Sondages et économétrie spatiale

Raphaël Lardeux & Thomas Merly-Alpa

Groupe de lecture : « Sondages et Économétrie »

15 juin 2015

Plan

Introduction

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.

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

Introduction

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.

Introduction

Comment appréhender la dimension spatiale des données ?

1. Définir des unités spatiales.
2. Définir une notion de « voisinage » : contiguïté, distance géographique, n 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.

Économétrie spatiale sur données d'enquête

Deux questions principales :

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

→ Est-il possible d'intégrer la dimension spatiale dès la constitution de l'échantillon ?

Pour approfondir la réflexion :
La dimension spatiale de l'échantillon est-elle la même que celle des données ? Y a t-il une endogénéité de la dimension spatiale dans l'échantillonnage vis-à-vis du problème étudié ? En quoi les propriétés des estimateurs sont-elles modifiées ou non par le choix d'un plan de sondage spatialisé ?

Introduction

Notre démarche :

1. Choix d'un espace géographique (comtés américains) et d'une matrice de pondération spatiale (par la distance principalement).
2. Simulation d'un jeu de données à partir d'un processus générateur de données spatial.
3. Tirage d'échantillons selon divers plans de sondage.
4. Estimation du même modèle sur l'échantillon.

Plan

Modèles d'économétrie spatiale

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.

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

Choix de $\mathcal{W}$

Comme mentionné précédemment, le choix de $\mathcal{W}$ est très important.

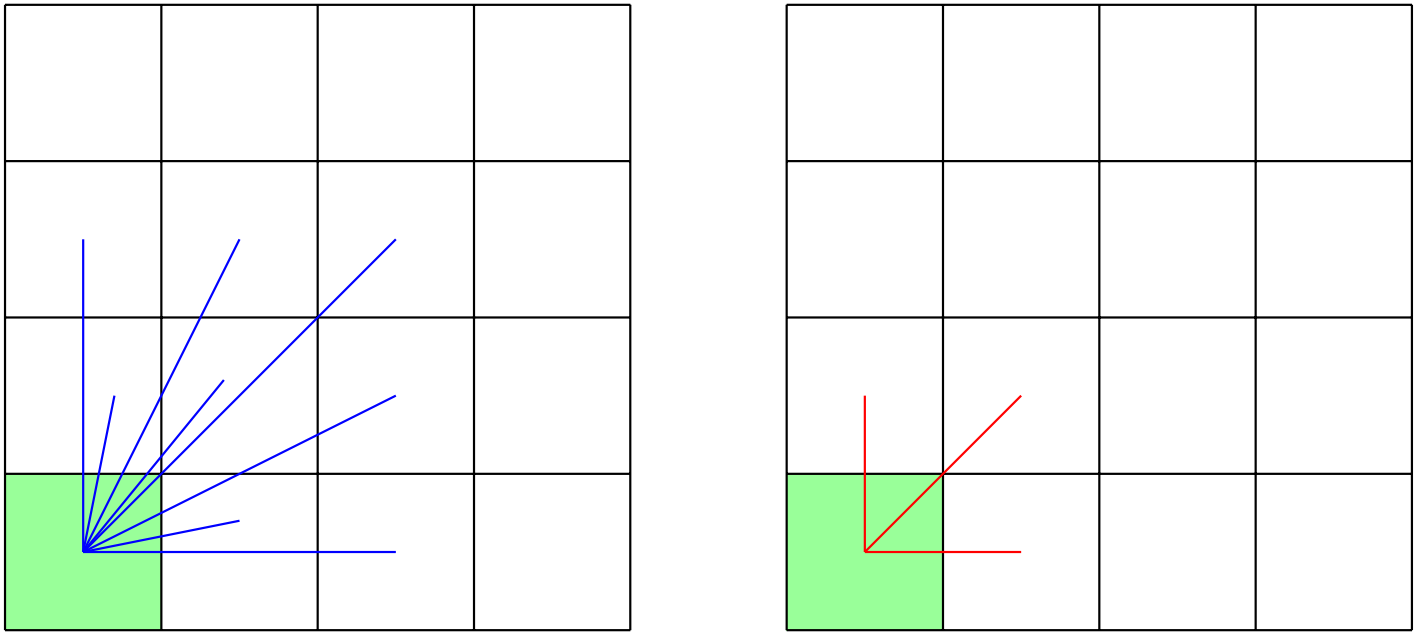


Figure: Gauche : matrice de distance bornée, Droite : 3 plus proches voisins

Choix de $\mathcal{W}$

En particulier 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.

