Autocorrélation spatiale locale

Statistiques locales de Moran : LISA

Statistiques locales de Moran

- Statistiques locales de Moran

$$I_i = \frac{(x_i - \mu)}{m_0} \sum_j w_{ij}(x_j - \mu) \qquad \text{avec } m_0 = \sum_i (x_i - \mu)^2 / n$$

$$\sum_i I_{i,t} = \frac{1}{m_0} \sum_i (x_{i,t} - \mu_t) \sum_j w_{ij}(x_{j,t} - \mu_t) \quad \Rightarrow \quad I = \sum_i I_i / S_0$$

  - Pour une matrice de pondérations spatiales standardisées en lignes $S_0 = n$:

$$\Rightarrow I = \frac{1}{n} \sum_i I_i$$

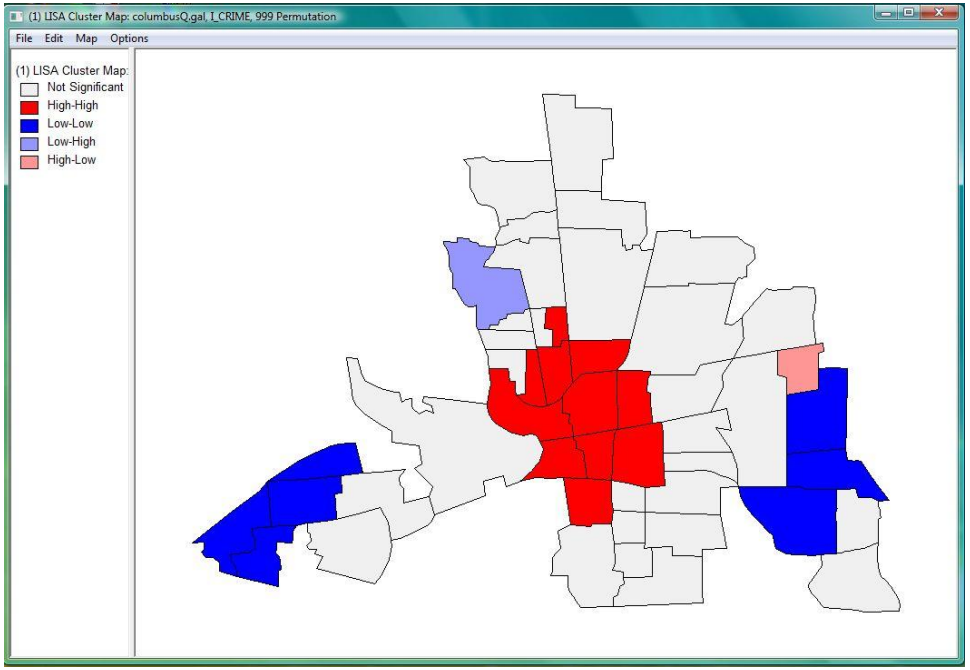
- Inférence
  - hypothèse de randomisation
  - permutation conditionnelle $\Rightarrow$ pseudo-significativité

    la valeur x_i est constante pour toute unité spatiale i, les autres valeurs sont aléatoirement permutées sur toutes les unités spatiales
  - dépendance locale ou hétérogénéité ?

- Visualisation
  - Carte des LISA et Carte de significativité de Moran

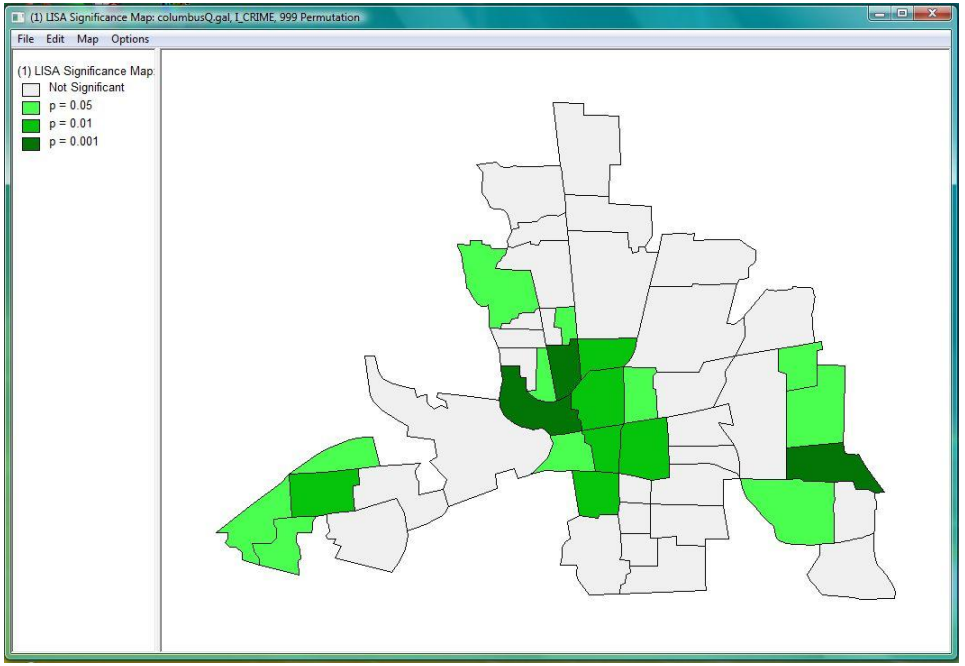
Cartes des LISA

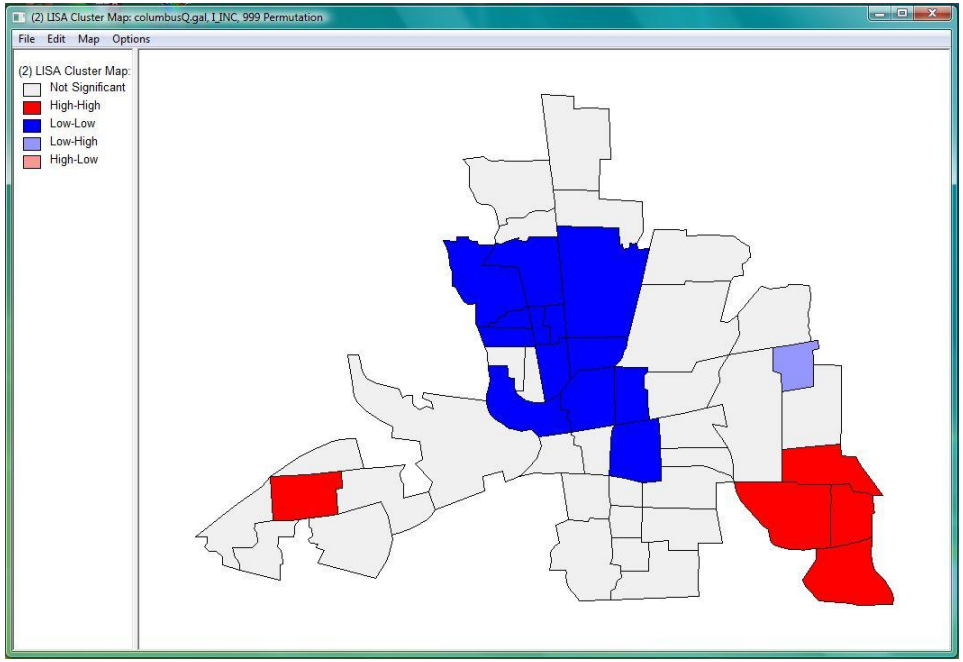
Crime à Columbus



Carte des clusters significatifs à 5%

Carte de significativité des LISA



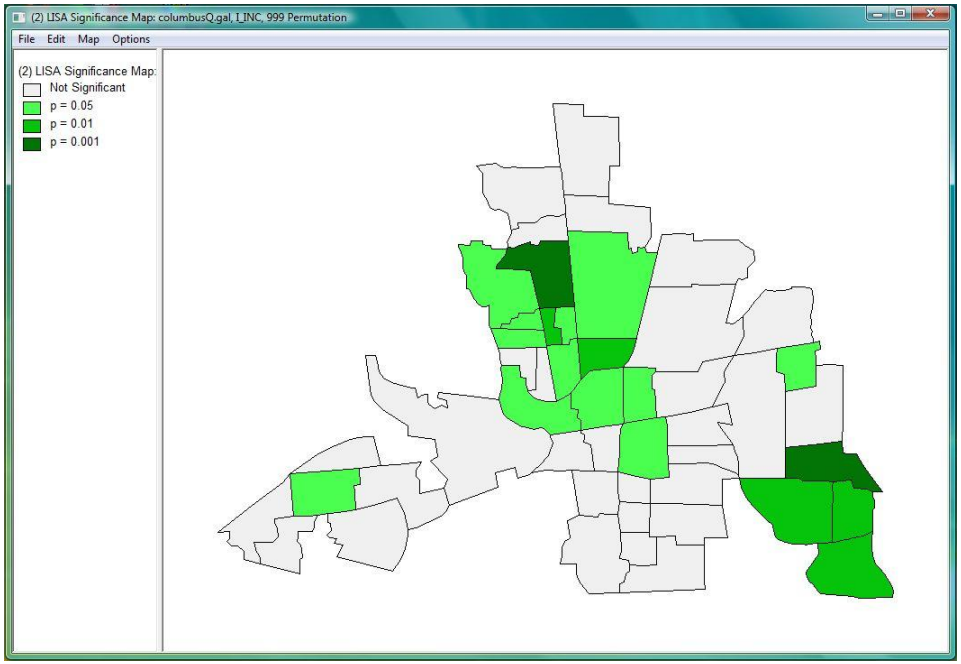


Cartes des LISA

Revenu à Columbus

Carte des clusters significatifs à 5%

Carte de significativité des LISA



Interprétation et limites (1)

