Économétrie spatiale sur données d'enquête

Thomas Merly-Alpa & Raphaël Lardeux

13èmes Journées de Méthodologie Statistique
Session 14 : Autocorrélation spatiale

13 juin 2018

Plan

1. Problématique
2. Modèles d'économétrie spatiale
3. Application : la production industrielle dans les Bouches-du-Rhône
4. Conclusions

Remerciements

Cette présentation s'inscrit dans le cadre du Manuel d'Analyse Spatiale publié par le Département des Méthodes Statistiques de l'INSEE, et en décrit le chapitre 11.

Nous remercions Marie-Pierre de Bellefon et Vincent Loonis pour la coordination du manuel.

Problématique

Pourquoi l'approche spatiale ?

- Les individus ne se positionnent pas au hasard sur un espace géographique.
- Phénomènes d'**autocorrélation spatiale** et d'**hétérogénéité spatiale**.
- Dynamiques de **diffusion spatiale** et d'**interaction spatiale**.
  Ex. Épidémies, migrations pendulaires.

$\hookrightarrow$ Étude de l'espace, non pas tant comme facteur explicatif, mais en tant que dimension intrinsèque de l'analyse économétrique.

Problématique

Pourquoi une économétrie particulière ?

- Il n'y a plus d'indépendance entre les individus.
- Les phénomènes spatiaux affectent les relations estimées.
- Risque de biais sur les paramètres des modèles.
- Interactions **multidirectionnelles** à estimer.

Problématique

Dans cette présentation, on se pose la question de faire de l'économétrie spatiale sur données d'enquête :

- Certaines informations ne sont disponibles que via des enquêtes sociales ;
- Applications plus larges : échantillonnage pour des données issues de réseaux sociaux.

Mais des questions se posent :

→ Peut-on retrouver des dynamiques spatiales à partir d'une estimation sur un échantillon ?

→ Est-il possible d'estimer un modèle d'économétrie spatiale à l'aide des résultats à une enquête par sondage ?

Problématique

Les données d'enquête proviennent d'un échantillon issu d'un sondage probabiliste :

- Ce processus de sondage est déterminé indépendamment de l'estimation (contrairement aux forages en géostatistique).
- Le taux de sondage est faible, il y a donc beaucoup de valeurs manquantes.
- Les individus échantillonnés peuvent être dispersés sur le territoire (sauf pour certaines enquête).

$\hookrightarrow$ L'étude sur un échantillon ne permet pas de se placer dans les condition idéales pour l'estimation.

Nous identifions deux effets :

(i) Un « **effet taille** » résultant de la restriction à un sous-ensemble localement complet d'un territoire ;

(ii) Un **effet de répartition spatiale des observations**, lié à l'omission aléatoire au sein de ce territoire d'unités spatialement corrélées avec les unités observées.

Illustrons ces deux effets :

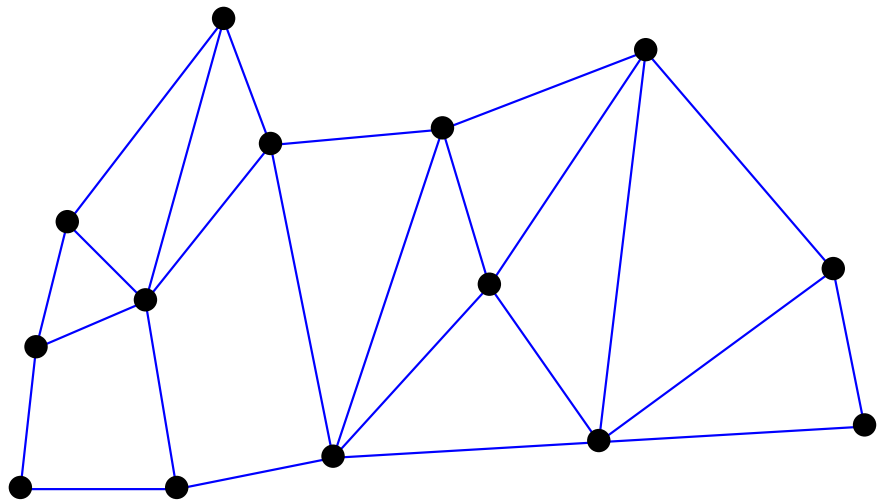


Figure – Graphe de voisinage d'une population ($n = 14$)

