coefficient θ est égal à la moyenne pondérée des y, la moyenne pondérée des x étant dans ce cas égale à 1.

On a $m_Y(s) = \sum_{i=1}^{n} w_i(s) Y_i$, les pondérations wi(s) étant des pondérations locales qui s'expriment de la façon suivante :

$$wi(s) = \frac{K\left(\frac{s - s_i}{h}\right)}{\sum_{k=1}^{n} K\left(\frac{s - s_k}{h}\right)} = \frac{\frac{1}{nh} K\left(\frac{s - s_i}{h}\right)}{\frac{1}{nh} \sum_{k=1}^{n} K\left(\frac{s - s_k}{h}\right)} = \frac{\frac{1}{nh} K\left(\frac{s - s_i}{h}\right)}{\hat{f}_A(s)}$$

On retrouve en effet au dénominateur l'estimation de la densité en s à partir de l'ensemble des points d'observation $\frac{1}{nh} \sum_{k=1}^{n} K\left(\frac{s - s_k}{h}\right)$, h désignait la fenêtre et n le nombre des points d'observation.

La nature particulière des valeurs de Y, 0 ou 1 selon que la personne est ou non au chômage, permet de reformuler $m_{Y(s)}$ de la façon suivante :

$$m_Y(s) = \sum_{i=1}^{n} w_i(s) Yi = \frac{1}{\hat{f}_A(s)} * \frac{1}{nh} \sum_{k=1}^{n} K\left(\frac{s - s_k}{h}\right) * yi = \frac{n_c}{n} * \frac{1}{\hat{f}_A(s)} * \frac{1}{n_c h} \sum_{k=1}^{n} K\left(\frac{s - s_k}{h}\right)$$

où n_c désigne le nombre de chômeurs.

On retrouve alors au dénominateur l'estimation de la densité des chômeurs, et on aboutit au résultat suivant $m_y(s) = \frac{n_C}{n} * \frac{\hat{f}_C(s)}{\hat{f}_A(s)}$.

A un coefficient près, qui correspond à la part des chômeurs sur le périmètre d'estimation, on retrouve le ratio des densités estimées. On retrouve donc le résultat présenté en 3 par des méthodes de régression non paramétrique.

6 - Détection de clusters d'activité ou d'équipements

Les fonctions de Ripley, présentées rapidement en 3-3, ont fait l'objet, depuis leur introduction, de nombreux développements. La littérature sur la concentration industrielle et les clusters est abondante, et elle a été profondément renouvelée par l'introduction des méthodes spatiales.

6-1 Indicateurs de Ripley et prolongements

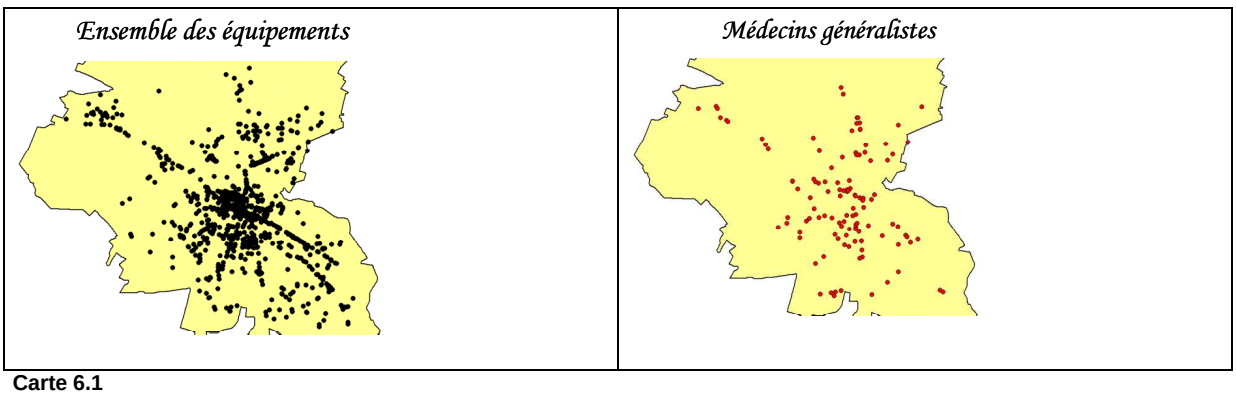
Les indicateurs de concentration industrielle comme ceux de Gini ou d'Herfindahl-Hirschman sont fondamentalement aspatiaux. Les travaux économiques récents ont utilisé les résultats proposés par Ellison et Glaeser, mais ces indices ne prennent pas en compte les distances, ni les localisations des zones les unes par rapport aux autres. Duranton et Overman (2005) ont profondément modifié l'approche en utilisant des estimateurs à noyaux. Dans le même temps, en venant d'autres disciplines, Marcon et Puech ont proposé des extensions des fonctions K de Ripley.