- Interprétation des LISA
  - Positif : clusters de valeurs similaires (élevées ou faibles)
  - Négatif : clusters de valeurs dissimilaires
    - permet d'évaluer l'absence d'aléas spatial
    - permet de suggérer des structures spatiales significatives

- Problèmes pour l'inférence statistique
  - comparaisons statistiques multiples : les statistiques locales entre deux localisations sont corrélées lorsque le voisinage de ces deux localisations contient des éléments communs
  - pseudo niveau de signification de Bonferroni $\Rightarrow$ a/*m* avec *m* = *n* ou *k*
  - en présence d'autocorrélation globale

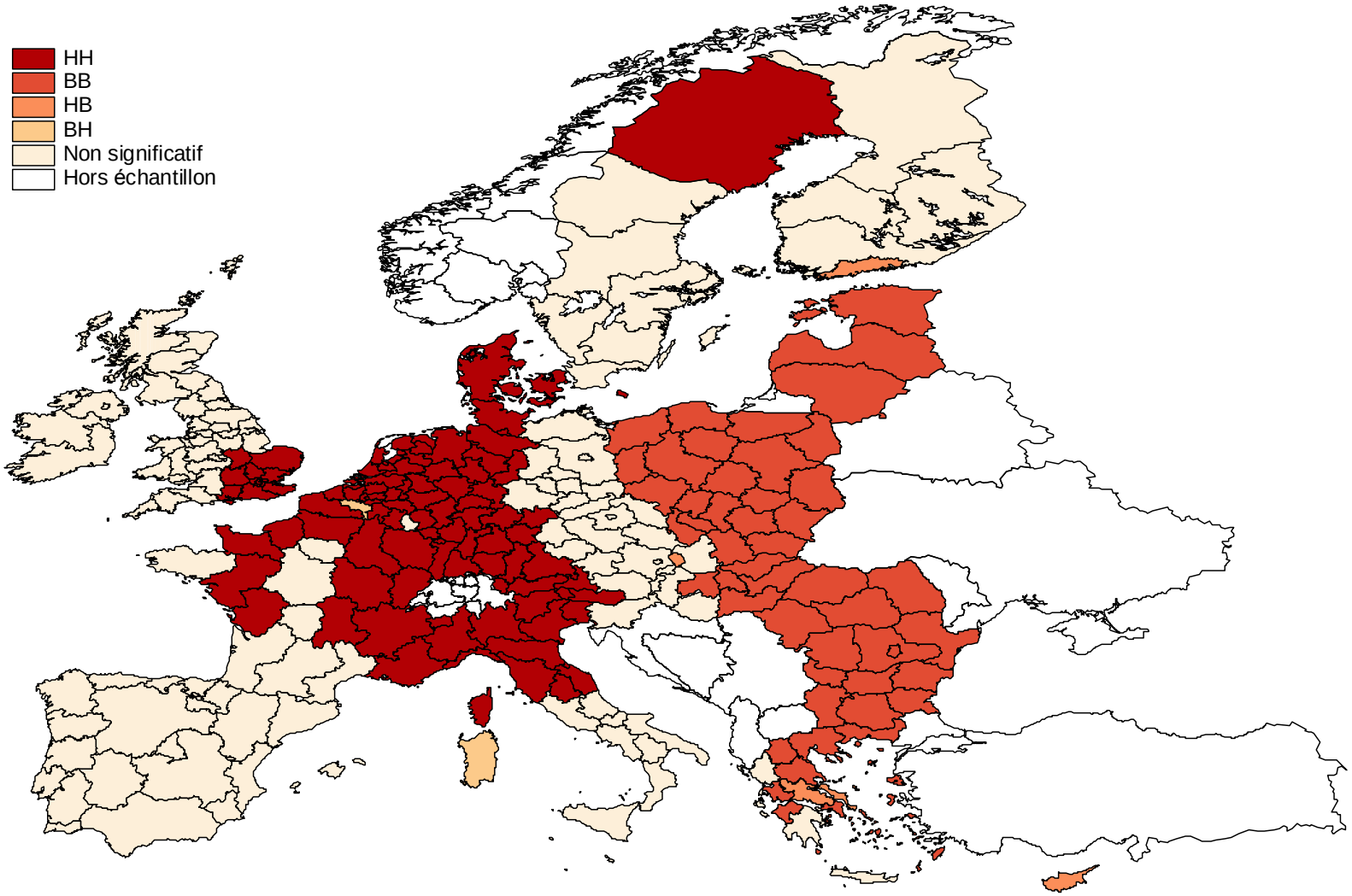
Interprétation et limites (2)

- Association multivariée
  - L'autocorrélation spatiale univariée peut résulter de :
    - association multivariée
    - problème d'échelle spatiale
  - Nécessité de contrôler l'effet d'autres variables

    $\Rightarrow$ régression spatiale

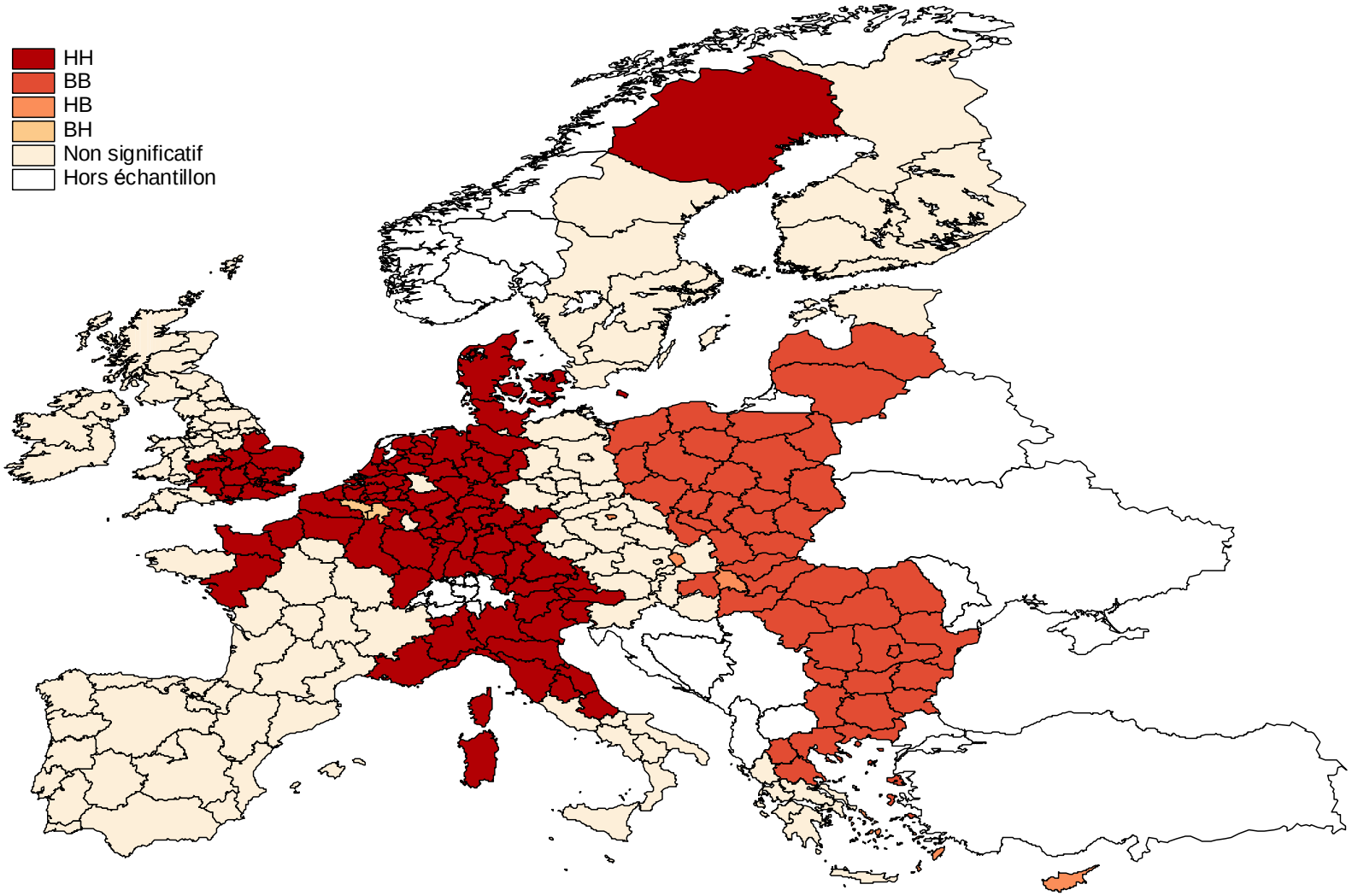
- Concentration LISA et "points chauds"
  - suggère des localisations intéressantes
  - n'explique pas