Un « effet taille » :

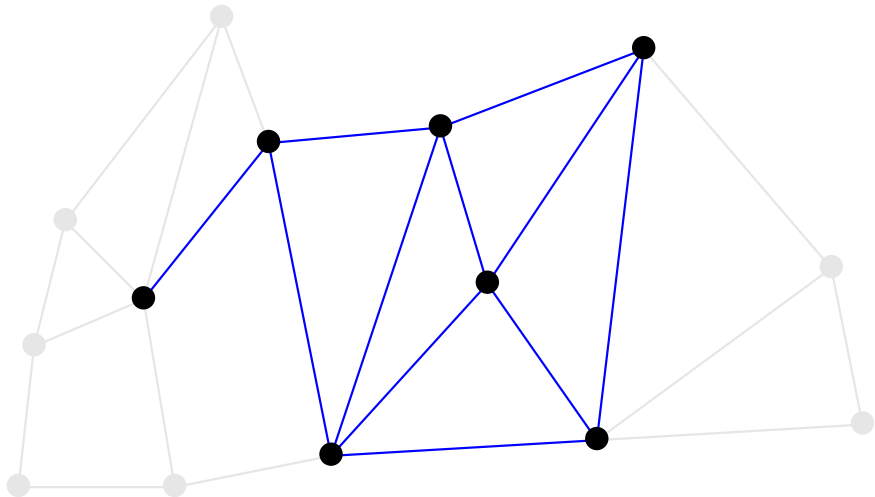


Figure – Échantillon localement complet ($n = 7$)

Auquel se rajoute un effet de répartition spatiale :

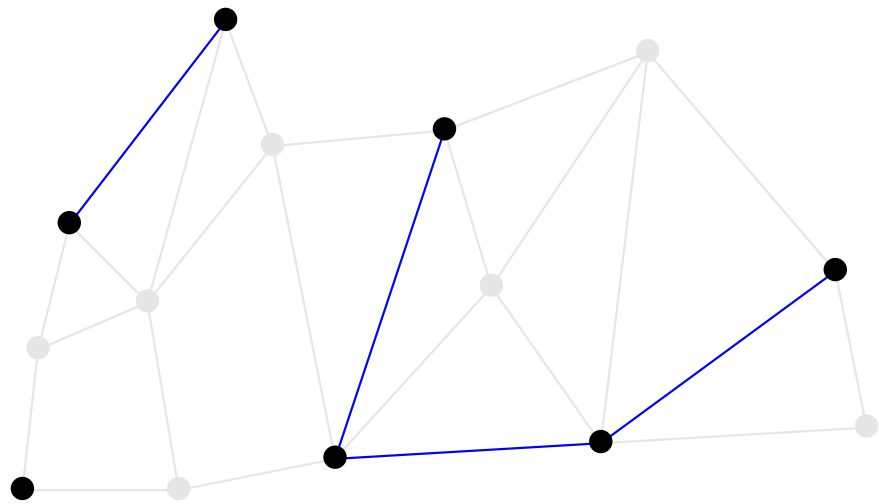


Figure – Échantillon réparti spatialement ($n = 7$)

Plan

1 Problématique

2 Modèles d'économétrie spatiale

3 Application : la production industrielle dans les Bouches-du-Rhône

4 Conclusions

Comment appréhender la dimension spatiale des données ?

1. Définir des unités spatiales.
2. Définir une notion de « voisinage » : contiguïté, distance géographique, k plus proches voisins, distance culturelle,...
3. En découle une **matrice de pondération spatiale ($\mathcal{W}$)**
   - ↪ 1 si les unités i et j sont voisines, 0 sinon (contiguité, PPV).
   - ↪ distance entre i et j tant que ces unités ne sont pas trop éloignées l'une de l'autre.
4. Celle-ci permet de prendre en compte des dynamiques spatiales.