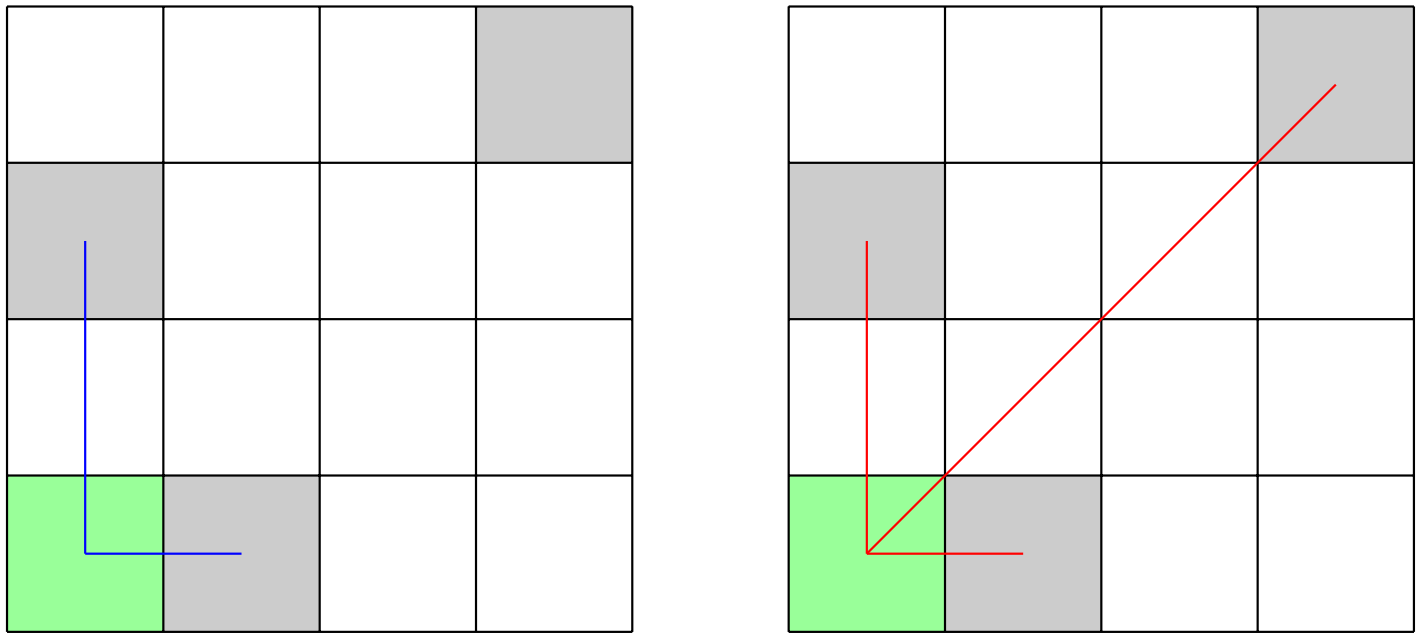
