Repérage de zones défavorisées
Comparaison de méthodes

Jean-Michel Floch

Insee-DG

Département de l'action régionale

jean-michel.floch@insee.fr

Les données spatialisées

- Depuis le travail de synthèse de Cressie (1993), on distingue trois types de données spatialisées:
  - Géostatistiques
  - Surfaciques
  - Ponctuelles
- Des méthodes spécifiques ont été développées dans chaque domaine
  - ex :Variographie, krigeage dans le cas de la géostatistique

Pour un même problème pratique…

- Plusieurs méthodes peuvent se retrouver en concurrence
- Laissons de côté les cas où l'on fait rentrer de force les données dans une modélisation non adaptée (quelques utilisations limite du krigeage en sciences sociales..)
- Cela peut provenir des données (valeurs brutes/taux..) ,du maillage territorial

Quelques questions

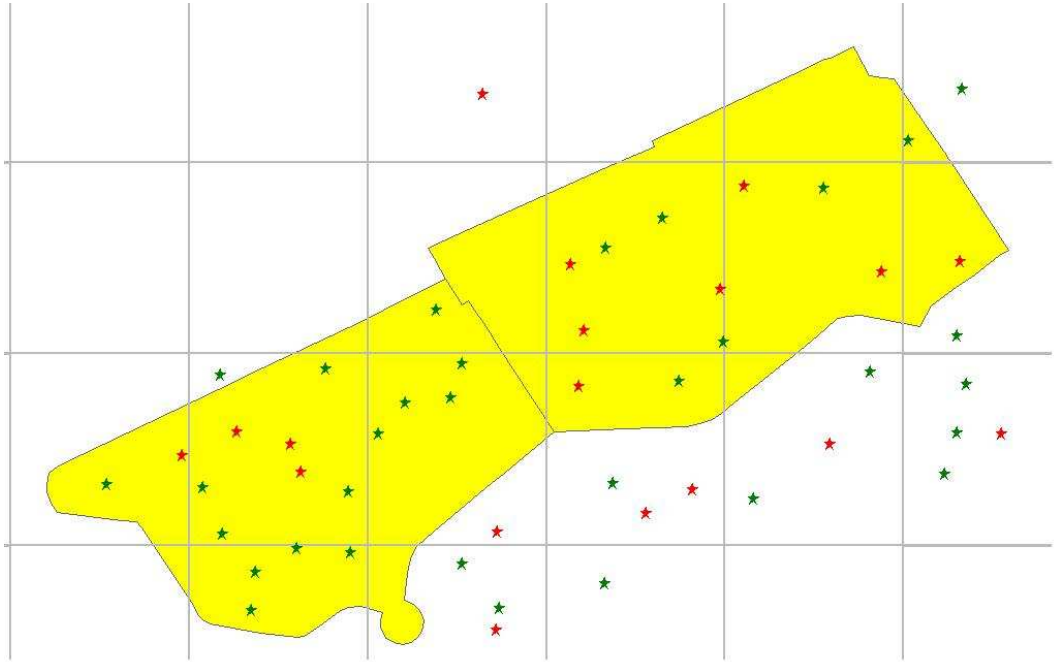
- A information initiale identique, comment départager les méthodes

- Lorsque l'on fait « avec ce que l'on a » (comme données, les données pouvant conditionner l'éventail des méthodes possibles), est-ce qu'on obtient des résultats acceptables?

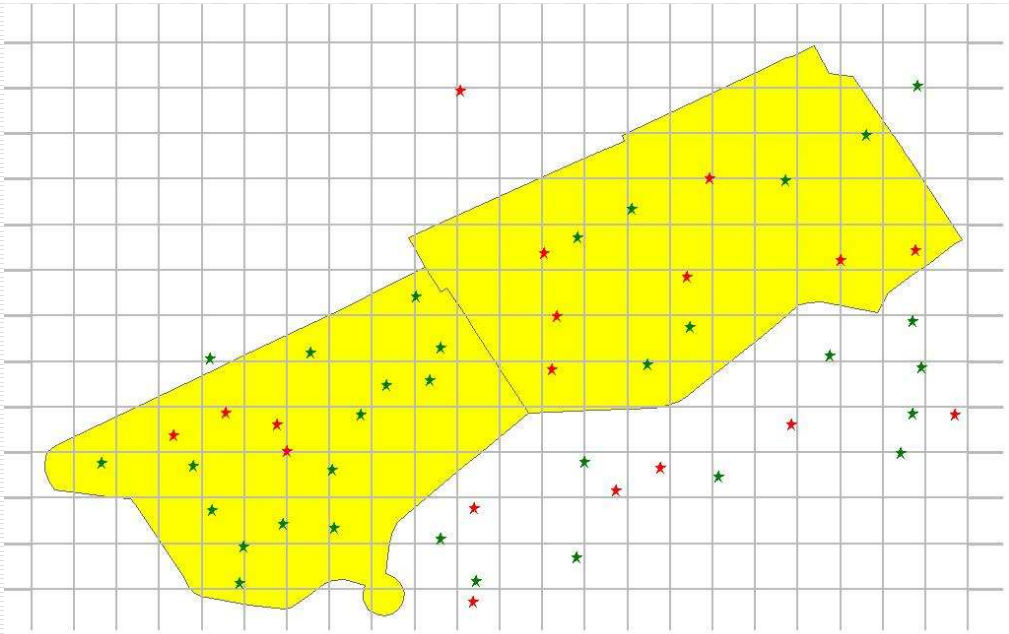
Un exemple : l'étude de la localisation des bas (ou hauts) revenus

Info.

