11. Économétrie spatiale sur données d'enquête

RAPHAËL LARDEUX, THOMAS MERLY-ALPA
Insee

Résumé

L'économétrie spatiale requiert des données exhaustives sur un territoire, ce qui interdit en principe l'utilisation de données d'enquête. Le présent chapitre présente les écueils liés à l'estimation d'un modèle spatial autorégressif (SAR) sur données échantillonnées et évalue les potentielles corrections proposées par la littérature empirique. Nous identifions deux sources de biais : *(i)* un "effet taille" résultant de la distorsion de la matrice de pondération spatiale et *(ii)* un effet résultant de l'omission d'unités spatialement corrélées avec les unités observées. Tous deux tendent à sous-estimer la corrélation spatiale. Le biais est cependant plus faible dans le cas d'un sondage par grappes et lorsque l'échantillon est suffisamment grand. Deux catégories de méthodes sont proposées par la littérature empirique afin de passer outre ces écueils : l'imputation des valeurs manquantes (régression linéaire, hot deck) et l'agrégation des données à une échelle supérieure. La difficulté est de reconstituer une information complexe à partir de peu d'observations, même si l'imputation par hot deck statistique semble constituer une piste prometteuse. La dernière partie de ce chapitre illustre cette problématique dans le cas concret de l'estimation d'externalités de production entre les industries du département français des Bouches-du-Rhône.

R La lecture préalable des chapitres 2 : "Codifier la structure de voisinage", 3 : "Indices d'autocorrélation spatiale" et 6 : "économétrie spatiale : modèles courants" est nécessaire pour comprendre ce chapitre.

Introduction

Les développement récents de l'économétrie spatiale et de la géolocalisation permettent l'analyse de phénomènes spatiaux à des échelles très locales (firmes, logements, ...), renforçant ainsi la précision des estimations. Les concepts issus de ce champ sont utilisés dans des domaines de plus en plus diversifiés : géostatistique, économie, analyse de réseaux. Cependant, l'application de ces méthodes d'analyse spatiale requiert des données exhaustives, qui ne sont pas toujours accessibles (non-réponse, temps de collecte trop important, ...) et ne peuvent pas aisément être traitées en un temps restreint. L'extension de l'économétrie spatiale aux données d'enquête permettrait de tirer pleinement parti d'une information détaillée pour mesurer finement l'incidence des corrélations spatiales sur les estimations économétriques [1]. Dans ce chapitre, nous discutons ainsi les développements récents relatifs à l'application des méthodes d'estimation spatiales lorsqu'une partie des observations est manquante, en particulier dans le cas de données d'enquête. Nous ne traitons ni la possibilité d'un sondage spatialisé, qui est complexe dans le cas des données sociales [2], ni les cas d'observations dont la localisation est inconnue. Pourquoi l'économétrie spatiale requiert-elle des données exhaustives ? L'économétrie classique repose sur une hypothèse d'indépendance mutuelle des observations. Estimer un modèle sur un sous-ensemble de données peut affecter la puissance des tests statistiques mais, en l'absence de problème de sélection, les estimateurs restent sans biais et efficaces. Au contraire, dans les modèles d'économétrie spatiale, les observations sont considérées comme corrélées entre elles : chaque unité est influencée par ses voisins. Supprimer des observations revient à omettre leurs liens avec les unités observées proches, ce qui introduit un biais dans l'estimation du paramètre de corrélation spatiale et des effets spatiaux estimés. Nous constatons que ce biais tend à atténuer la valeur du paramètre de corrélation spatiale, puisque certains liens de voisinage ne sont alors plus pris en compte dans l'estimation.

Conceptuellement, l'économétrie spatiale se distingue de l'économétrie classique par la façon dont elle considère les observations. En économétrie classique, les observations s'apparentent à un échantillon aléatoire représentatif d'une population et sont interchangeables. L'analyse spatiale les conçoit comme l'unique réalisation d'un processus spatial, chaque observation étant alors nécessaire à l'estimation du processus sous-jacent [3]. L'économétrie spatiale a été développée dans le cadre très pur des modèles de CLIFF et al. 1972, caractérisé par une information exhaustive et parfaite sur les unités spatiales et par l'absence de données manquantes (ARBIA et al. 2016). En pratique, ces conditions ne sont quasiment jamais réunies et appliquer directement des techniques d'estimation spatiale peut fortement altérer les résultats.