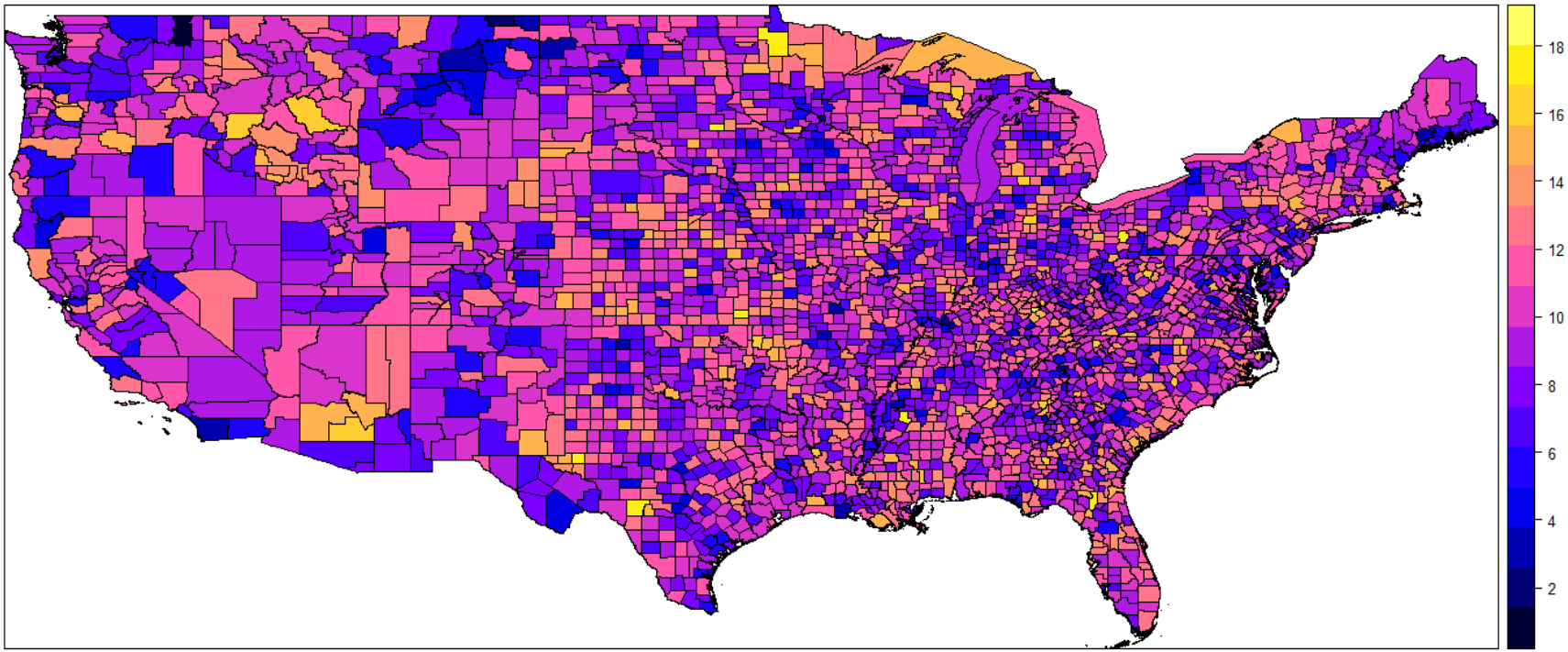