Statistiques Locales de Moran (1)



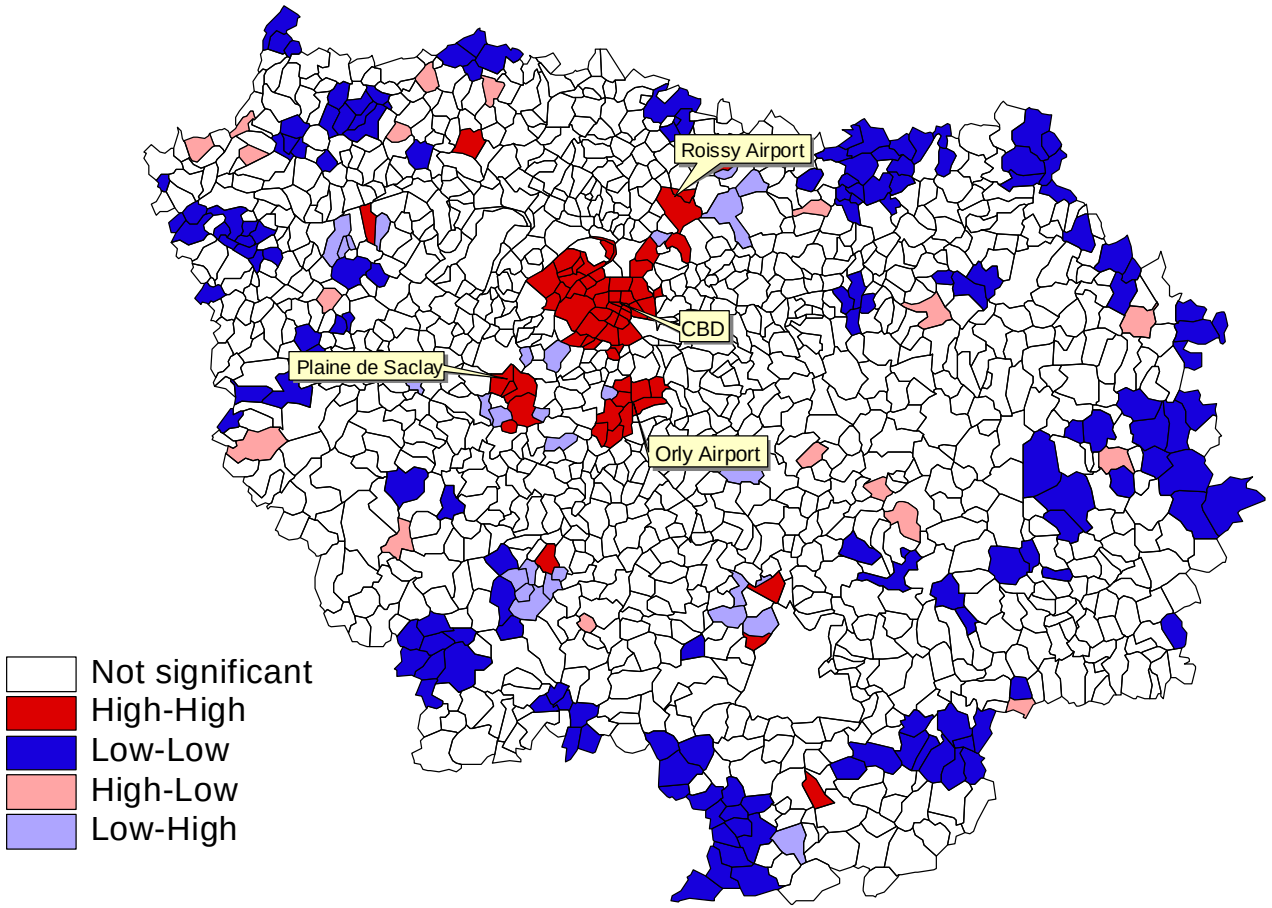
Statistiques locales de Moran significatives à 5%
PIB par tête en logarithmes et en SPA pour 1995 et pour l'Europe des 27

Statistiques Locales de Moran (2)



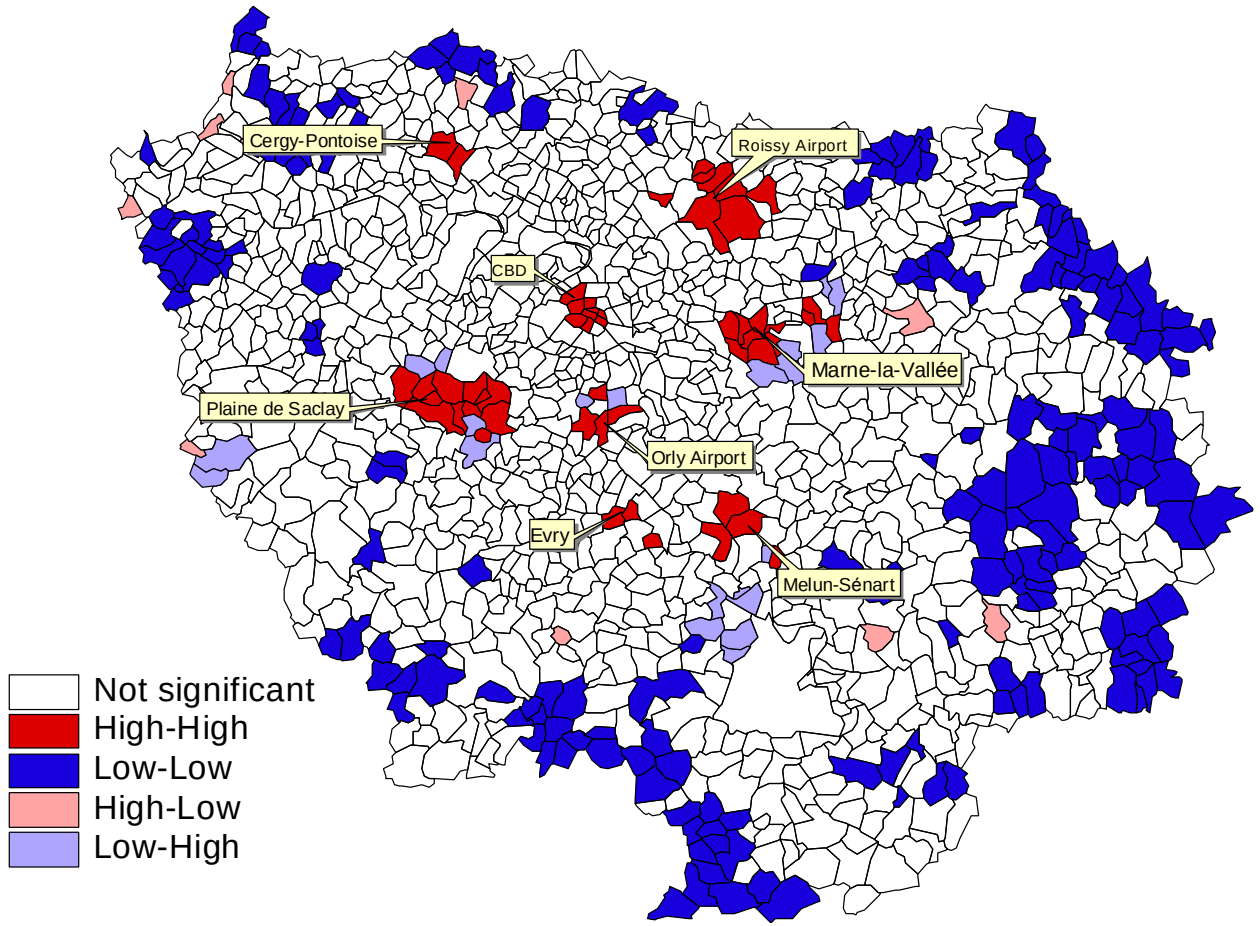
Statistiques locales de Moran significatives à 5%
PIB par tête en logarithmes et en SPA pour 2000 et pour l'Europe des 27

Statistiques Locales de Moran (3)



Statistiques locales de Moran significatives à 5%
Ratio emploi sur population en Ile de France, 1978

Statistiques Locales de Moran (3)



Statistiques locales de Moran significatives à 5%
Ratio emploi sur population en Ile de France, 1997