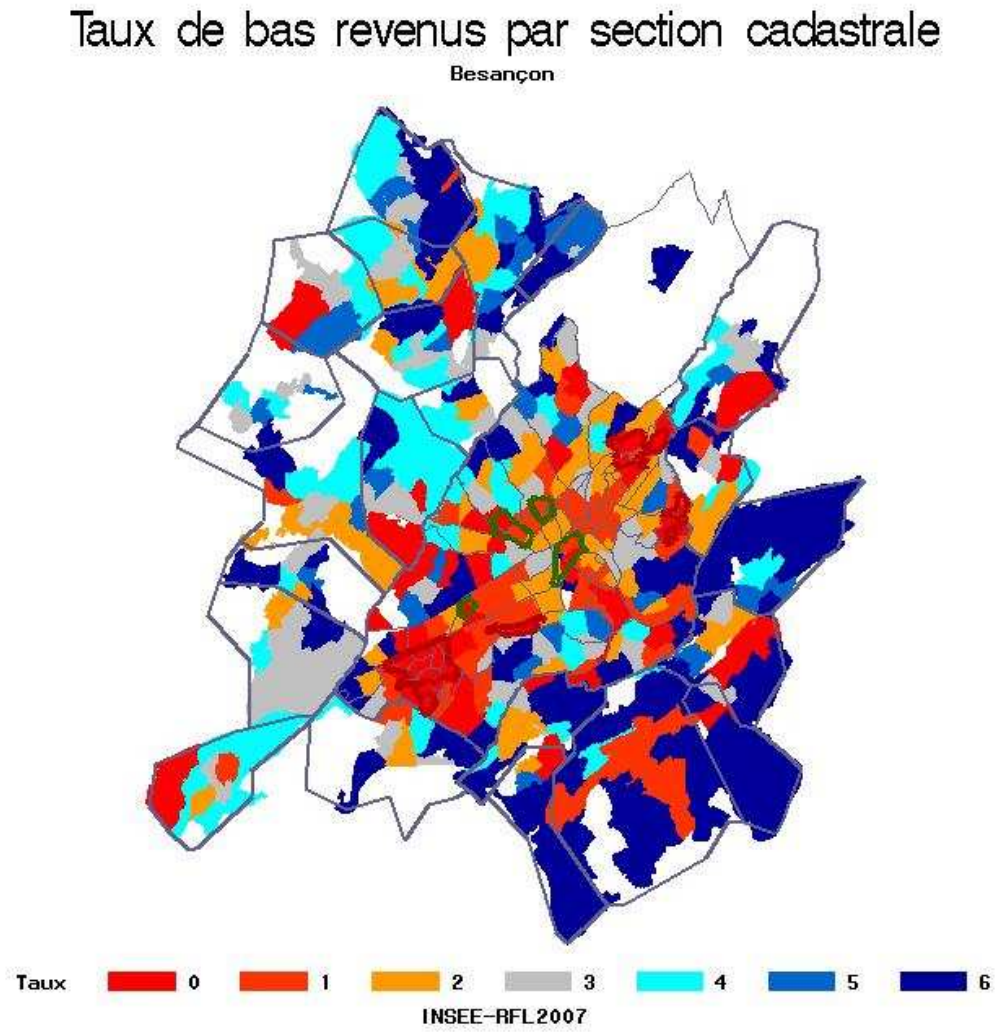
- On dispose pour chaque ménage de variables indicatrices indiquant que le revenu/UC est inférieur ou supérieur à un certain seuil, ainsi que de la localisation géographique (x,y). C'est l'information la plus complète que l'on puisse avoir
- On peut avoir des comptages sur un carroyage
- On peut enfin avoir des taux sur des découpages administratifs



La réduction de la maille fait qu'on se « rapproche » des données individuelles



La situation étudiée



Les méthodes étudiées pour résoudre ce problème

- Les LISA (Surfacique)
- Ratios de densité estimée (Ponctuel)
- Indicateurs « M » de Marcon et Puech (Ponctuel)

Les LISA

(Local indicators of spatial association
L.Anselin Geographical analysis 1995)

Le I de Moran : mesure globale de l'autocorrélation

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

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

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

$$E(I) = \frac{-1}{N-1}$$

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

$$\mathrm{V(I)} = \frac{\mathrm{N^2 S1 - N S_2 + 3 S_0^{\,2}}}{\mathrm{(N-1)(N+1) S_0^{\,2}}} - \left(\frac{1}{\mathrm{N-1}}\right)^2$$

$$\frac{I - E(I)}{\sqrt{V(I)}} \approx N(0,1)$$

Global et local

- S'il n'y avait pas d'autocorrélation, on n'étudierait pas le territoire

- D'où l'intérêt de mesures qui montrent d'où vient cette autocorrélation :développement des LISA et mise en évidence de points chauds (hot spots)

La version « Lisa » du I de Moran

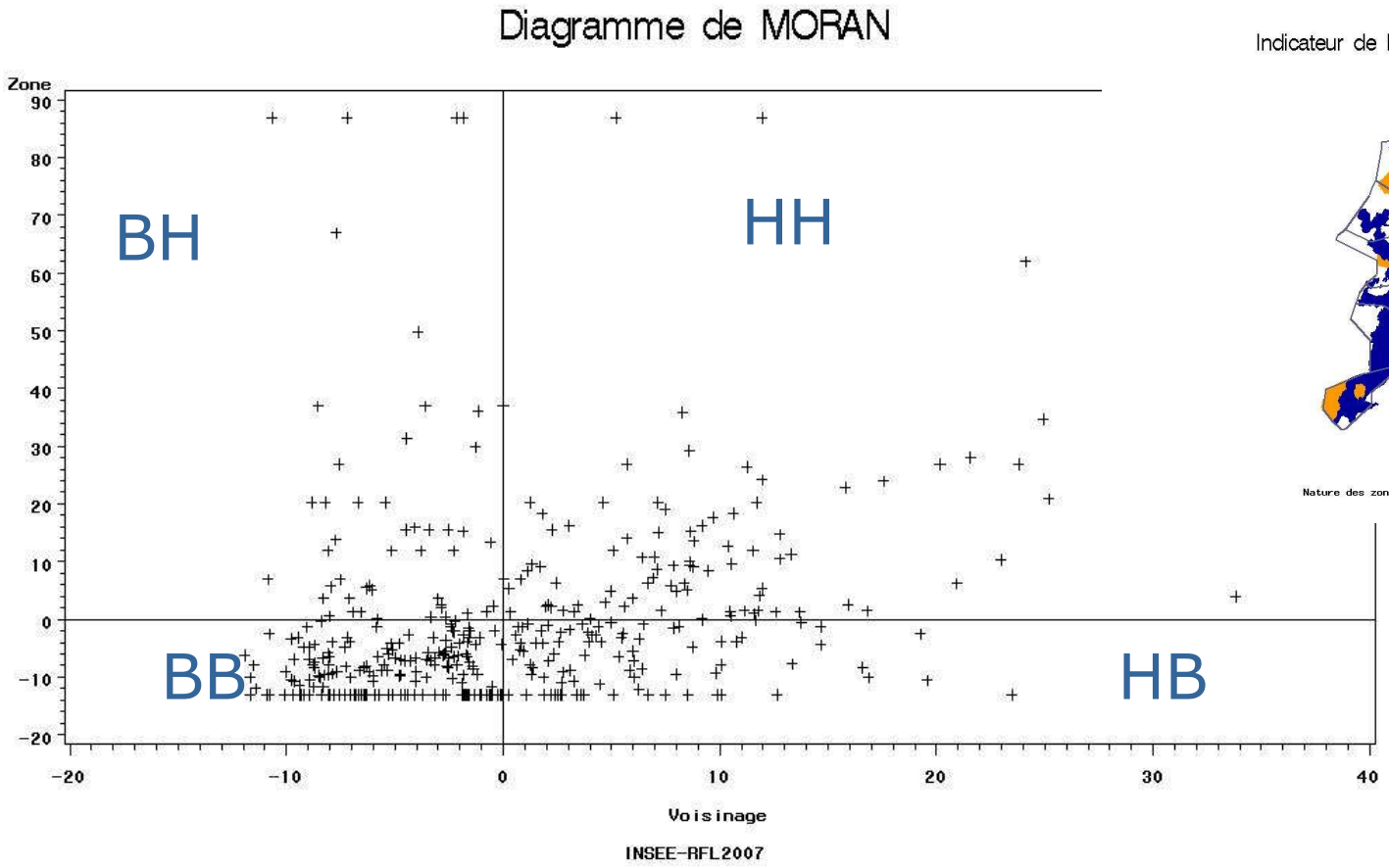
- Les Lisa ont deux propriétés.
- -ils traduisent l'importance de la propension à former des grappes autour d'une zone considérée
- -leur somme est proportionnelle à un indicateur global d'association spatiale.