Figure: Gauche : matrice de distance bornée, Droite : 3 plus proches voisins

Un exemple

Simulation d'un SAR avec $\beta = 1$ et $\rho = 0.5$ sur les comtés des US :

Figure: Y simulés selon un modèle SAR

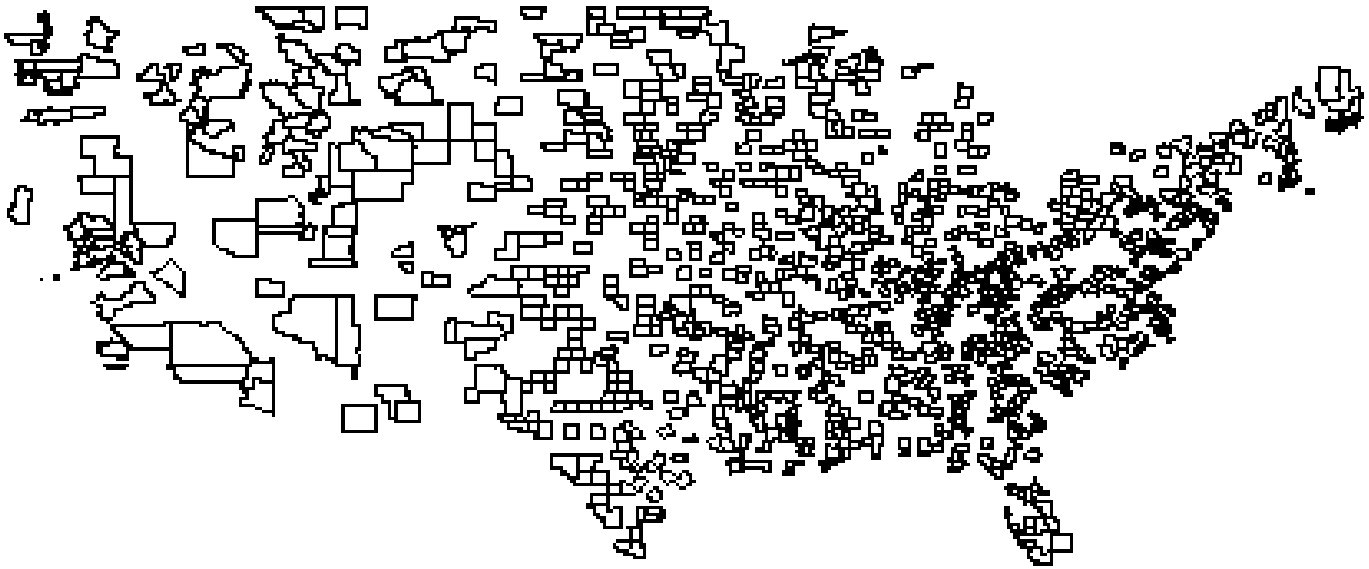


Plan

Sondage aléatoire simple

On s'intéresse pour l'instant au sondage aléatoire simple parmi les 3114 comtés. Empiriquement, un échantillon est de la forme suivante (ici pour $n = 1000$) :

Figure: Un échantillon pour $n = 1000$



Sondage aléatoire simple

On simule 100 modèles et 100 échantillons pour chacun des modèles, et on obtient les résultats suivants :

Table: Estimation des paramètres par Monte Carlo - Modèle SAR

	ρ	β
n = 100	0.016671	1.038009***
	(0.026453)	(0.069029)
n = 200	0.025659	1.037564***
	(0.018862)	(0.048234)
n = 300	0.035970**	1.039415***
	(0.017495)	(0.038486)
n = 500	0.059361***	1.035146***
	(0.017689)	(0.029143)

Sondage aléatoire simple

Quel effet du choix de la matrice de pondération spatiale ?

Table: Estimation des paramètres par Monte Carlo - Modèle SAR

	ρ			β		
	k=2	k=5	Distance	k=2	k=5	Distance
n = 100	0.023977 (0.074101)	0.018829 (0.111609)	0.016671 (0.026453)	1.116647*** (0.077982)	1.057082*** (0.064966)	1.038009*** (0.069029)
n = 200	0.059829 (0.052176)	0.069620 (0.075278)	0.025659 (0.018862)	1.113500*** (0.054116)	1.058051*** (0.045142)	1.037564*** (0.048234)
n = 300	0.086660** (0.042170)	0.110825* (0.060644)	0.035970** (0.017495)	1.115209*** (0.043950)	1.055183*** (0.036459)	1.039415*** (0.038486)
n = 500	0.132350*** (0.032475)	0.171898*** (0.044434)	0.059361*** (0.017689)	1.109467*** (0.033858)	1.051829*** (0.027841)	1.035146*** (0.029143)

Sondage aléatoire simple

Estimation des effets directs, indirects et totaux :

Table: Modèle SAR, matrice de pondération par la distance

	SAR			MCO
	Direct	Indirect	Total	$\hat{\beta}$
n = 100	1.036958***	0.006097	1.043055***	1.036729***
	(0.069028)	(0.009526)	(0.070342)	(0.068418)
n = 200	1.037350***	0.014418	1.051769***	1.036345***
	(0.047689)	(0.011092)	(0.049848)	(0.047666)
n = 300	1.037416***	0.025114**	1.06253***	1.039613***
	(0.038930)	(0.012686)	(0.041767)	(0.039344)
n = 500	1.039222***	0.053410***	1.092632***	1.041152***
	(0.030304)	(0.015721)	(0.035342)	(0.030960)

Sondage aléatoire simple

Quelques remarques sur les résultats précédents :