LISA généralisés

- Généralisation des statistiques locales de Moran

$$z_{1i} \times \sum_j w_{ij} z_{2j}$$

  - z_1 et z_2 sont des variables différentes
  - ou même variable pour des périodes différentes

- Inférence
  - Hypothèse nulle
    - affectation aléatoire de la valeur de z_1 en i,t aux valeurs « voisines » de z_2

Schémas dans l'Espace-Temps

- Concentration dans l'espace-temps = contagion
  - valeur élevée (par rapport à la moyenne) en une localisation i, t entourée de valeurs élevées en $j, s \neq t$
    - Comparé à élevé-élevé pour le même t
  - Similaire pour faible-faible
- Outliers dans l'espace-temps = changement
  - Valeur élevée (par rapport à la moyenne) en une localisation i, t entourée de valeurs faibles en $j, s \neq t$
  - Similaire pour faible-élevé
- Significativité basée sur la permutation

Autocorrélation spatiale locale

Statistiques de Getis et Ord

Statistiques de Getis -Ord (1992)

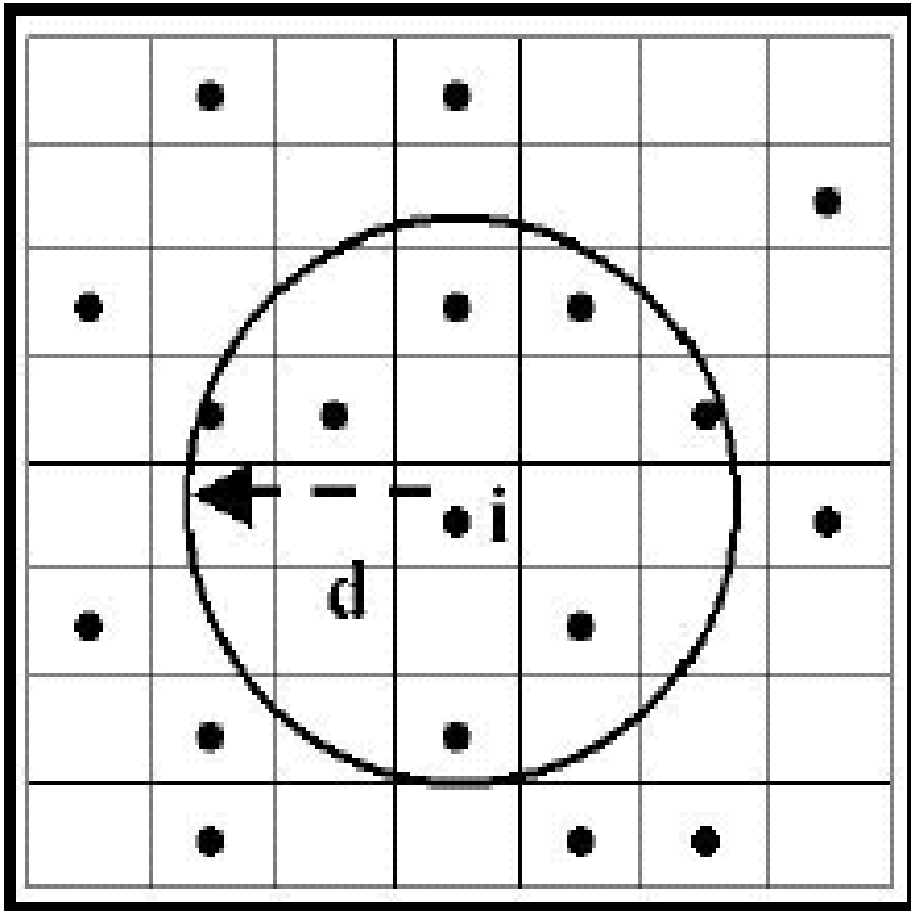
- Statistiques de « distance » pour l'association spatiale locale
- Statistique de Getis – Ord (1992) G_i et G_i^* :

$$G_i(d) = \sum_{j \neq i} w_{ij}(d) x_j \Big/ \sum_{j \neq i} x_j$$

  - une statistique pour chaque localisation
  - $G_i(d)$ dépend de d : voisinage au sens des bandes de distance : $w_{ij}(d)$ ou du nombre de plus proches voisins k : $G_i(k)$ avec $w_{ij}(k)$
  - à utiliser avec des matrices binaires
  - à utiliser avec des variables x à valeurs positives
  - G_i n'inclut pas l'observation i dans la somme
  - G_i^* inclut l'observation i dans la somme

Statistiques de Getis -Ord (1992)

- voisinage au sens des bandes de distance pour le calcul de $w_{ij}(d)$



Statistiques de Getis-Ord (1995)

- Généralisation
- Statistiques de Getis – Ord (1995) G_i et G_i^* :

$$G_i(d) = \frac{\sum_{j \neq i} w_{ij}(d) x_j - W_i \mu}{\sigma \left\{ \left[(n-1) S_{1i} - W_i^2 \right] / (n-2) \right\}^{1/2}}$$

avec $W_i = \sum_j w_{ij}$; $S_{1i} = \sum_j w_{ij}^2$ pour $j \neq i$

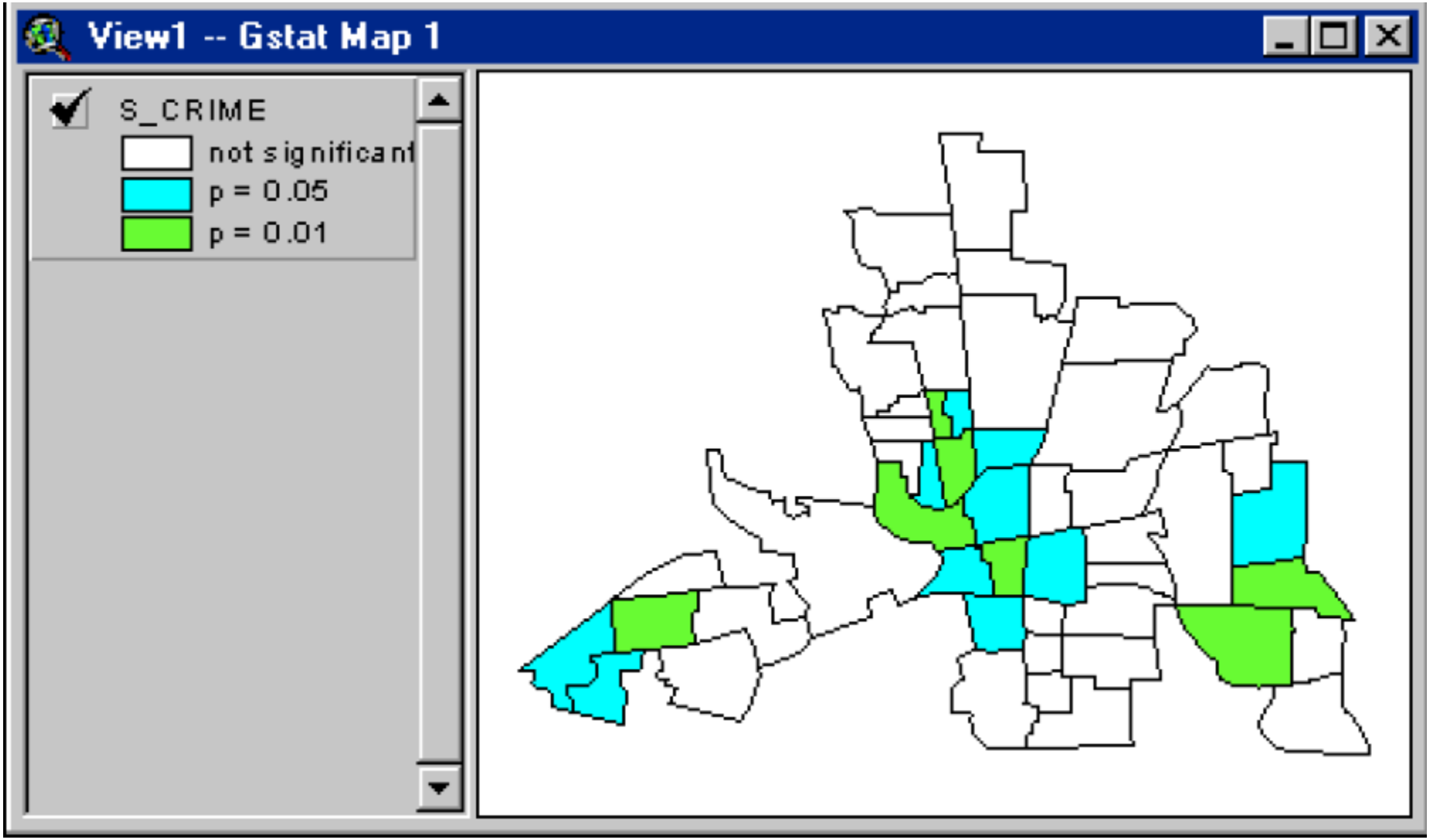
μ et σ sont respect. la moyenne et l'écart-type de l'échantillon de taille *n-1* excluant *i*