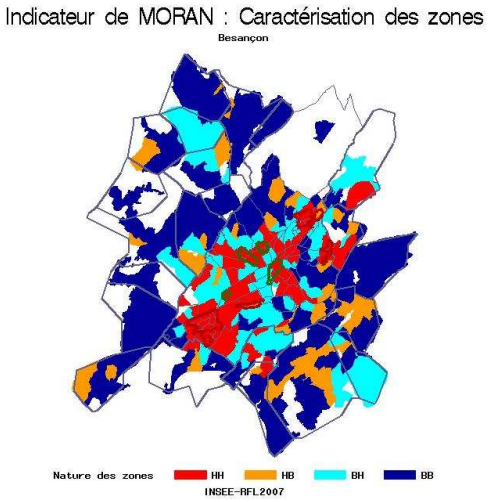
$$I_i = \frac{z_i}{m_2} \sum_{j=1}^{n} w_{ij} z_j$$

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

Valeur observée dans le voisinage

Graphique de Moran et partition de la population (HH,BB,HB,BH)





Des tests

- Mise en évidence des zones qui se détachent significativement des autres

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

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

$$Z(I_i) = \frac{I_i - E(I_i)}{\sqrt{Var(I_i)}}$$

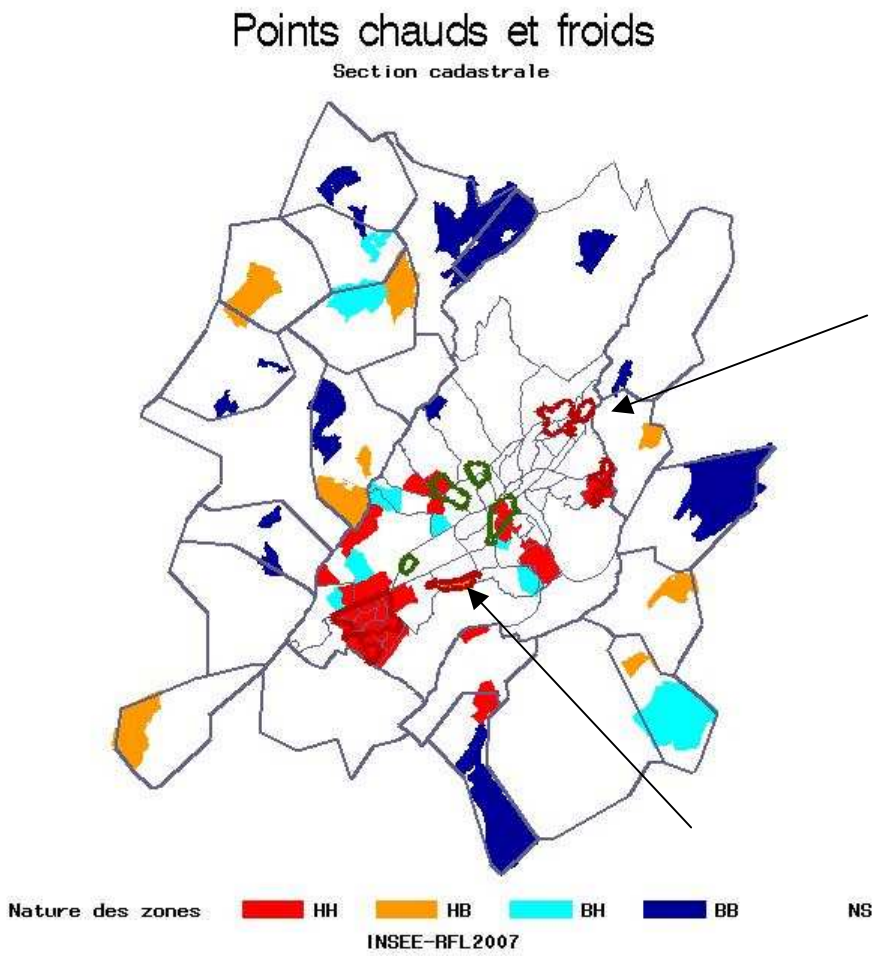
$$b_2 = \frac{m_4}{m_2^2}$$

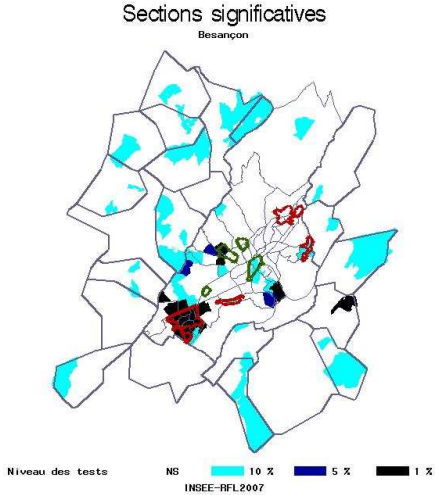
$$m_r = \sum_{i=1}^{n} \frac{z_i^r}{n}$$

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

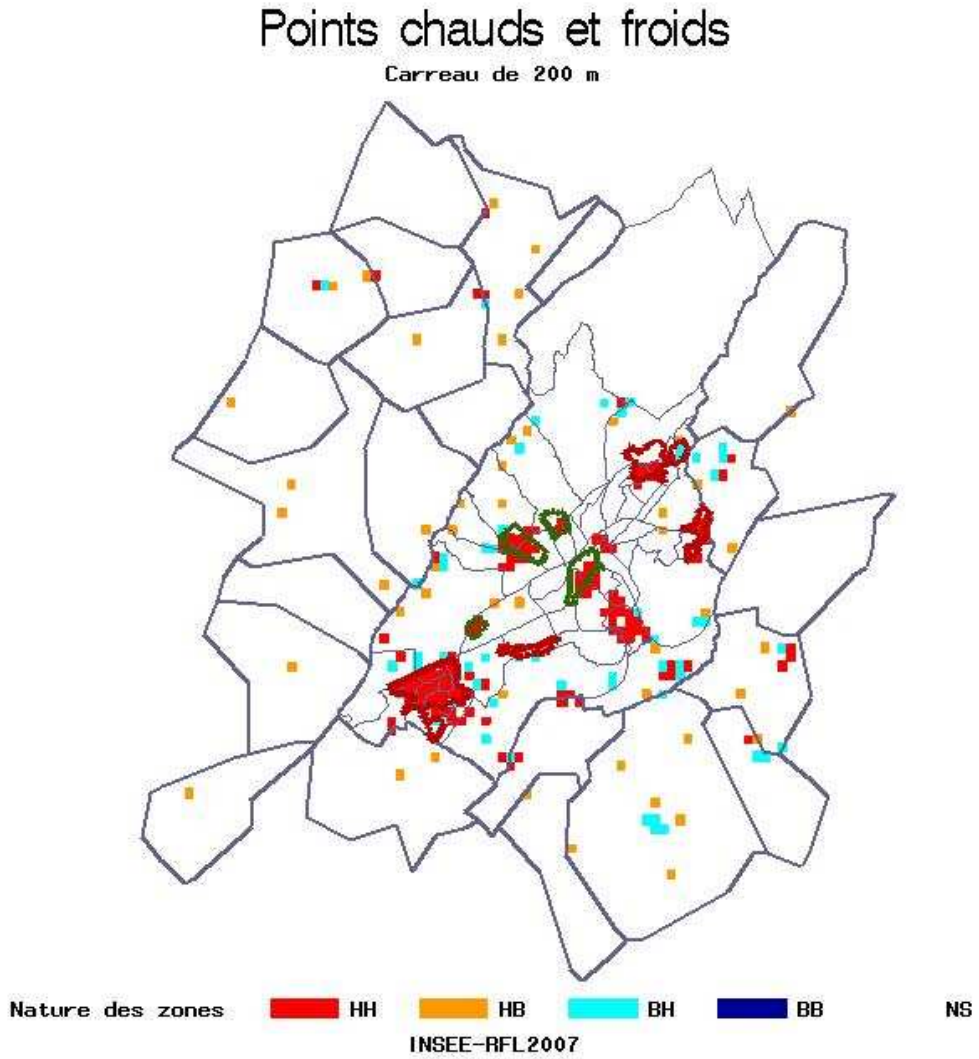
$$w_{i(kh)} = \sum_{k=}^{n} \sum_{h=1}^{n} w_{ik}\, w_{ih}$$