Il existe plusieurs choix de $\mathcal{W}$:

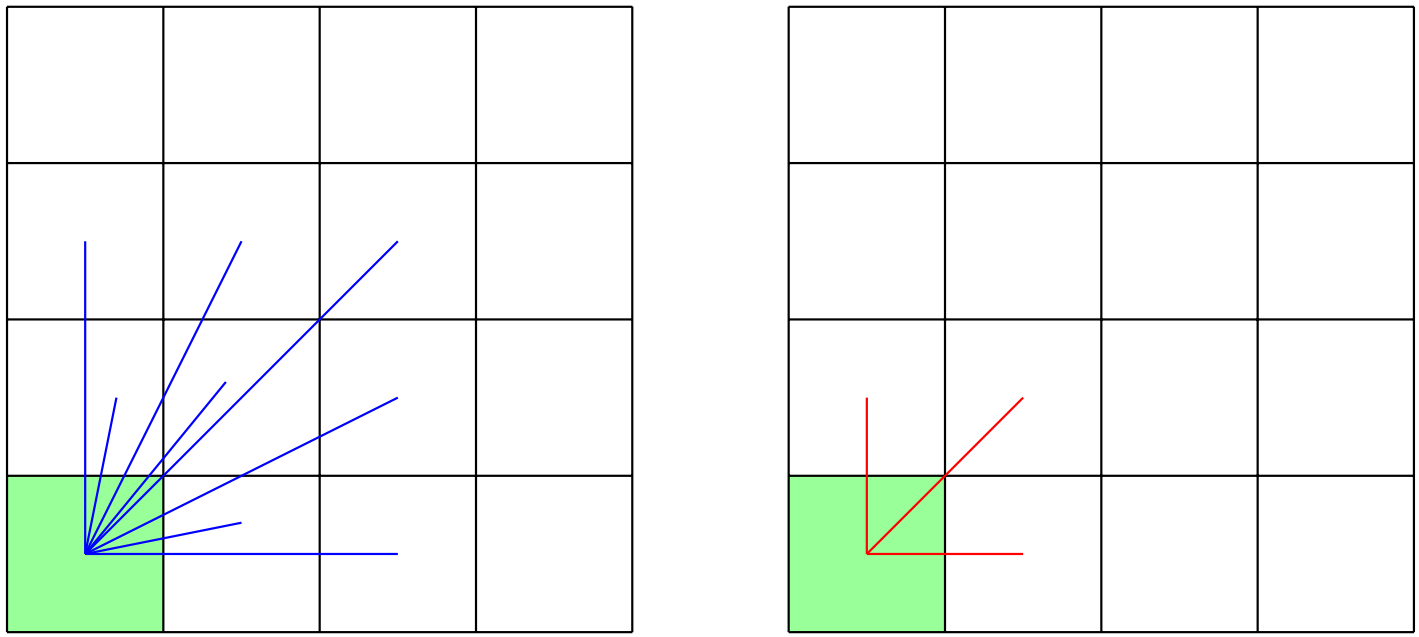


Figure – Gauche : matrice de distance bornée, Droite : 3 plus proches voisins

Quand on considère un échantillon S, on fait le choix de la matrice de distance bornée qui garde la structure de voisinage de la population générale :