- Le paramètre β est correctement estimé. Il correspond globalement à l'effet direct du modèle.
- L'effet indirect est négligeable dans cette situation.
- On détecte de l'autocorrélation spatiale à partir de $n = 300$, soit un taux de sondage de l'ordre de $f = 1/10$. Cependant, les valeurs de ρ estimées restent très faibles.
- Le choix de la matrice de pondération spatiale affecte peu les résultats.
- (Cas particulier : ici, l'estimation par les MCO coïncide avec l'effet direct)

Sondage stratifié

Constitution de 2 strates (Est/Ouest) puis tirage par SAS au sein de chacune d'entre elles :

Table: Estimation d'un SAR pour le sondage stratifié

	ρ	β
n = 100	0.017161	1.04039***
	(0.026699)	(0.06847)
n = 200	0.026301	1.04037***
	(0.018834)	(0.048659)
n = 300	0.0348883**	1.03617***
	(0.017348)	(0.038440)
n = 500	0.060053***	1.03579***
	(0.017685)	(0.02941)

Sondage par grappes

Le sondage par grappes consiste à échantillonner d'autres comtés qui sont proches des comtés initialement tirés. On construit les grappes de comtés en suivant les règles suivantes :

- Toutes les grappes ont même taille : elles contiennent 18 comtés.
- On veut minimiser le diamètre des grappes.
- On ne fixe pas de règle de contiguité des grappes.

On obtient finalement 173 grappes.

Sondage par grappes

Table: Estimation d'un SAR pour le sondage par grappe

	ρ	β
p = 5	0.263558**	1.017854***
n = 90	(0.128856)	(0.061604)
p = 11	0.283748***	1.014822***
n = 198	(0.096696)	(0.041012)
p = 16	0.295382***	1.016142***
n = 288	(0.081984)	(0.034007)
p = 28	0.314812***	1.014269***
n = 504	(0.063799)	(0.025807)

Sondage par grappes

Table: Effets directs, indirects et totaux pour le modèle par grappes

	Direct	Indirect	Total
p = 5	1.030521***	0.3849629*	1.415484**
n = 90	(0.07739719)	(0.2298276)	(0.2770181)
p = 11	1.035111***	0.4175277**	1.452639***
n = 198	(0.04465074)	(0.1631477)	(0.1772829)
p = 16	1.039306***	0.4162175**	1.455523***
n = 288	(0.03346941)	(0.1536485)	(0.1657557)
p = 28	1.030511***	0.4402968**	1.470807***
n = 504	(0.02717901)	(0.1202532)	(0.1288771)

Sondage par grappes

Des résultats frappants...

- L'autocorrélation spatiale est captée pour des taux de sondage de l'ordre de $f = 3/100$ (c'était $1/10$ pour le SAS).
- Le paramètre spatial estimé est proche de sa vraie valeur.
- L'effet indirect est significatif et de grande ampleur.

...mais de potentiels problèmes :

- Tendance à détecter facilement de l'autocorrélation spatiale.
- Biais potentiels en présence d'hétérogénéité spatiale.

Plan

Conclusions

Les résultats de nos simulations semblent indiquer que :

- Dans le cadre d'un sondage aléatoire simple, ou plus largement non défini spatialement, pour un taux de sondage classique il est inutile d'employer plus que les moindres carrés ordinaires.
- Pour des sondages en grappe, les méthodes d'estimation utilisées en économétrie spatiale sont pertinentes même pour des taux de sondage assez faibles.

Ces résultats ne sont évidemment qu'indicatifs, car ils sont basés uniquement sur des simulations, et dans un cadre favorable ($\mathrm{Var}(\epsilon)$ faible).

Conclusions

L'application de ces résultats aux enquêtes de l'INSEE dépend de leur champ. Pour ce qui concerne les enquêtes ménages,

- l'EEC étant réalisée par grappes, on peut utiliser des méthodes d'économétrie spatiale.
- il faut sinon agréger les données sur des zones géographiques : département, IRIS...

Pour les enquêtes entreprises,

- les taux de sondage pouvant être plus importants, on pourrait avoir des résultats significatifs.
- les plans de sondage stratifiés selon l'effectif $\times$ secteur d'activité, qui sont des variables potentiellement spatialement hétérogènes également, seraient à étudier.