Résultats sur les données à la section cadastrale





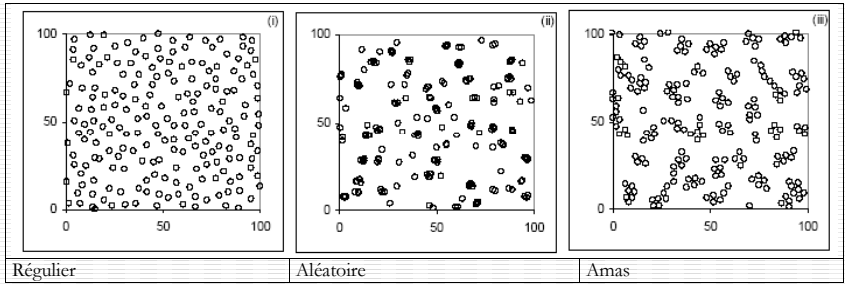
Résultats sur un carroyage



Les ratios de densité estimée

Le cadre théorique

- Les données : une série de localisation, chaque point pouvant être doté d'un certain nombre d'attributs



- Une caractéristique du premier ordre, l'intensité

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

N(S) est le nombre de points dans l'Aire S

Processus homogène

- Si λ=cte, on peut définir un processus homogène de Poisson (l'aléatoire dans le plan)

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

- Dans le cas contraire (la règle pour nous) , on a un processus inhomogène (cas d'un processus inhomogène de Poisson)

$$P(N(S)=n)=e^{-\nu(S)}\frac{(\nu(S))^n}{n!} \qquad \text{avec} \qquad \nu(S)=\int_S \lambda(s)ds$$

Intensité du processus et densité de probabilité

- On va utiliser des techniques non paramétriques d'estimation de la densité
- Lien entre λ(s) et f(s)(densité de probabilité de la répartition des points)
  - $\int \lambda(s)ds = n$
  - $\int f(s)ds = 1$
- Article de Diggle et Marron (JASA 1988) sur l'équivalence de l'estimation de l'intensité et de la densité

Ratio de densité, risque relatif

- Les épidémiologistes définissent un risque relatif

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

Cas/contrôle (0)

- Ce qui fait apparaître le rapport des intensités. On va ici introduire le rapport des densités, qui va traduire la sur(sous)représentation de la population d'intérêt par rapport à la population de référence

Propriété du rapport de densité

- Dans un cadre bien délimité (ici population d'intérêt sous-ensemble de la population de référence), le ratio des densités de probabilité f_B et f_P (f_b et f_p étant définis par rapport à la mesure de Lebesgue) est lui même la densité de probablité de la distribution de B par rapport à la mesure $f_P.\mu$ (théorème de Radon-Nikodym)
- Interprétation possible également en terme d'espérance conditionnelle

L'estimation du ratio

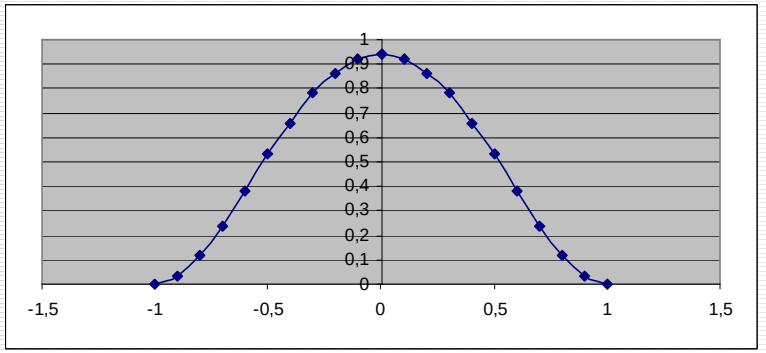
- On va quitter le monde merveilleux des probabiltés et estimer $r = f_B / f_P$ par $\hat{r} = \hat{f}_B / \hat{f}_P$

- On va utiliser pour cela les estimateurs non-paramétriques de la densité (les exemples seront présentés sur la droite et non dans le plan, mais ils peuvent se généraliser (au prix quand même de quelques difficultés)

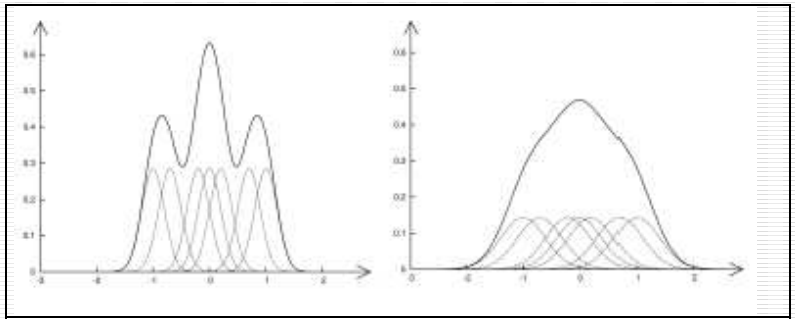
L'estimation de la densité

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

Exemple de noyau
Biweight de Tukey



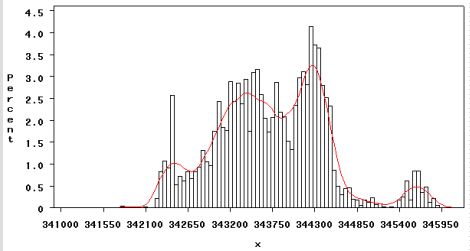
- h Fenêtre
- K fonction Noyau (kernel)
- x point d'estimation
- X_i point d'observation



Représentation comme une somme de bosses

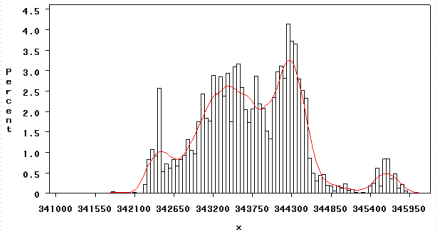
L'épineux problème de la fenêtre

- Un consensus sur le fait que le noyau importe peu (à condition qu'il soit décroissant et symétrique)