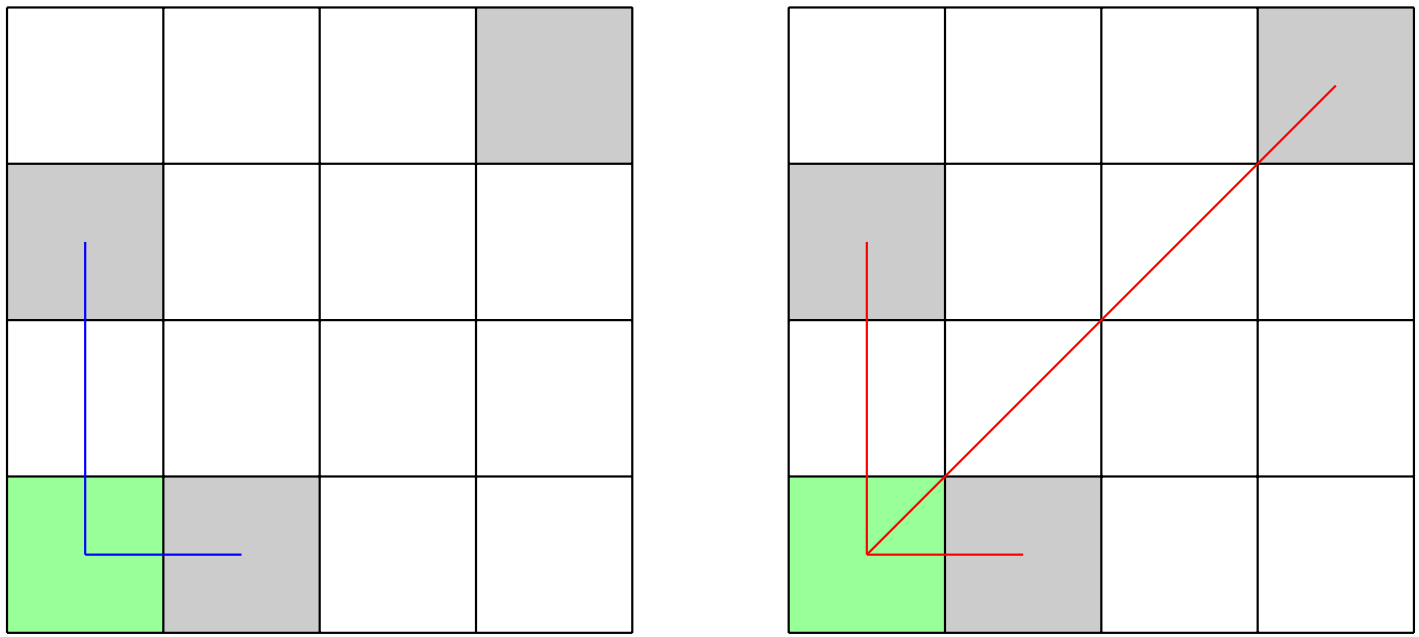


Figure – Gauche : matrice de distance bornée, Droite : 3 plus proches voisins

Deux modèles canoniques :

- « Spatial Auto-Regressive » model (SAR) : effet de diffusion + effet multiplicateur.

$$Y = \rho \mathcal{W} Y + X\beta + \varepsilon \tag{1}$$

- « Spatial Error Model » (SEM) : effet de diffusion.

$$\begin{cases} Y & = X\beta + \varepsilon \\ \varepsilon & = \lambda \mathcal{W} \varepsilon + u \end{cases} \tag{2}$$