L'application de méthodes spatiales à des données non exhaustives pose plusieurs problèmes. Premièrement, les estimations sont perturbées par un "effet taille". L'existence de m données manquantes parmi une population de taille n donne lieu à une matrice de pondération de taille $(n-m)\times(n-m)$ au lieu de la vraie matrice de taille $n\times n$, ce qui biaise le paramètre de corrélation spatiale du simple fait du changement de dimension (ARBIA et al. 2016). Nous illustrons par la suite ce phénomène à partir d'un échantillon localement exhaustif de données simulées par un modèle SAR, en montrant qu'appliquer le même modèle à cet échantillon ne permet pas de retrouver la valeur du paramètre de corrélation spatiale. Deuxièmement, l'existence de données manquantes

1. PINKSE et al. 2010 qualifient ces perspectives de "futur de l'économétrie spatiale".
2. Sur ces questions, on pourra se reporter au chapitre 10 : "échantillonnage spatial".
3. L'analyse spatiale se rapproche en cela des séries temporelles, où le jeu de données observé est issu d'un processus stochastique.

engendre une erreur de mesure sur l'effet du voisinage (régresseur $\mathbf{W}y$) qui biaise les paramètres estimés. Par simulation, nous montrons qu'au-delà de l' "effet taille", ce biais a des conséquences importantes.

Différentes corrections ont été proposées, sans qu'aucune ne s'impose radicalement [4]. Lorsque la localisation des individus est connue, les solutions par imputation sont généralement privilégiées (RUBIN 1976 ; LITTLE 1988 ; LITTLE et al. 2002). Cependant, une imputation naïve, par exemple par un modèle linéaire, ne permet pas de corriger les biais (BELOTTI et al. 2017a). Pour contourner ce problème, KELEJIAN et al. 2010b développent des estimateurs lorsque seul un sous-ensemble incomplet d'une population est disponible. WANG et al. 2013a mettent en place une méthode d'imputation par moindres carrés en deux étapes dans un cadre où des valeurs de la variable dépendante sont aléatoirement manquantes. Dans ce même contexte, LESAGE et al. 2004 recourent à l'algorithme EM (DEMPSTER et al. 1977) : une phase "E" (espérance) assigne une valeur aux données manquantes, conditionnellement aux observables et aux paramètres du modèle spatial sous-jacent, puis une phase "M" (maximisation) détermine la valeur de ces paramètres par maximisation de la vraisemblance du modèle. Par itération, cette procédure permet de tirer d'un modèle estimé l'ensemble de l'information disponible pour imputer des valeurs manquantes. Les travaux plus récents de BOEHMKE et al. 2015 étendent cette procédure au cas d'observations manquantes (variables dépendante et indépendantes inconnues).

Des travaux empiriques récents illustrent l'importance de ces corrections. Dans un modèle de prix hédoniques, LESAGE et al. 2004 appliquent l'algorithme EM pour prédire la valeur des logements non vendus. Dans un modèle de réseaux avec autocorrélation spatiale, LIU et al. 2017 montrent que la détection d'un effet de pair requiert de prendre en compte le processus d'échantillonnage. Les méthodes complexes d'imputation selon un modèle estimé (*model-based*) sont cependant encore peu appliquées. Lorsque certaines données sont manquantes, la solution généralement retenue est de supprimer du champ de l'analyse les observations correspondantes, au risque d'engendrer un biais d'atténuation de la corrélation spatiale. Certains travaux se restreignent à un sous-ensemble, notamment une région ou un groupe particulier (REVELLI et al. 2007), ce qui peut poser un problème d' "effet taille" et amener à sous-estimer les corrélations à la bordure de l'espace considéré (KELEJIAN et al. 2010b). Enfin, la plupart des applications sont réalisées sur données agrégées pour bénéficier de données exhaustives sur une échelle plus large, mais cette solution peut provoquer des erreurs positionnelles [5] (ARBIA et al. 2016) ainsi qu'un biais écologique (ANSELIN 2002b). Nous discutons par la suite l'incidence de ces diverses méthodes sur les estimations spatiales.