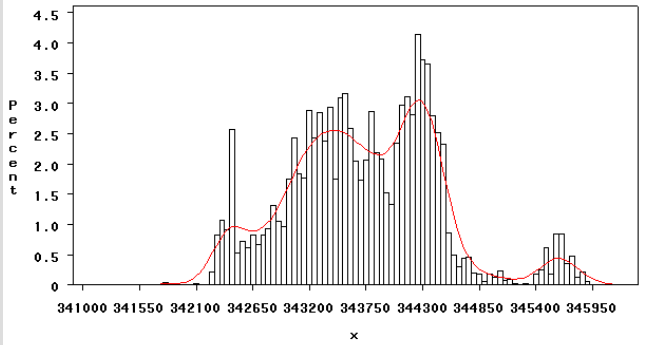


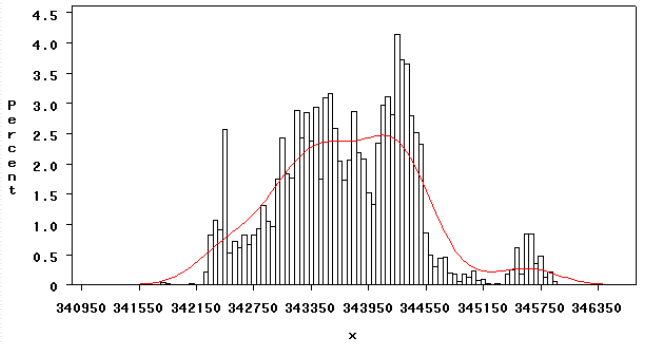
T

Bw



- La fenêtre est plus difficile à déterminer





Quelques difficultés

- Biais des estimateurs

- Fenêtre optimale pour estimer à la fois la densité « numérateur » et la densité « dénominateur »

Des résultats d'optimalité problématiques

- Dans SAS, la procédure KDE, qui utilise le noyau gaussien, qui permet des résultats ultra-rapides (FFT), le noyau optimal est

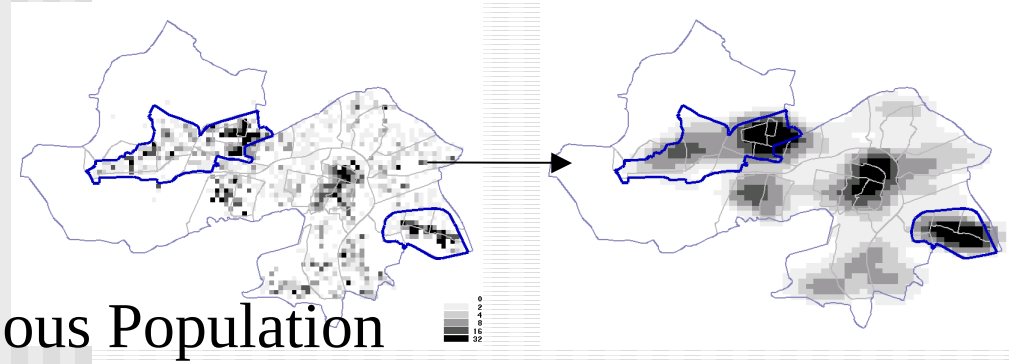
$$\frac{1}{2\pi h_x h_y}\exp\left(-\frac{\left(\frac{x}{h_x}\right)^2+\left(\frac{y}{h_y}\right)^2}{2}\right)$$

$$h_x = \hat{\sigma}_x \eta^{-\frac{1}{6}} \text{ (resp. y)}$$

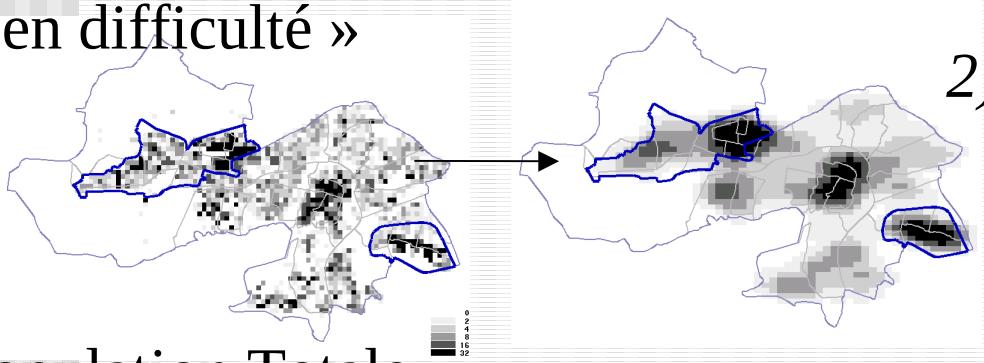
- L'expérience et la confrontation avec la connaissance du terrain nous a conduit à diminuer par deux la bande passante

Un outil standard

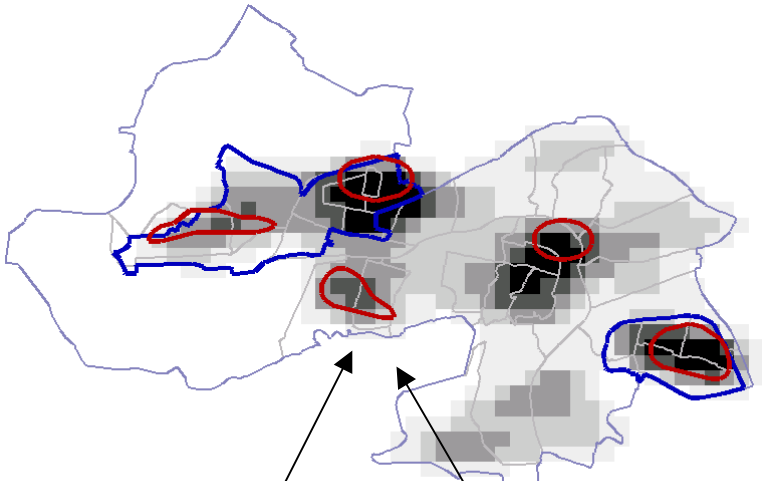
1) Simplification



Sous Population « en difficulté »