Plan

1 Problématique

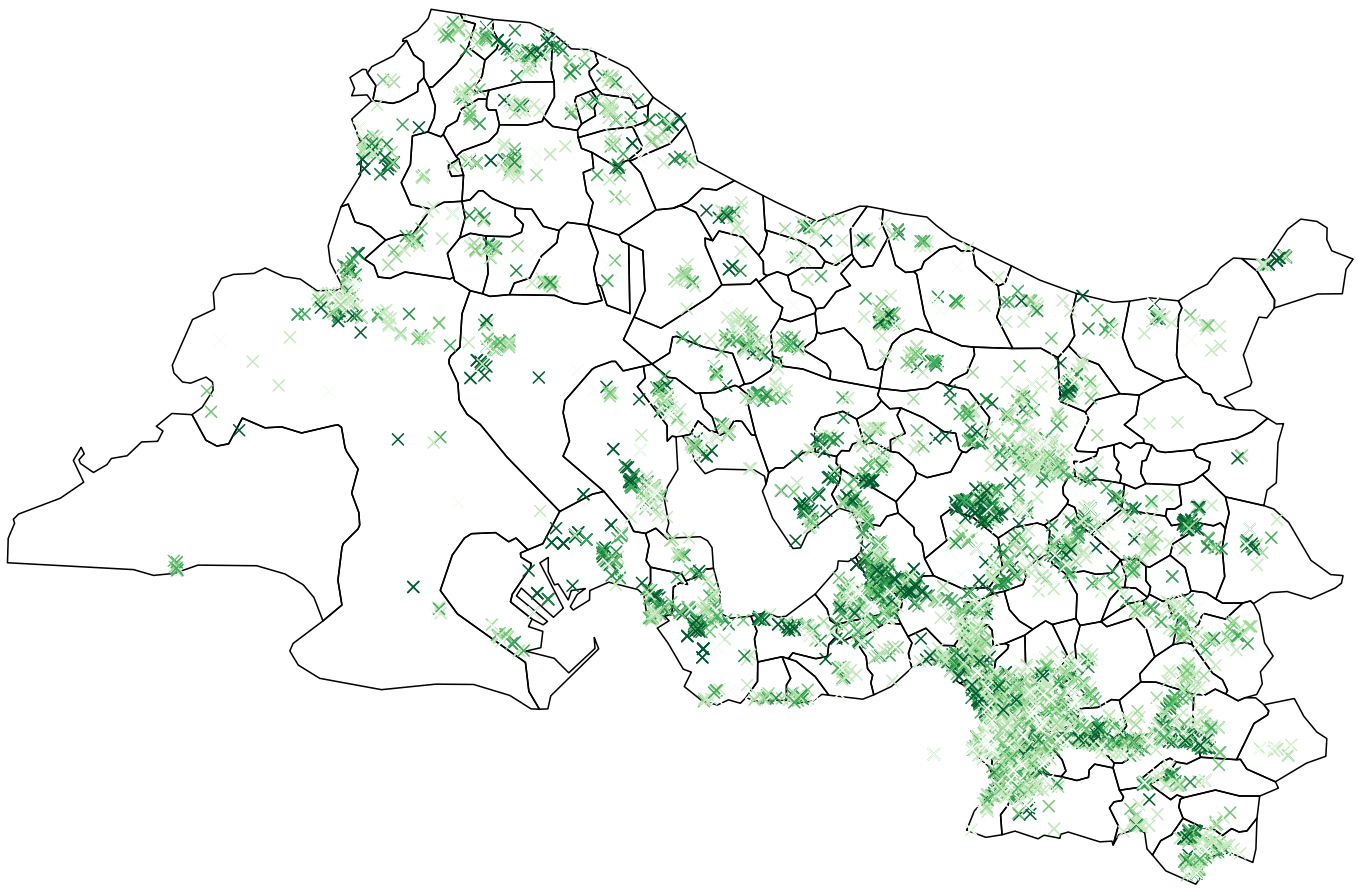
2 Modèles d'économétrie spatiale

3 Application : la production industrielle dans les Bouches-du-Rhône

4 Conclusions

Modèle

On s'intéresse à la production des 6 306 entreprises industrielles des Bouches-du-Rhône :



La production Y_i d'une entreprise i peut s'exprimer selon une fonction de type Cobb-Douglas :

$$Y_i = A L_i^{\beta_L} K_i^{\beta_K}$$

où :

- L_i est son effectif salarié ;
- K_i son capital ;
- $A = \exp(\beta_0) \prod_{j \in v_i} Y_j^{\rho \omega_{ij}}$ la productivité générale des facteurs, dépendant de la production des entreprises voisines Y_j ;
- β_L et β_K représentent les élasticités de la production au travail et au capital.

En passant au logarithme, on retrouve un modèle de type SAR :

$$\widetilde{Y} = \beta_0 + \rho \mathcal{W} \widetilde{Y} + \beta_L \widetilde{L} + \beta_K \widetilde{K} + \varepsilon$$

On estime ce modèle sur l'ensemble des entreprises :

$\hat{\beta}_0$	$\hat{\beta}_L$	$\hat{\beta}_K$	$\hat{\rho}$
0.42	0.54	0.77	0.05
(0.05)	(0.02)	(0.01)	(0.01)

On va réaliser deux familles de plans de sondages sur cette population :

1. Des sondages aléatoires simples, comme base de référence ;
2. Des sondages stratifiés sur l'effectif salarié, en utilisant une allocation de Neyman basée sur la dispersion du chiffre d'affaires.

On fera varier dans chacun des cas la taille d'échantillon, ici 500 ou 2 000 entreprises.

Résultats de l'estimation du modèle à partir d'un échantillon (par Monte Carlo) :

	Echantillon aléatoire (SAS)			Echantillon stratifié		
n	$\hat{\rho}$	$\hat{\beta}_L$	$\hat{\beta}_K$	$\hat{\rho}$	$\hat{\beta}_L$	$\hat{\beta}_K$
500	0.02	0.55^{***}	0.77^{***}	0.02^{**}	0.37^{***}	0.80^{***}
	(0.02)	(0.07)	(0.05)	(0.01)	(0.05)	(0.04)
2000	0.03^{***}	0.54^{***}	0.77^{***}	0.04^{***}	0.46^{***}	0.79^{***}
	(0.01)	(0.03)	(0.02)	(0.01)	(0.03)	(0.02)

Pour rappel, sur la population totale :

$\hat{\rho}$	$\hat{\beta}_L$	$\hat{\beta}_K$
0.05	0.54	0.77
(0.01)	(0.02)	(0.01)

Les conclusions de ces simulations sont les suivantes :

- Pour un échantillon de petite taille, on ne détecte aucun effet spatial significatif ;
- Le paramètre d'autocorrélation spatiale $\hat{\rho}$ est systématiquement sous-évalué ;
- En revanche, les coefficients $\hat{\beta}_L$ et $\hat{\beta}_K$ sont correctement estimés dans le cadre du sondage aléatoire simple.