On trouvera dans un article récent de Marcon et Puech (2010) une typologie des indicateurs de concentration spatiale basé sur les distances. Les travaux partent en général des indicateurs de Ripley et Besag, qui permettent de mettre en évidence des grappes d'équipements. Mais ces indicateurs ne permettent pas de savoir, ou pas directement, si cette concentration n'est pas due à un autre processus sous-jacent.

6-2 Le M de Marcon et Puech

Les indicateurs proposés par Marcon et Puech (2004) permettent d'aller plus loin en ce sens et de mesurer cette concentration par rapport à un ensemble d'équipements, ou par rapport à la population. Un équipement directement lié à la population peut sembler concentré spatialement parce que la population qu'il dessert est elle-même concentrée. Une sur-représentation, ou sur-concentration a lieu lorsque les équipements seront relativement plus nombreux que à la population de référence.

Illustrons ceci par la répartition des médecins et l'ensemble des équipements (Amiens):



Carte 6.1

L'idée de Marcon et Puech (2005), avec leur indicateur M, est de comparer la tendance de la sous-population à se concentrer à celle de la population d'ensemble, que l'on peut schématiser de la façon suivante :

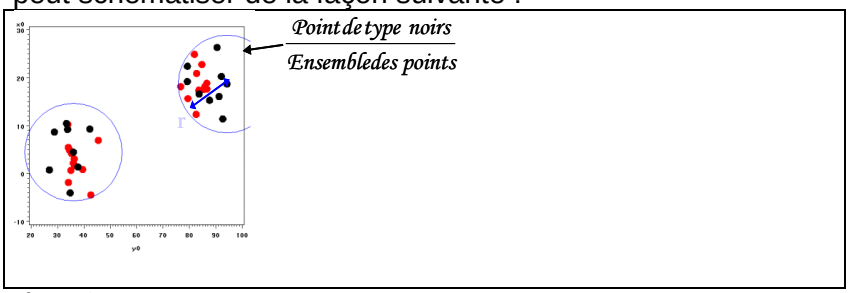


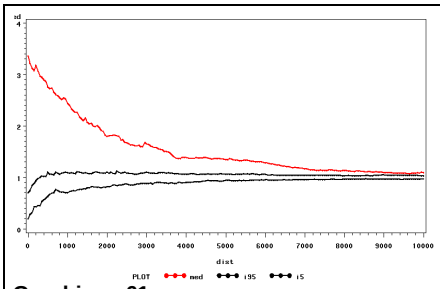
Figure 6.1

La formule de l'indicateur M est la suivante $M_{Sk}(r) = \dfrac{\dfrac{1}{N_{Sk}} \sum_{i=1}^{N_{Sk}} \dfrac{\sum_{j=1,\, j\neq i}^{N_{Sk}} c_{Sk}(i,j,r)}{\sum_{j=1,\, j\neq i}^{N} c(i,j,r)}}{\dfrac{N_{Sk}-1}{N-1}}$

Dans cette formule, r désigne le rayon autour duquel on effectue les observation, N le nombre total d'équipements, N_{Sk} le nombre d'équipements de type k $c_{Sk}(i,j,r)$ est une indicatrice qui vaut 1 lorsqu'un équipement de type k se trouve à une distance inférieure à r de l'équipement de type k à partir duquel est effectué le comptage, tandis que $c(i,j,r)$ vaut 1 si c'est un équipement quelconque qui est situé dans le cercle de rayon r.