- extension à des matrices non binaires
- extension à toute variable *x*
- G_i n'inclut pas l'observation *i* dans la somme
- G_i^* inclut l'observation *i* dans la somme

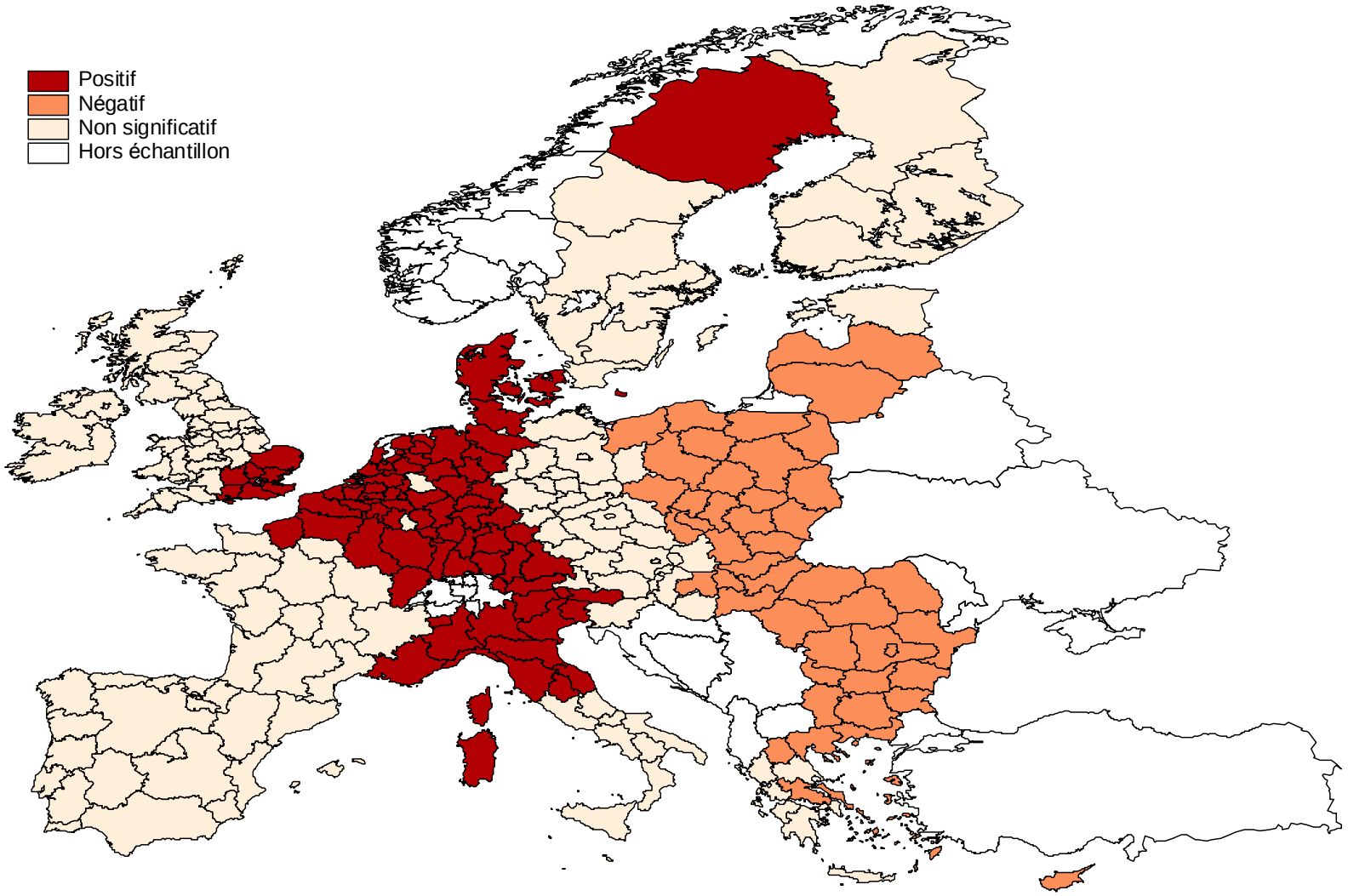
Interprétation des statistiques G_i

- Association spatiale locale
  - positive: clusters de valeurs élevées
  - négative: clusters de valeurs faibles
- Inférence
  - distribution asymptotique normale
- Visualisation
  - cartes des localisations avec des G_i ou G_i^* significatives
- Problèmes
  - pour le calcul : choix du paramètre d
  - pour l'inférence statistique :
    - comparaisons statistiques multiples : les statistiques locales entre deux localisations sont corrélées lorsque le voisinage de ces deux localisations contient des éléments communs
    - niveau de pseudo significativité de Sidák $\Rightarrow$ $1 - (1 - \alpha)^{1/m}$ avec $m = n$ ou k
    - en présence d'autocorrélation spatiale globale

Carte de significativité des G_i^*

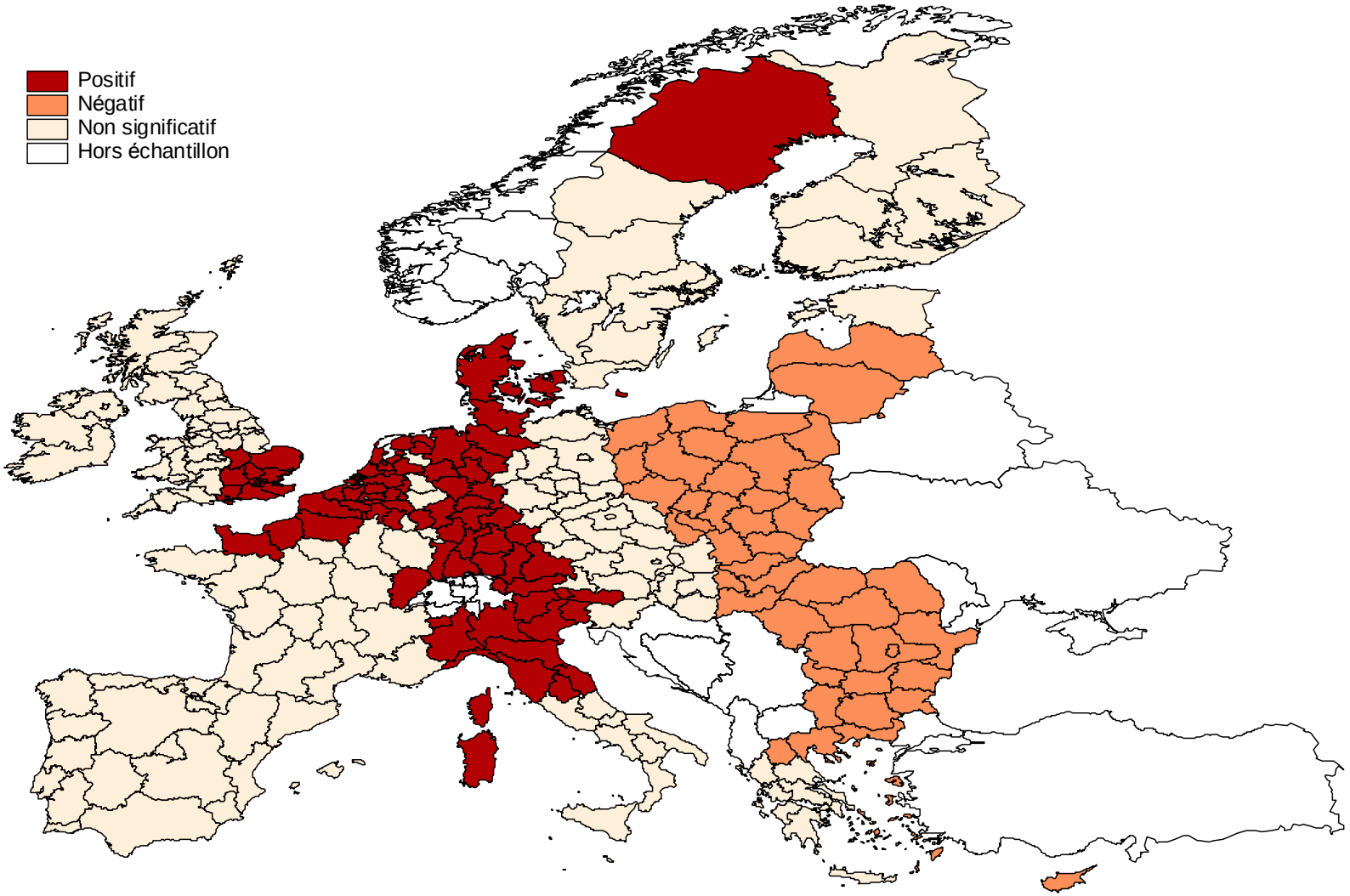


Statistiques de Getis-Ord $G_i(10)$ (1)