Remarque : on n'utilise pas ici les poids de sondage dans l'estimation, car il n'existe pas de consensus sur la manière d'estimer un SAR pondéré dans la littérature.

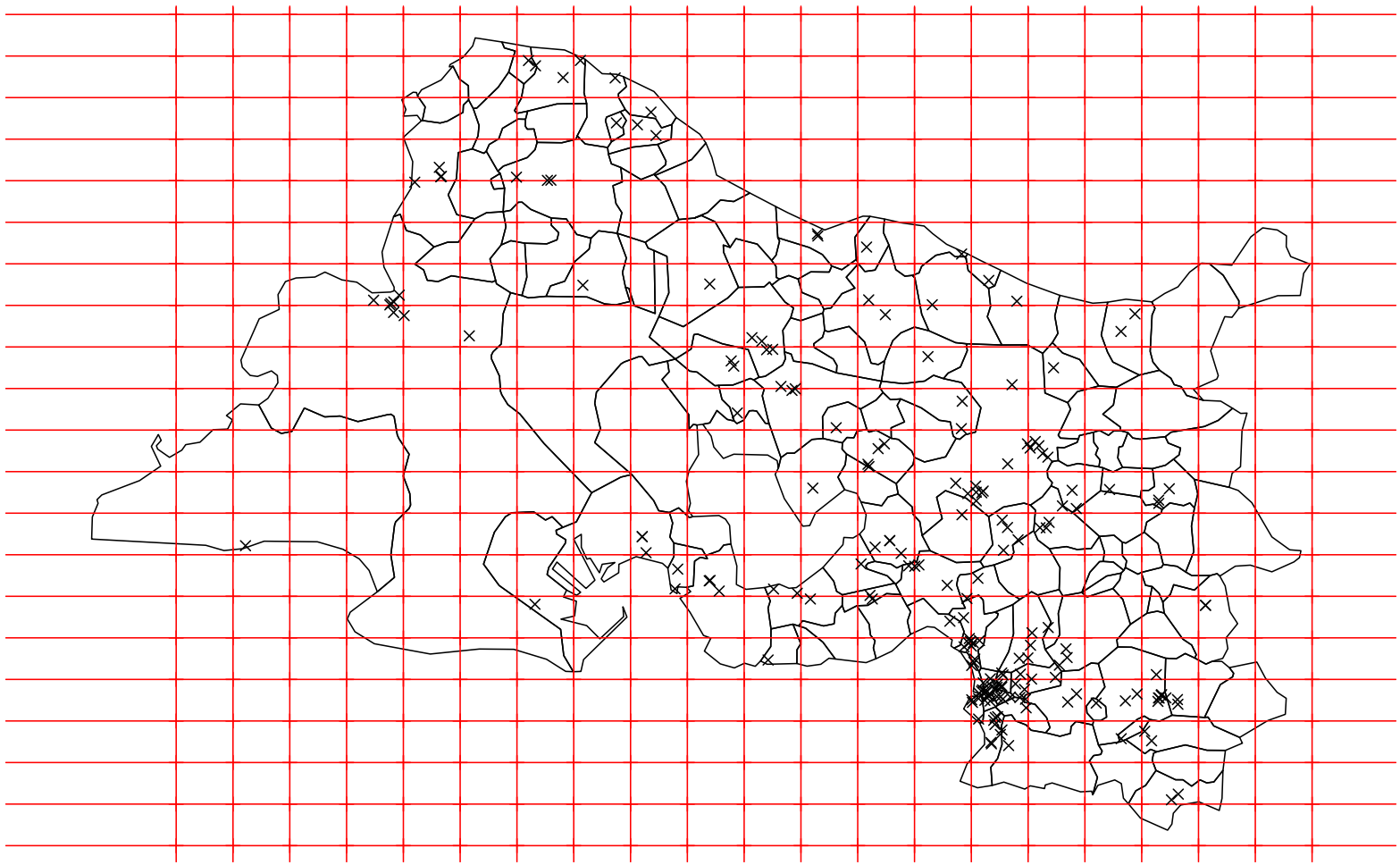
Méthodes de correction

Nous proposons deux types de méthodes de correction :

1. Agréger les données à un niveau supérieur, pour limiter les valeurs manquantes et recréer une structure géographique cohérente ;
2. Imputer les valeurs manquantes avec des valeurs plausibles afin d'estimer le modèle économétrique sur un jeu de données complet.