Par construction, ce rapport est inférieur ou égal à 1 Lorsqu'on le norme par l'expression $\dfrac{N_{Sk}-1}{N-1}$, on a une idée de la concentration locale de l'équipement Lorsque $\dfrac{\sum_{j=1,\, j\neq i}^{N_{Sk}} c_{Sk}(i,j,r)}{\sum_{j=1,\, j\neq i}^{N} c(i,j,r)} >= \dfrac{N_{Sk}-1}{N-1}$, l'équipement étudié sera localement plus représenté que l'ensemble, le rapport $\dfrac{\dfrac{\sum_{j=1,\, j\neq i}^{N_{Sk}} c_{Sk}(i,j,r)}{\sum_{j=1,\, j\neq i}^{N} c(i,j,r)}}{\dfrac{N_{Sk}-1}{N-1}}$ étant alors supérieur à 1. On a une expression qui n'est pas sans rapport avec les « Lisa ». L'indicateur local indique un point chaud dans la répartition spatiale des phénomènes, tandis que l'indicateur global, qui est une agrégation des indicateurs locaux, mesure la tendance globale à la concentration.

Dans le cas de la répartition des médecins, on obtient les résultats suivants :

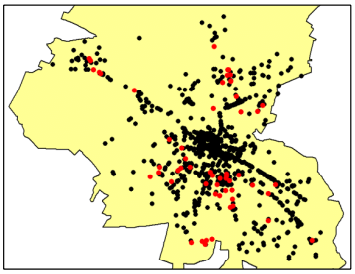


Graphique 61

Par des méthodes de type Monte-Carlo, on peut calculer des intervalles de confiance permettant de dire, si, et jusquà quelle distance, il y a concentration spatiale de l'équipement étudié, relativement à l'ensemble des équipements

Dans le cas des médecins, la localisation des médecins diffère sensiblement de celle des autres équipements

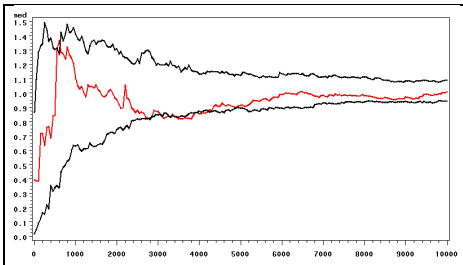
Médecins



Carte 6.2

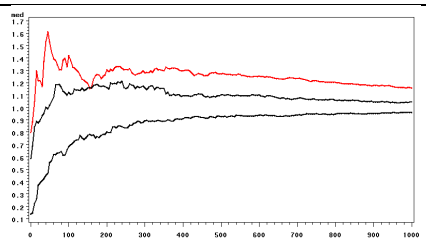
Les points rouges représentent les équipements pour lesquels le rapport local est supérieur à 1

Les indicateurs de Marcon et Puech, comme les indicateurs de Ripley permettent des comparaisons On a chois de présenter les résultats pour trois types d'équipement : médecins (Graphique 6.1), boulangeries (Graphique 6.2) et coiffeurs (Graphique 6.3)



Boulangeries

Graphique 6.2



Salons de coiffure

Graphique 6.3

Dans le cas des boulangeries, la courbe des situe à l'intérieur de l'intervalle de confiance.Lla localisation ne se distingue donc pas de celle de l'ensemble des équipements. Dans le cas des salons de coiffures, il y a une tendance à la surreprésentation locale, mais celle ci n'est pas très marquée.