Statistiques de Getis-Ord significatives à 5%
PIB par tête en logarithmes et en SPA pour 1995 et pour l'Europe des 27

Statistiques de Getis-Ord $G_i(10)$ (2)



Statistiques de Getis-Ord significatives à 5%
PIB par tête en logarithmes et en SPA pour 2000 et pour l'Europe des 27

Le cas particulier des taux ou proportions

Le cas particulier des taux ou proportions

- La proportion comme estimateur du risque
  - Souvent, les taux ou les proportions ne sont pas des variables d'intérêt mais servent à estimer le risque sous-jacent
    - Par exemple, on cherche à connaître quel est le risque de mourir d'un type particulier de cancer
    - Afin d'estimer ce risque, on divise le nombre de personnes effectivement mortes de ce cancer par la population à risque

- La proportion p est l'estimateur du maximum de vraisemblance du « vrai » risque π (taux de mortalité brut : raw rate)

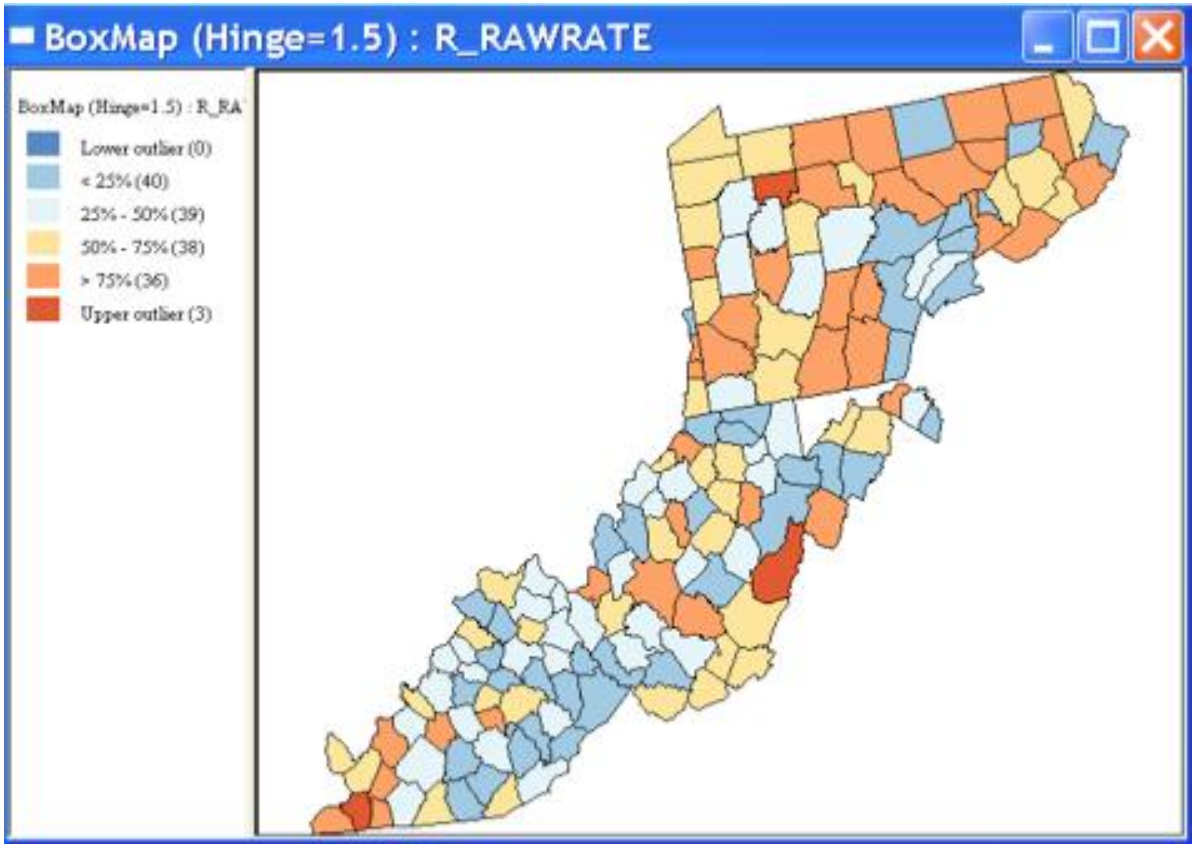
$$p = \left(E / P \right) \times S$$

E : nombre « d'évènements »

P : population à risque

S : facteur d'échelle (1 pour 100 000 etc.)

Taux brut



Taux brut de cancer de la prostate dans les comtés des Appalaches (Anselin, 2005)

Ajustement du risque

- Questions de recherche
  - Hétérogénéité spatiale
    - La répartition spatiale joue un rôle
    - Le risque n'est pas uniforme à travers l'espace
    - Le risque dépend de facteurs explicatifs
  - Clusters, concentrations spatiales
    - Caractéristiques générales / globales ou caractéristiques spécifiques / locales
    - On s'intéresse aux déterminants d'un risque élevé

- Ajustement du risque
  - Le taux brut peut seulement refléter des caractéristiques locales de la distribution de la population : nécessité de contrôler par rapport à d'autres facteurs explicatifs
  - Par exemple : effet sexe, effet âge
    - Taux de cancer de la prostate élevé dans un département dont la population est essentiellement constitué d'hommes âgés
  - Il faut procéder à un ajustement du taux brut

Taux ajusté

- Ajustement pour l'âge
  - Proportion de la classe d'âge m dans la population globale

$$\pi_m = P_{mG} / P_G$$

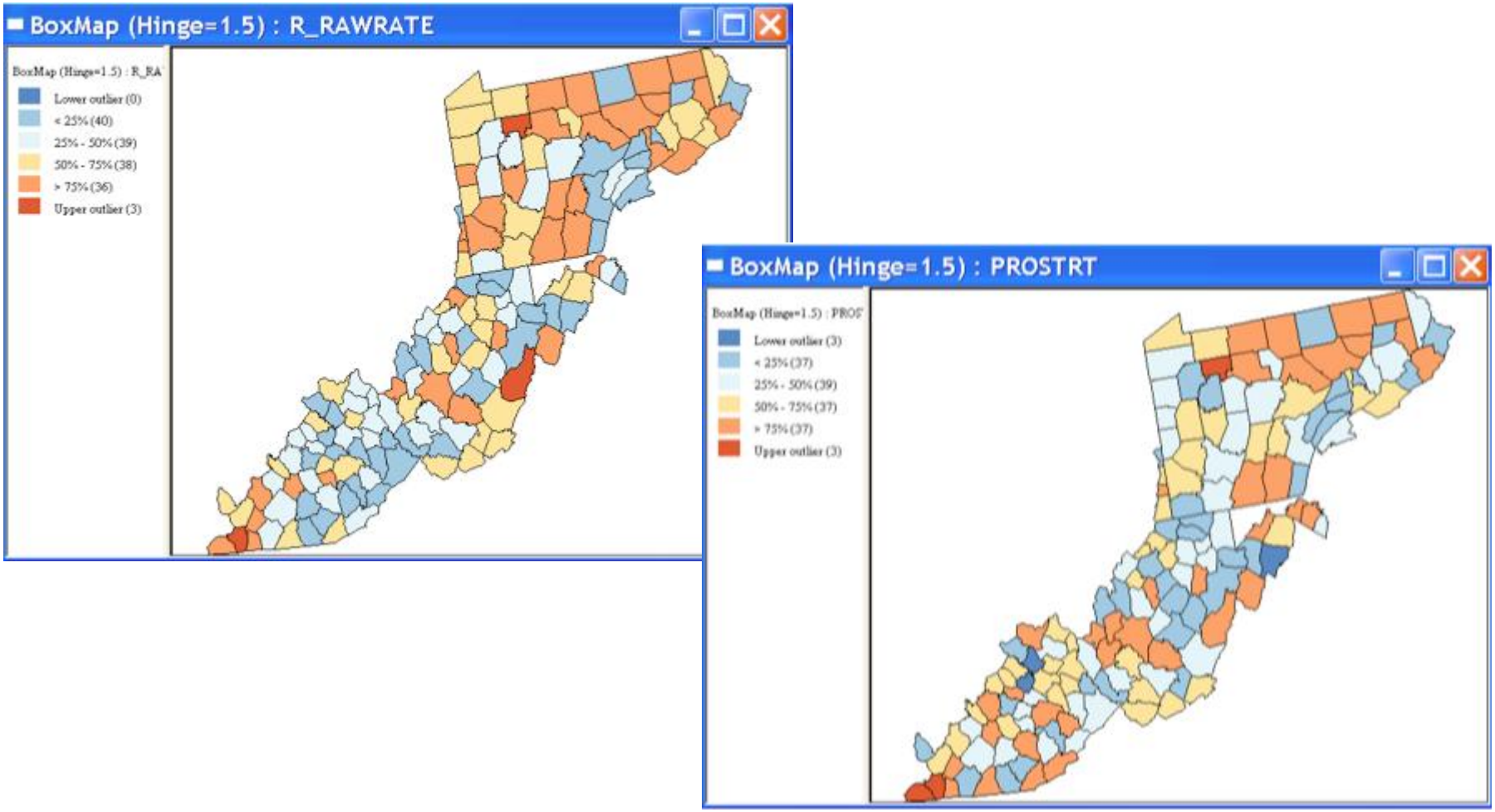
  - Taux brut non ajusté pour l'unité spatiale *i*

$$r_i = \sum_m E_{i,m} / P_{i,m}$$

  - Taux brut ajusté pour l'unité spatiale *i*

$$r_{i,adj} = \sum_m \pi_{i,m} E_{i,m} / P_{i,m}$$

Taux brut et taux ajusté



Taux de cancer de la prostate brut et ajusté dans les comtés des Appalaches (Anselin, 2005)