Le problème des valeurs manquantes dans un cadre d'observations non indépendantes a été mis en avant par des champs proches de l'économétrie spatiale : séries temporelles et géostatistique d'une part, économétrie des réseaux d'autre part. Les séries temporelles et la géostatistique se rapprochent du traitement des données spatiales continues. Le problème des données manquantes a été abordé très tôt dans le domaine des séries temporelles (CHOW et al. 1976, FERREIRO 1987). JONES 1980 ; HARVEY et al. 1984 recommandent l'utilisation d'un filtre de Kalman pour simultanément estimer un modèle et imputer des valeurs. L'analyse géostatistique corrige des

4. En particulier, ces méthodes varient selon les hypothèses sous-jacentes portant sur les données manquantes : selon que la valeur et/ou la localisation des observations est manquante, que les variables dépendantes et/ou indépendantes sont affectées et selon que la probabilité pour une donnée d'être manquante dépend des corrélations avec les données observables et/ou inobservables. La littérature sur l'incidence des données manquantes établit ainsi une distinction entre *Missing at Random* (MAR), *Missing Completely at Random* (MCAR) et *Missing Not at random* (MNAR). cf RUBIN 1976, HUISMAN 2014

5. ARBIA et al. 2016 proposent ce concept pour désigner les cas où la position d'une observation (X,Y) n'est pas connue précisément. Par exemple, manque de précision dans la mesure, mesure brouillée pour des questions de confidentialité, adresses manquantes.

jeux de données incomplets soit en amont par des méthodes d'échantillonnage spatialisé, soit en prédisant la valeur d'une variable spatiale continue en une position inconnue (interpolation spatiale ou krigeage, voir chapitre 5 :"Géostatistique"). Des approches spatio-temporelles croisant krigeage et filtre de Kalman ont également été développées (MARDIA et al. 1998). Cependant, ces méthodes propres aux données continues ne peuvent être transposées à l'analyse économique et sociale, où les données sont fondamentalement discrètes. De plus, le recours à ces techniques de sondage spatialisé irait à l'encontre des principes fondamentaux de la collecte de données sociales tels que l'équipondération et l'utilisation de bases de sondages déterministes. L'économétrie des réseaux a très vite souligné les biais engendrés par des observations manquantes (BURT 1987 ; STORK et al. 1992 ; KOSSINETS 2006), mais les solutions pratiques restent rares, même si les enjeux liés à l'estimation de l'autocorrélation spatiale sur un échantillon d'un réseau prennent de l'ampleur avec l'utilisation croissante des réseaux sociaux (ZHOU et al. 2017). De même qu'en économétrie spatiale, la principale difficulté est de reconstituer l'information sur les données inobservées à partir des données observées, sans connaître l'effet des premières sur les secondes (KOSKINEN et al. 2010). En particulier, HUISMAN 2014 ne tranche pas entre diverses stratégies d'imputation classiques et montre que celles-ci ne fonctionnent que dans des cas spécifiques. Des solutions fondées sur des méthodes d'échantillonnage ont également été proposées afin de collecter des données sur les populations d'intérêt (GILE et al. 2010).

Le présent chapitre se concentre sur deux questions : quels sont les biais engendrés par l'application de méthodes spatialisées à des données d'enquête ? quelles sont les conséquences des diverses solutions classiques (suppression des données, imputation, agrégation) ? Ces questions sont abordées par ARBIA et al. 2016, qui procèdent par simulation et observent une incidence plus marquée des données manquantes lorsqu'elles sont regroupées en grappes, auquel cas l'intégralité de phénomènes locaux peut être perdue. Ils considèrent cependant des cas où les données manquantes représentent au maximum 25 % de la population, ce qui est très faible par rapport aux données d'enquête, où elles atteignent généralement plus de 90 % de la population.

La section 11.1 présente les biais issus de l'application de méthodes spatiales à un échantillon non exhaustif de données, selon la part des observations échantillonnées et le type de sondage. La section 11.2 discute les conséquences de quelques solutions usuelles : le passage à l'échelle supérieure par agrégation et l'imputation des valeurs manquantes. La section 11.3 illustre ces biais à partir de l'estimation d'une équation de production avec externalités sur les industries du département français des Bouches-du-Rhône.

11.1 Première approche par simulations

Dans cette première partie, nous mettons en évidence, par des simulations de type Monte-Carlo, l'existence d'un biais dans l'estimation d'un modèle autorégressif spatial (SAR) sur des données d'échantillon. D'abord, nous simulons sur un espace géographique des données spatialement corrélées en fixant la valeur du paramètre de corrélation spatiale. Nous procédons par la suite à des tirages d'échantillon, à partir desquels nous estimons la valeur du paramètre de corrélation spatiale.