Méthodes de correction – Agrégation

On utilise une grille $G \times G$ pour agréger les données économiques :



Méthodes de correction – Agrégation

Les résultats obtenus ne permettent pas de détecter un effet spatial significatif :

n \ G	20	30	50	60
100	0.01	0.01	0.01	0.02
	(0.02)	(0.02)	(0.02)	(0.02)
200	0.01	0.01	0.02	0.02
	(0.02)	(0.02)	(0.02)	(0.02)
500	0.02	0.02	0.01	0.01
	(0.03)	(0.02)	(0.01)	(0.01)
1000	0.03	0.06*	0.02*	0.01
	(0.03)	(0.04)	(0.02)	(0.01)

Méthodes de correction – Imputation

Différentes méthodes d'imputation sont considérées :

- Par le ratio : on estime un modèle linéaire entre $\tilde{Y}$, $\tilde{L}$ et $\tilde{K}$, qu'on utilise pour obtenir des valeurs plausibles ;
- Par hot-deck statistique : on tire aléatoirement parmi les « voisins » statistiques de l'unité manquante, c'est à dire les entreprises ayant les mêmes caractéristiques ;
- Par hot-deck géographique : on tire aléatoirement parmi les « voisins » de l'unité manquante, c'est à dire les entreprises les plus proches.

Méthodes de correction – Imputation

- L'imputation par le ratio ne permet pas de détecter un effet spatial ;
- La méthode de hot-deck statistique semble fonctionner.

	Ratio			**Hot Deck Statistique**		
n	$\hat{\rho}$	$\hat{\beta}_L$	$\hat{\beta}_K$	$\hat{\rho}$	$\hat{\beta}_L$	$\hat{\beta}_K$
500	0.00	0.55^{***}	0.77^{***}	0.04^{***}	0.61^{***}	0.70^{***}
	(0.00)	(0.08)	(0.06)	(0.01)	(0.06)	(0.04)
2000	0.02^{***}	0.54^{***}	0.77^{***}	0.04^{***}	0.56^{***}	0.75^{***}
	(0.00)	(0.03)	(0.02)	(0.01)	(0.03)	(0.02)

Pour rappel, sur la population totale :

$\hat{\rho}$	$\hat{\beta}_L$	$\hat{\beta}_K$
0.05	0.54	0.77
(0.01)	(0.02)	(0.01)

Méthodes de correction – Imputation

- L'imputation par hot deck géographique conduit à des résultats aberrants.

	Hot Deck Géographique		
n	$\hat{\rho}$	$\hat{\beta}_L$	$\hat{\beta}_K$
500	0.41^{***}	0.06^{*}	0.15^{***}
	(0.04)	(0.03)	(0.02)
2000	0.33^{***}	0.20^{***}	0.34^{***}
	(0.02)	(0.03)	(0.02)

Pour rappel, sur la population totale :

$\hat{\rho}$	$\hat{\beta}_L$	$\hat{\beta}_K$
0.05	0.54	0.77
(0.01)	(0.02)	(0.01)

Plan

1 Problématique

2 Modèles d'économétrie spatiale

3 Application : la production industrielle dans les Bouches-du-Rhône

4 Conclusions

Conclusions

Lors de l'estimation d'un modèle d'économétrie spatiale de type SAR sur des données issues d'enquête, les deux effets identifiés (taille et répartition spatiale) conduisent à :

- Sous-estimer l'ampleur du facteur d'autocorrélation spatiale...
- voire conclure à sa non-significativité.

Hors cas particuliers (tels que l'Enquête Emploi en Continu, dont l'échantillon est par grappes), il convient donc d'être très précautionneux dans l'interprétation des résultats obtenus.

Conclusions

Les pistes de solution abordées ici ne sont pas idéales :

- L'agrégation spatiale peut conduire à des biais écologiques, c'est à dire à ne pas s'intéresser au niveau d'interaction économique pertinent ;
- L'imputation semble plus prometteuse, mais nous n'avons aucun résultat solide permettant de conclure sur son efficacité de manière générale.

Merci pour votre attention !