Taux de mortalité standardisé

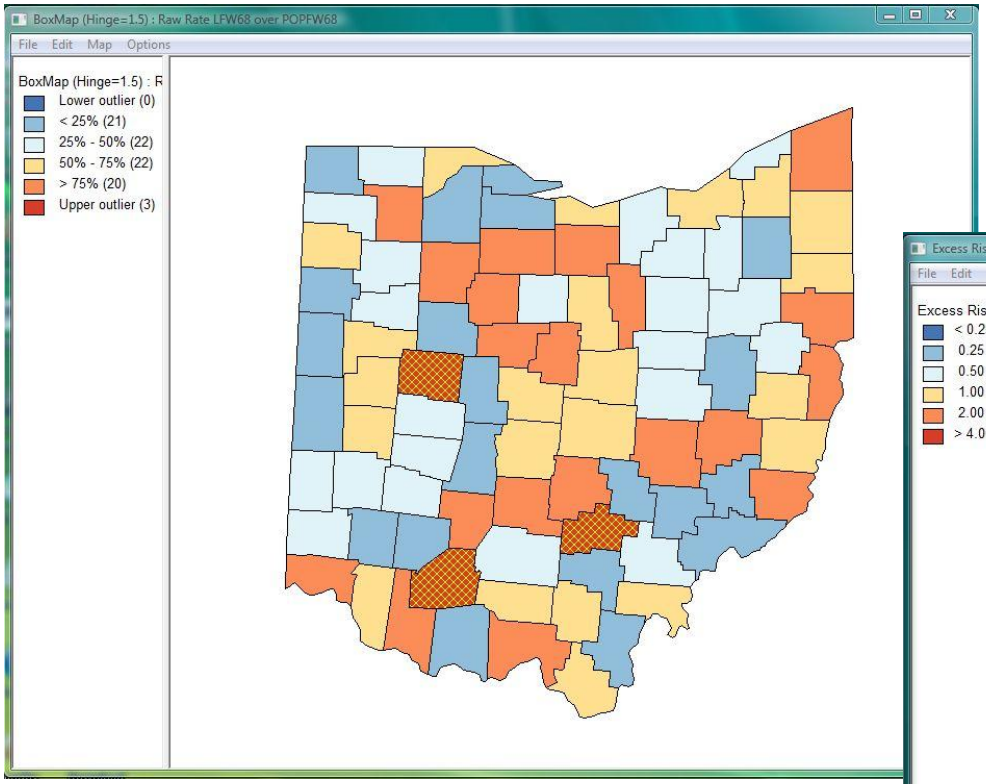
- Taux de mortalité standardisé
- Comparaison évènements observés et évènements espérés
  - Risque moyen pour la population de référence :

    risque moyen = nb total d'évènements / population totale
  - Le nb d'évènements espérés est calculé à l'aide du risque moyen

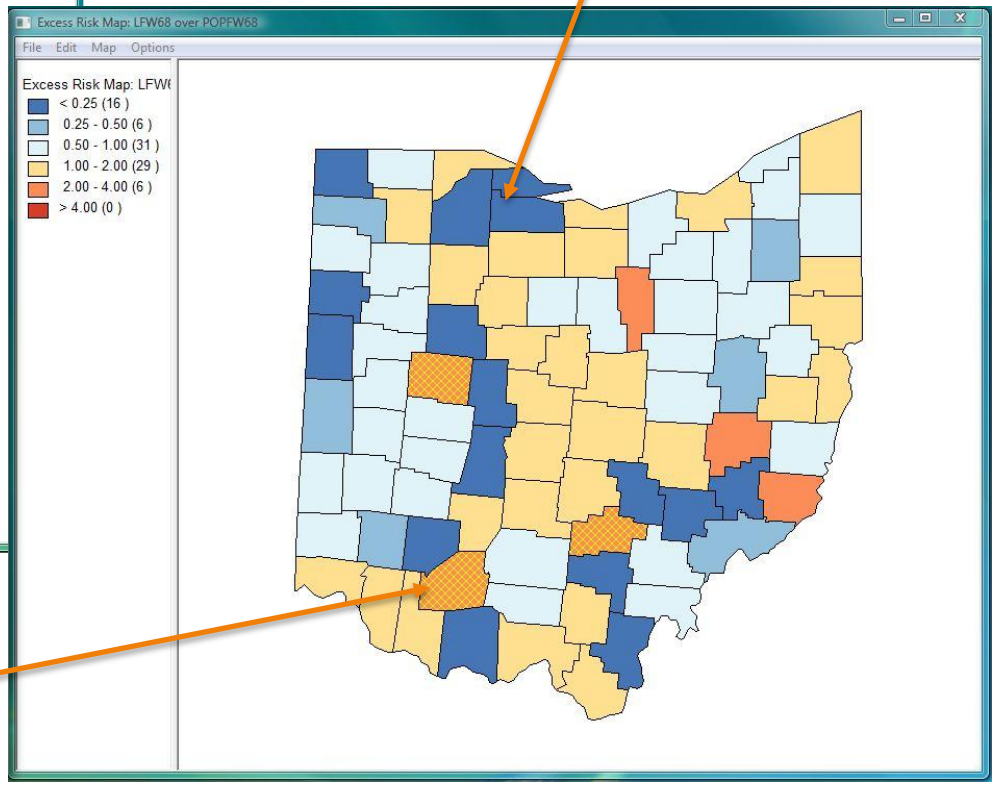
    nb d'événement espéré = risque moyen × population
  - On calcule le risque relatif comme le ratio du nombre d'évènements observés sur le nombre d'évènements espérés

    risque relatif = nb d'événement observés / nb d'événements espérés
- Interprétation :
  - Des valeurs du risque relatif inférieures à 1 indiquent les localisations où le nombre d'évènements est inférieur au nombre espéré
  - Des valeurs du risque relatif supérieures à 1 indiquent les localisations où le nombre d'évènements est supérieur au nombre espéré
  - **Excess risk map**

Box map et Excess risk map



En ton bleu : les comtés où le risque est inférieur au risque moyen de l'Etat d'Ohio



En ton orangé : les comtés où le risque est supérieur au risque moyen de l'Etat d'Ohio

Taux de mortalité brut et taux de mortalité standardisé (Excess risk) dû au cancer du poumon dans les comtés de l'Ohio (USA) en 1968

Propriétés de l'estimateur

- Propriétés de l'estimateur

$$p = (E / P)$$

$$E(E / P) = \pi \quad \Rightarrow \quad \text{estimateur centré}$$

$$V(E / P) = \pi(1-\pi) / P$$

- Deux problèmes :
  - Dépendance la variance par rapport à l'espérance
    - La variance dépend de l'espérance inconnue π, le vrai risque
    - Il faut utiliser des transformations
  - Instabilité de la variance
    - La variance est une fonction inverse de la taille de la population à risque P
      - Plus P est petit, plus la variance est importante et moins l'estimateur est précis
      - Plus P est grand, plus la variance est faible et plus l'estimateur est précis
      - Si P varie beaucoup, comparaison difficile voire impossible entre les unités
      - Les observations apparaissant comme « extrêmes » peuvent être fallacieuses : elles peuvent simplement être une conséquence de la variabilité de la précision des estimateurs