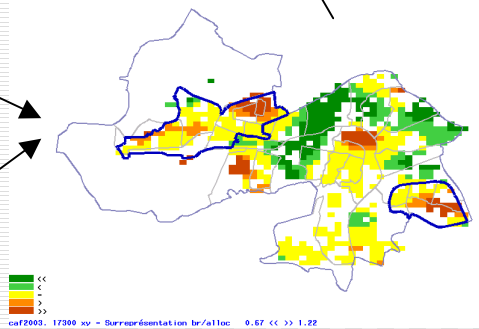
Population Totale



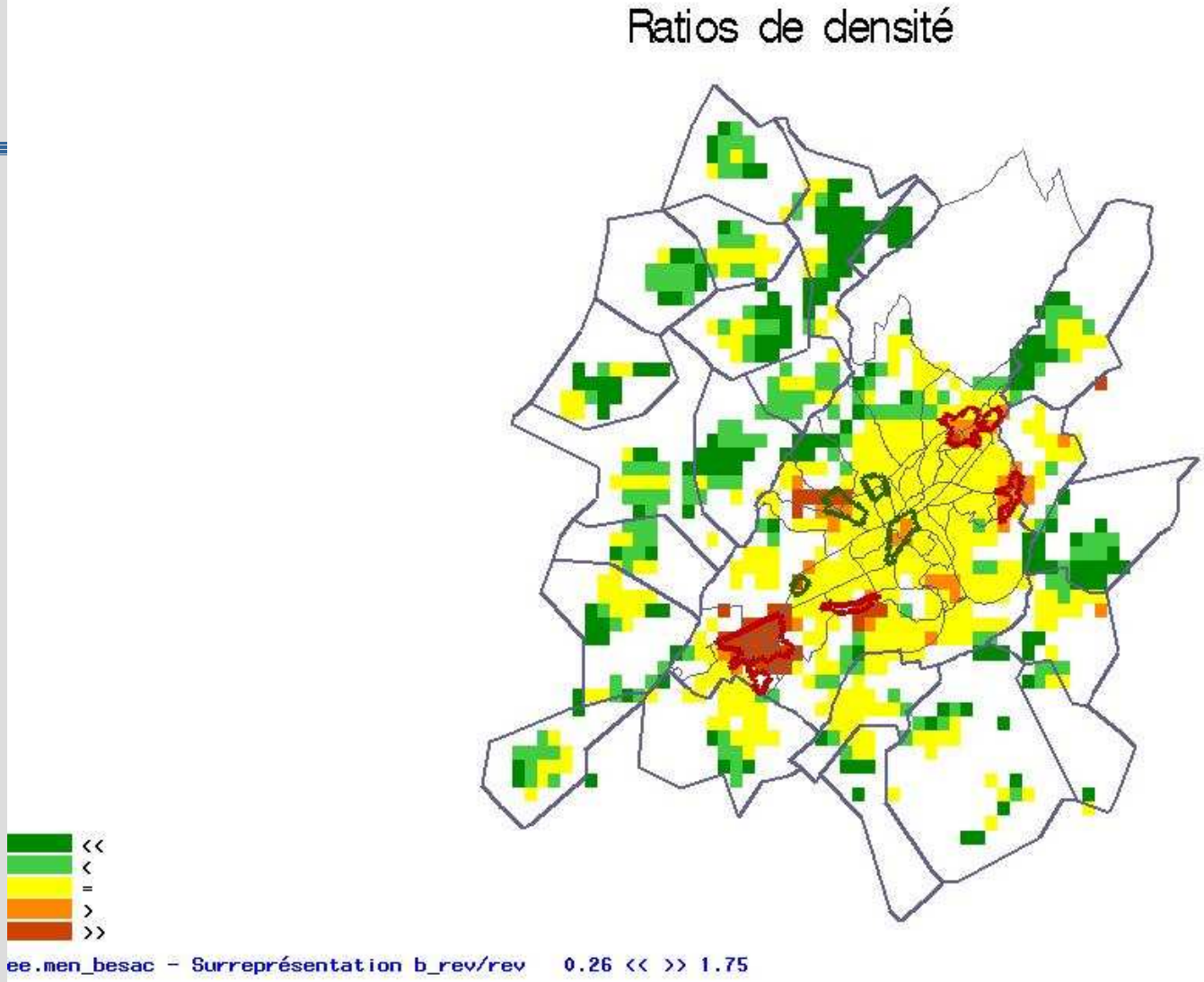
3) Extraction des contours

Superposition

2) SP/PT



Le résultat


Ratios de densité
<<
<
=
>
>>
ee.men_besac - Surreprésentation b_rev/rev 0.26 << >> 1.75

Indicateurs de Marcon et Puech

Les propriétés du second ordre

- Intensité d'un processus ponctuel

$$\lambda(s) = \lim_{ds \to 0} \left\{ \frac{E(N(ds)}{ds} \right\}$$ ← 1° ordre

- Propriétés du second ordre

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

  - Inutilisable en tant que telle
  - Ce sont les fonctions K de Ripley (et leurs variantes) qui servent à aborder ces questions

La fonction K de Ripley

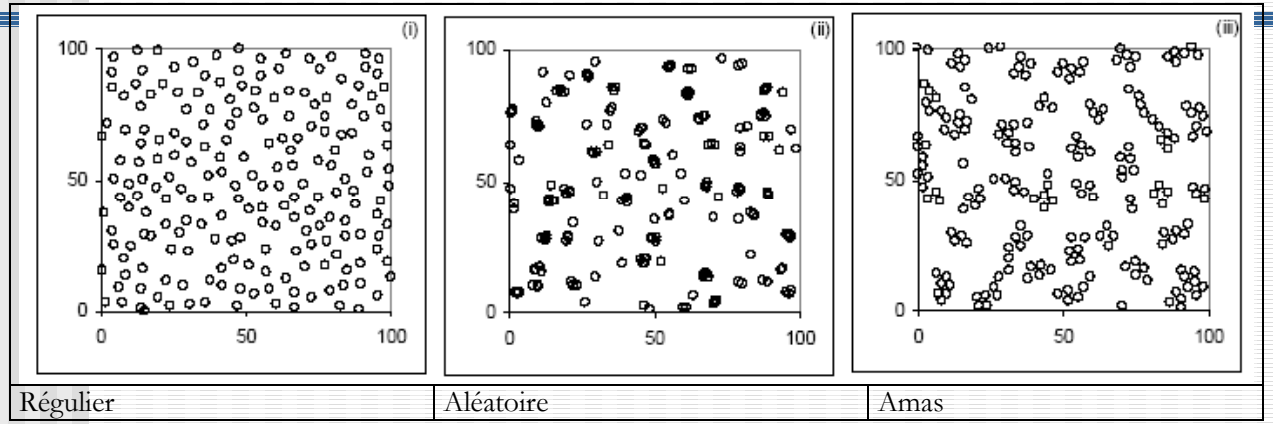
$$\frac{K(r)}{\pi r^2} = \frac{\frac{1}{N}\sum_{i=1}^{N}\sum_{j\neq i} d_{ij}(r)}{\frac{N-1}{D}}$$

N : Nbre de points du semis

$d_{ij}(r)$: indicatrice valant 1 si j est à une distance inférieure à r de i

D : Aire du domaine

Trois configurations classiques, leur K (Ripley), leur L(Besag)

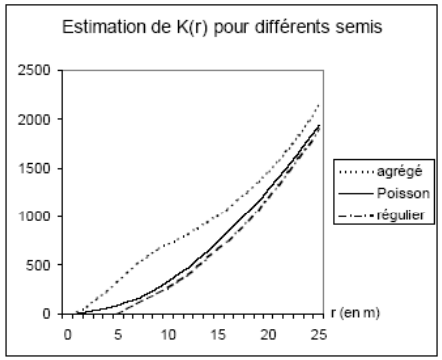


Régulier	Aléatoire	Amas

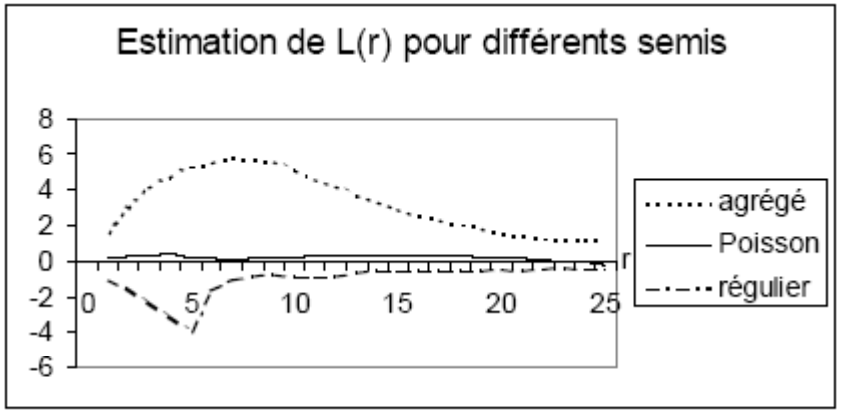
(Source : Thèse de François GOREAUD)

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

K



Estimation de la fonction K(r) pour un semis agrégé, un semis de Poisson, et un semis régulier.