Bibliographie

Anselin,L(1995) *Local indicators of spatial association:LISA,* Geographical analysis 27(2),93-116

Ardilly,P.(2006) *Panorama des méthodes d'estimation sur petits domaines*, Insee, document de travail de l'UMS, N°M0602

Bailey,T.C, Gattrell,AC, (1995)- *Interactive spatial data analysis* - Prentice & Hall

Brunsdon,C., Fotheringham,A.S., Charlton,M., *Geographically weighted regression* -The statistician(1998),47,Part 3, 431-443

Cliff, A.D, Ord,J.K, (1973) *Spatial autocorrelation* ,Pion

Cressie, N,(1993) *Statistics for spatial data*, Wileyl

Delecroix,M. - *Estimation non paramétrique des densités et des régressions* (1997) dans INSEE-Méthodes 59-60-61

Diggle,P,J (2003)- *Statistics analysis of spatial point patterns* (2003)- HArnold

Diggle,P.J., Zheng,P. , Durr,P.(2005)- *Nonparametric estimation of spatial segregation in a multivariate point process* - Applied statistics,54,part3, pp645-658

Diggle,P,J, et Marron,J,S - Equivalence of smoothing parameter selectors in density and intensity estimation - JASA-Sept 1988- Vol83 N°403

Droesbeke,J .J, Lejeune,M.,Saporta,G. *Analyse statistique des données spatiales*(2005), Technip

Duranton & Overman (2005), *Testing for localisation using microgeographis data*,Review of economic studies,72(4),1077-1106

Escofier,B., Benali,H.,Bachar,K,(1990), *Comment introduire de la contiguïté en analyse des données*, Statistique et analyse des données, Vol15,N°3

Feitosa, F.F., G. Camara, A.M.V. Monteiro, T. Koschitzki,and M.P.S. Silva. (2004). *Spatial measurement of Residential Segregation*, in Proceedings of GeoInfo 2004, pp. 59–73Geneve: IFIP

Fotheringham,A.S., Brunsdon,C., Charlton,M., (2002) *Gegraphically weighted regression*,J.Wiley

Getis,A, Ord,JK,(1996), The analysis of spatial association by distance statistics, Geographical analysis, 24(3) 189-207

Goreaud,F (1998)*"Apports de l'analyse de la structure spatiale en forêt tempérée à l'étude et la modélisation des peuplements complexes"* Cemagref (Thèse de doctorat)

Gotway,CA - Schabenberger,O.(2005) - *Statistical methods for spatial data analysis* - Chapman & Hall

Gotway,C,A ,Waller,LA,(2004) *Applied statistics for public health data*;Wiley

Hardle,W,(1990) - *Applied nonparametric regression*, Cambridge university press

Hardle,W(1998) - *A course on non and semi parametric modelling* - Polycopié ENSAE

Kelsall,J, Diggle,PJ,(1995),*Kernel estimation of relative risk*,Bernoulli,Vol.1,N°1-2, p3-16

Le Toqueux(2004) - *Les indices de ségrégation spatiale de Reardon et O'Sullivan* - Note méthodologique Insee

Lipatz,J.L. (2009),*Mode d'emploi de la boite à outils d'analyse spatiale*, ftp-consultation.intranet.insee.fr\infracommunal\kde

Lloyd,C(2006), *Local model for spatial analysis*, CRC Press

Marcon,E, Puech,F,(2004) *Generalizing Ripley'sK function to inhomogeneus populations*, http://halshsarchives-ouvertesfr/halshs-00372631

Marcon,E, Puech,F,A (2010) *A Typology of Distance-Based Measures of Spatial concentration*, http://halshsarchives-ouvertesfr/docs/00/67/99/93/PDF/HAL-Marcon_Puech-A_Typology_of_Distance-Based_Measures_of_Spatpdf

Massey, DS,Denton,N,A,(1988) *The dimension of residential segregation*, Social forces,Vol67,N°2

Rathelot,R.,Sillard,P., (2010), *L'apport des méthodes à noyaux pour mesurer la concentration*
Reardon,S.F., O'Sullivan,D(2004), *Measures of Spatial Segregation* - Working paper, Pennsylvania state university
Ripley,B.,D.(1977), *Modelling spatial patterns*,Journal of the Royal Statistical Society, Series B, 1977, p. 172-212
Silverman,BW(1986) - *Density estimation for statistics and data analysis* - Chapman & Hall
Tannier,C.,Pumain,D.(2005) *Fractals in urban geography : a theoretical outline and an empirical example*, Cybergeo, revue electronique de géographie,http://cybergeo.revues.org/3275