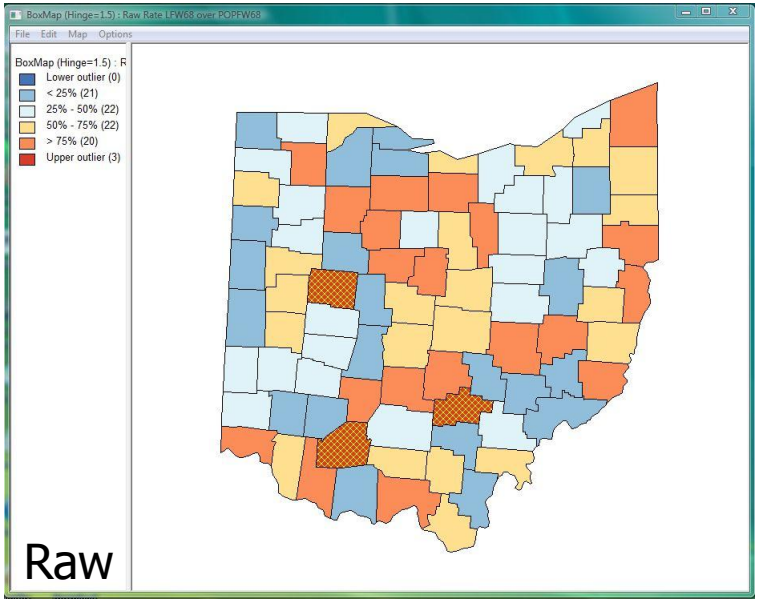
Problèmes sur la variance - Transformations

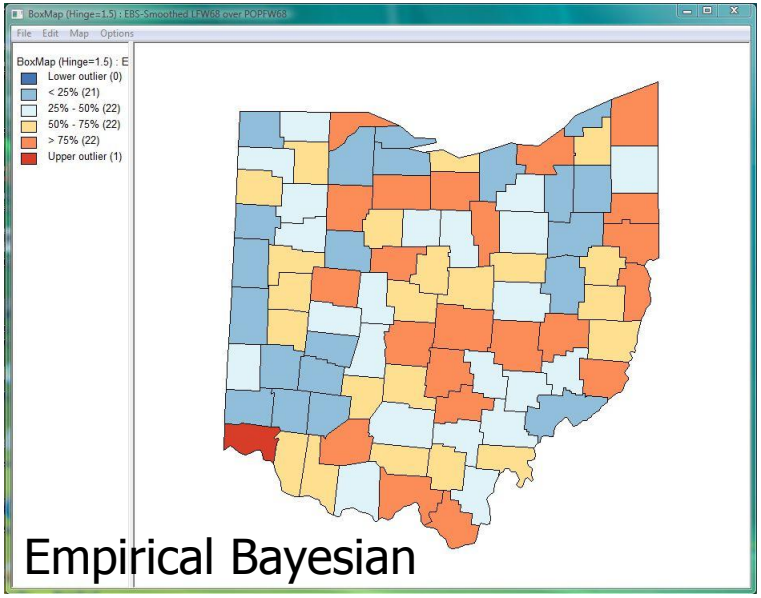
- Transformation
  - Le principe est d'obtenir une nouvelle variable aléatoire qui ne souffre pas du problème de dépendance
  - Cependant, l'échelle est différente et l'interprétation parfois difficile
  - Exemples :
    - Transformation de Freeman-Tukey (1950)
    - Standardisation ArcSin (Anscombe, 1948)
    - Standardisation Anscombe (1948)
    - Standardisation bayésienne empirique (Assunção et Reis, 1999) : méthode de correction de la statistique I de Moran lorsque les densités de population varient selon les observations et que la variable d'intérêt est une proportion

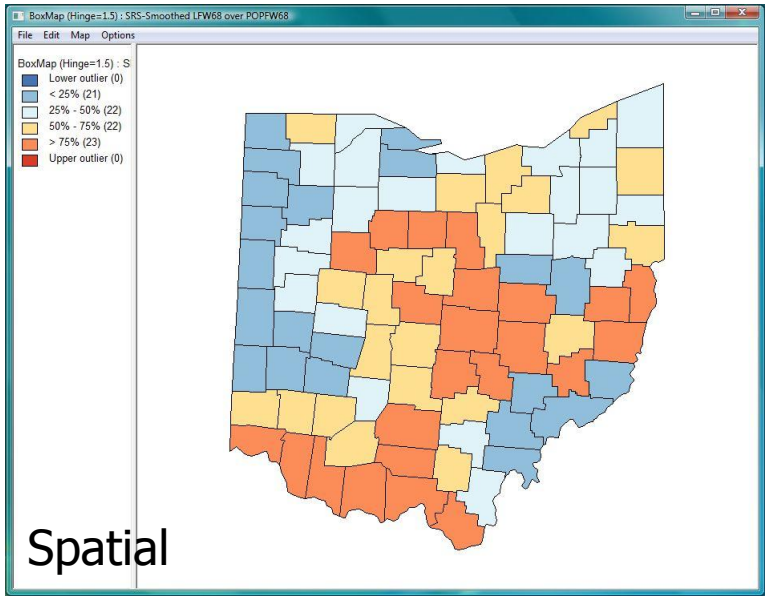
Problèmes sur la variance - Lissages

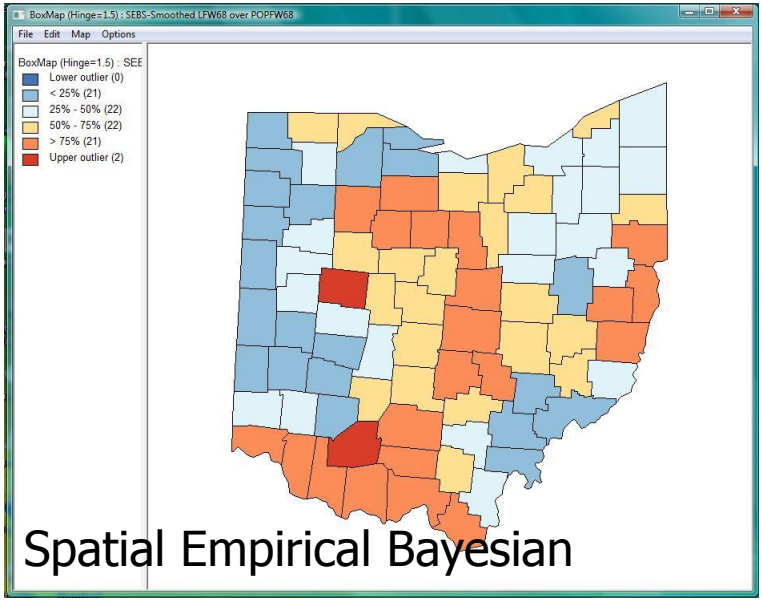
- Lissage
  - Le principe est d'ajuster l'estimateur du risque dans une unité spatiale en utilisant l'information fournie par les autres unités spatiales
  - Différents ajustements possibles
    - Lissage bayésien empirique
      - On utilise l'information de tout l'échantillon
      - Le taux brut est déplacé vers la moyenne globale, le déplacement étant une fonction inverse de la variance
      - Si la variance est faible (population importante) : peu de changement
      - Si la variance est élevée (population faible) : changement plus important
    - Lissage spatial
      - On calcule l'estimateur du risque pour une unité à l'aide des informations associées à cette unité mais aussi des unités voisines
      - Utile pour lisser les données mais pas pour identifier les observations extrêmes
    - Lissage spatial bayésien empirique
      - Même principe que le lissage bayésien empirique mais la référence n'est pas la moyenne globale mais la moyenne des observations voisines

Lissages – Box-maps









Quelques problèmes méthodologiques

Quelques problèmes méthodologiques (1)

- Les résultats sont conditionnés par le choix de la matrice de poids

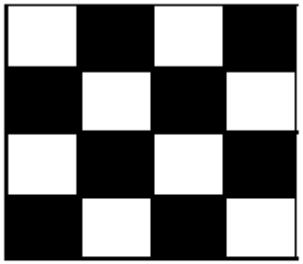


Figure 2.3 : Echiquier
Source : Cliff et Ord (1973, p. 16)

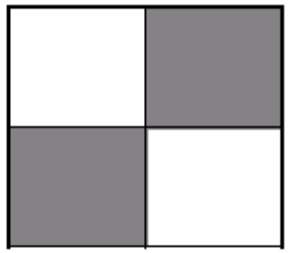
Matrice de contiguïté

Tour : autocorrélation spatiale < 0

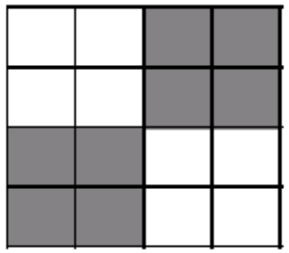
Fou : autocorrélation spatiale > 0

Dame : pas d' autocorrélation spatiale

- Le problème MAUP (Modifiable Areal Unit Problem)
  - Effet de forme
  - Effet d'échelle



$I = -1$ (4 régions)



$I = 0{,}33$ (16 régions)

Quelques problèmes méthodologiques (2)

- Dissonance entre échelle de mesure d'une variable et processus sous-jacent
- Dissonance entre collecte d'informations spatiales (points) et variables spatiales (aires)
- Illusion ou biais écologique

  ⇒ Probs. liés à l'agrégation des données individuelles sur des unités spatiales arbitraires d'origine administrative

- Conclusion : la structure et la collecte des données qui contiennent une dimension spatiale nécessitent l'élaboration d'une nouvelle perspective économétrique