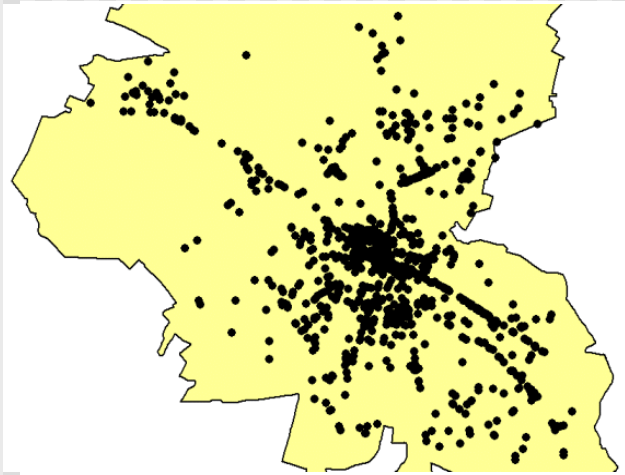


L

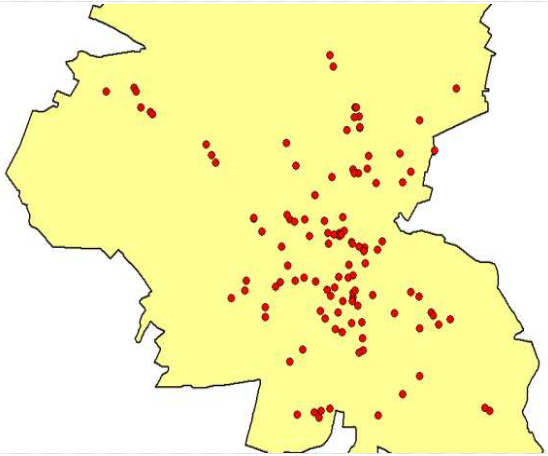
De K et L à M….

- Les fonctions K et L permettent de savoir si le processus est ou n'est pas homogène, mais comme toujours, par rapport à quoi (espace « amorphe », R^2 ou espace structuré)

Ensemble des équipements



Médecins généralistes



Les travaux de E.Marcon et F.Puech

- Ils visent :
  - À prendre en compte la structure sous-jacente (dans notre cas, la population des bas revenus relativement à un espace peuplé et non à R^2)
  - A permettre (comme dans le cas des LISA de faire apparaître les zones de non-stationnarité

La formulation de M

Proportion moyenne de points de type Sk au voisinage de chaque entité de Sk

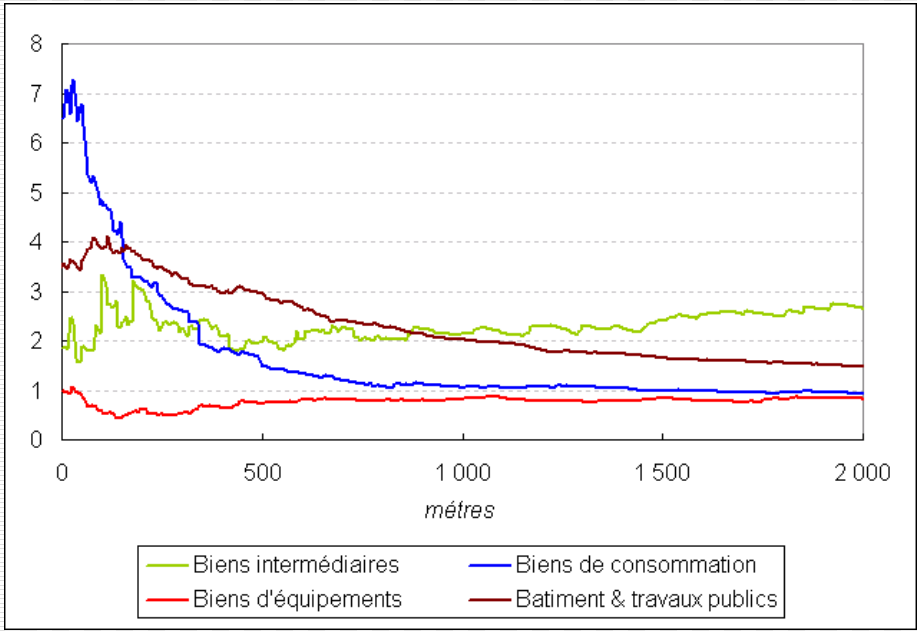
Valeurs individuelles de M

$$M_{Sk}(r) = \frac{1}{N_{Sk}} \sum_{i=1}^{N_{Sk}} \frac{\sum_{j=1, j\neq i}^{N_{Sk}} c_{Sk}(i,j,r)}{\sum_{j=1, j\neq i}^{N} c(i,j,r)} \cdot \frac{N_{Sk}-1}{N-1}$$

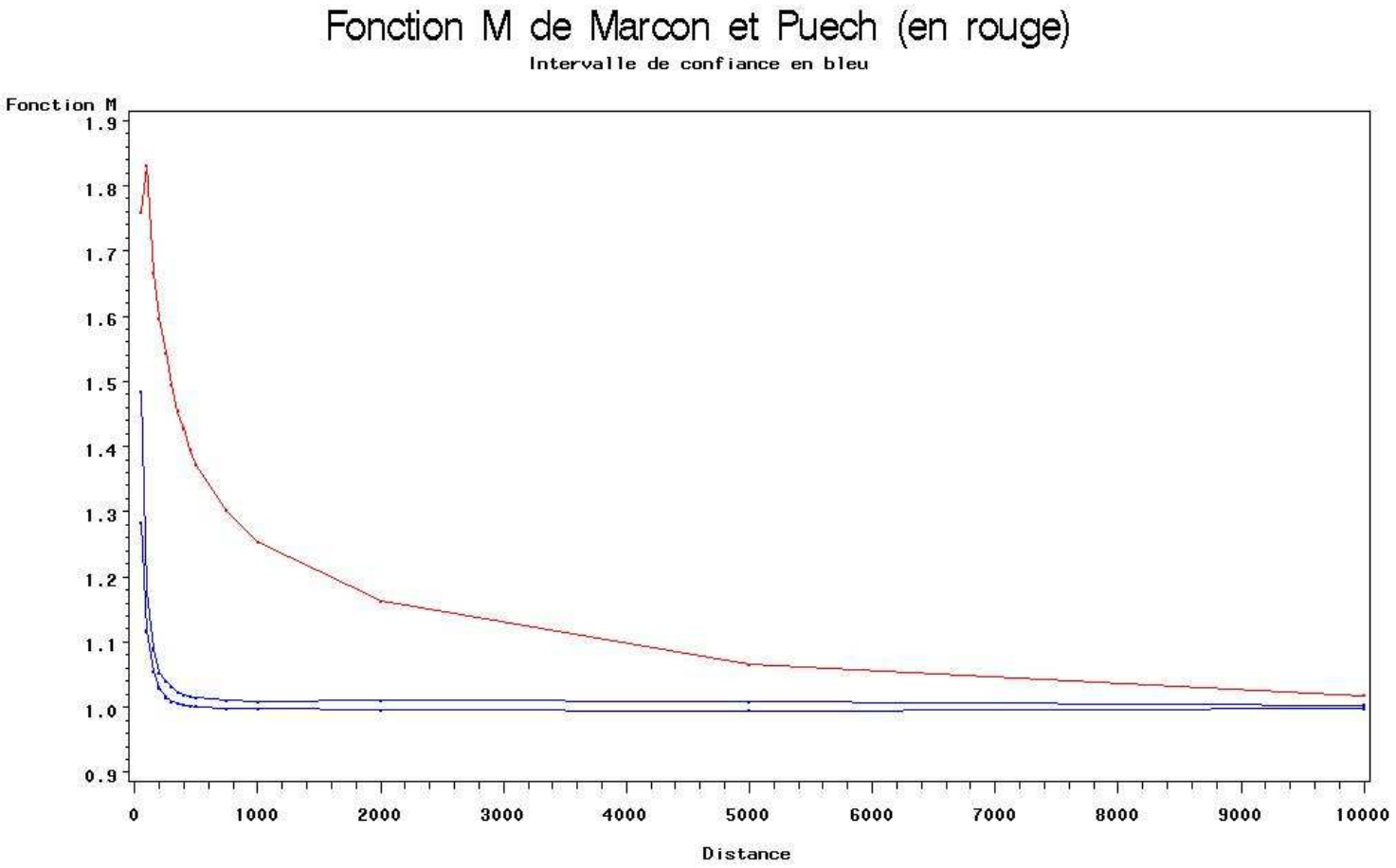
Lecture

- La fonction M est une mesure de concentration relative
- La référence pour la lecture de la fonction M est la valeur 1

Un exemple dans un autre champ

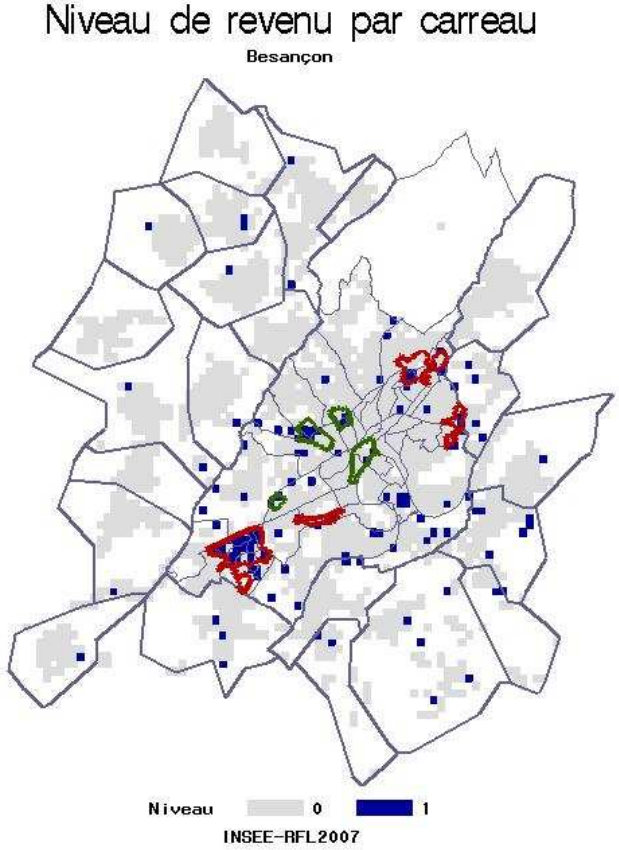


La fonction M pour les bas revenus

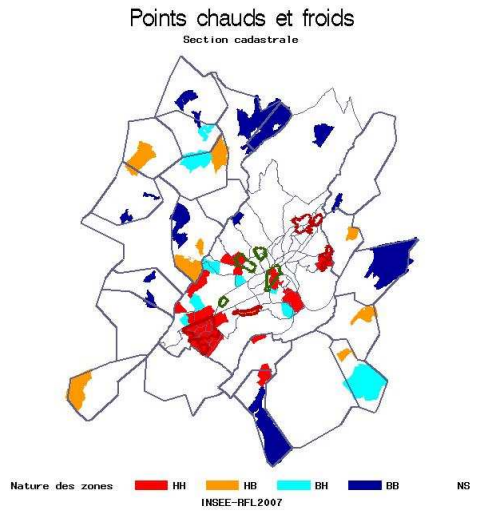


Le repérage des zones

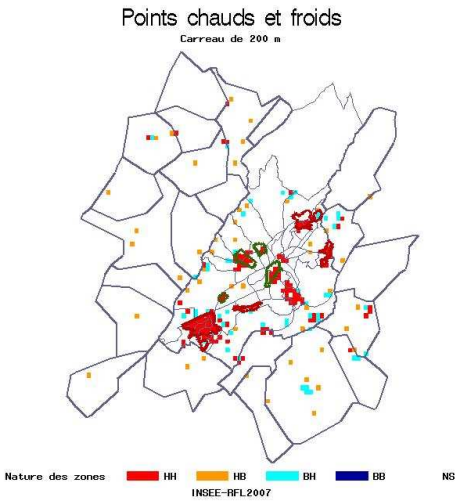
- Valeur locale Mi>1.8, à 100 m de distance



Conclusion



Lisa section

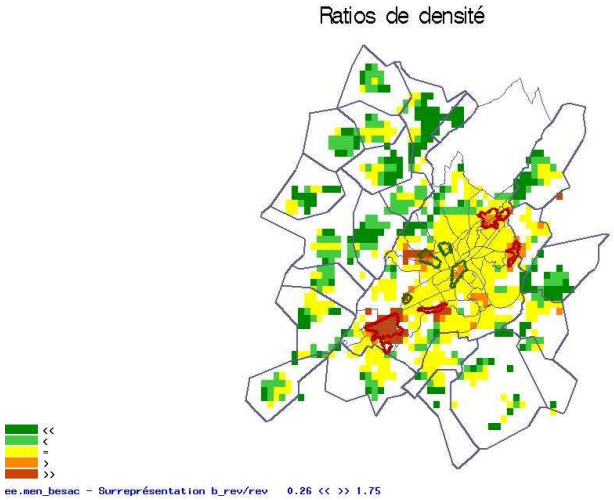


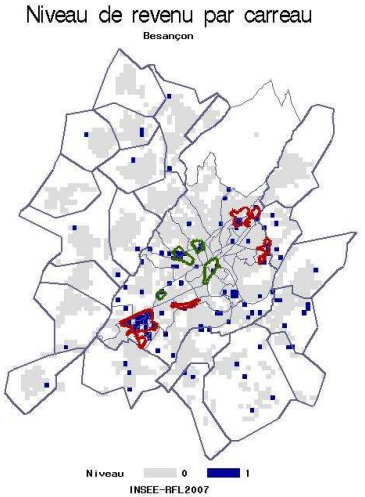
Lisa carreau

Bas revenus

Ratios

Marcon & Puech





Quelques remarques

- Trois représentations cohérentes, ne laissant pas de côté de zones bien identifiées de difficultés sociales
- Représentation plus approximative dans le cas des sections cadastrales
- Les Lisa, qui n'utilisent que des taux peuvent contribuer à l'analyse. La disponibilité de données sur un maillage fin permet de meilleurs résultats.

Merci de
votre
attention

Merci de votre attention