11.1.1 Simulation d'un SAR

L'espace géographique retenu est une carte de l'Europe [6], détaillée au niveau administratif NUTS3 (échelon le plus bas dans la hiérarchie NUTS définie par Eurostat, qui correspond à des petites zones sur lesquelles peuvent être menées des études spécifiques : les départements français, par exemple) de laquelle nous retirons les îles les plus éloignées ainsi que l'Islande afin de conserver

6. Cette carte est diffusée sur le site : `http://ec.europa.eu/eurostat/fr/web/gisco/geodata/reference-data/administrative-units-statistical-units`.

un espace géographique homogène et compact. À partir du *shapefile* de l'Europe, nous construisons une matrice de voisinage $\mathbf{W}$ fondée sur la distance, de telle sorte que le poids associé à deux unités voisines décroît selon le carré de la distance et s'annule lorsque cette distance dépasse un seuil limite. Les résidus et une variable explicative sont simulés dans des lois normales : $\varepsilon \sim \mathcal{N}(0,1)$ et $X \sim \mathcal{N}(5,2)$. Cela nous permet de finalement simuler une variable Y suivant un modèle de type SAR (*Spatial Auto-Regressive*) :

$$Y = (1-\rho\mathbf{W})^{-1} X\beta + (1-\rho\mathbf{W})^{-1}\varepsilon \tag{11.1}$$

avec $\beta = 1$ et $\rho = 0.5$, paramètres de référence que nous cherchons à retrouver par l'estimation de modèles SAR sur des échantillons. Les données des variables simulées Y sont représentées sur la figure 11.1. La présence de zones colorées concentrées est caractéristique de l'autocorrélation spatiale positive résultant du processus générateur des données.

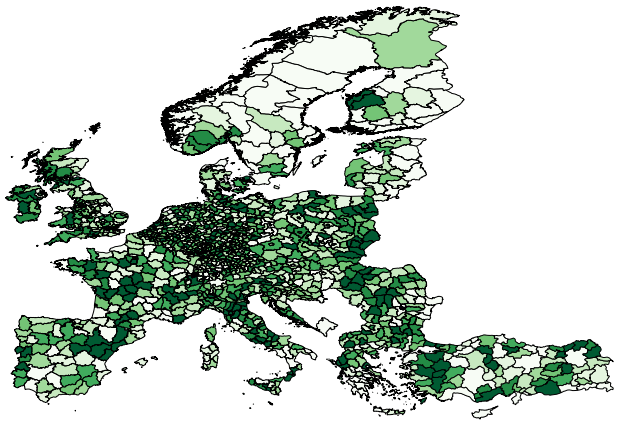

FIGURE 11.1 – Y simulé selon un modèle SAR
Copyright : EuroGeographics pour les limites administratives

La table 11.1 présente les résultats de l'estimation d'un modèle SAR sur l'ensemble des zones NUTS3 d'Europe. Ils confirment la validité de cette simulation, puisque les paramètres β et ρ estimés sont très proches des valeurs calibrées initialement.

β	ρ	Direct	Indirect	Total
0.989	0.494	1.043	0.860	1.902

TABLE 11.1 – Paramètres estimés par SAR sur l'ensemble des zones

Encadré 11.1.1 — Simulation d'un SAR avec R. Pour simuler un SAR en R, l'étape la plus importante est de formater sa matrice de voisinage $\mathbf{W}$ de la façon suivante :

```
D <- nb2listw(W, style="W", zero.policy=TRUE)
```

Une fois la matrice de voisinage au format `listw`, il faut alors inverser $1 - \rho\mathbf{W}$ en utilisant la fonction suivante, dont ρ est l'un des paramètres :

```
InvD <- invIrW(D,rho)
```

Attention, cette étape peut être chronophage. Il ne reste alors plus qu'à simuler notre variable Y :

```
Y <- (InvD %*% X) + (InvD %*% eps)
```

11.1.2 Procédures de sondage

L'enjeu est d'examiner la capacité des modèles spatiaux à correctement estimer ρ et β à partir d'échantillons tirés dans ces données simulées. Nous discutons en particulier l'effet que peut avoir l'échantillonnage de certaines de ces zones sur l'estimation du modèle sous-jacent.

Un sondage consiste à sélectionner de façon aléatoire en suivant une procédure dite plan d'échantillonnage un ensemble de n unités au sein d'une population de N, où n est souvent bien plus petit que N afin de limiter les coûts liés à la collecte d'information. La théorie des sondages affirme que les estimations réalisées à l'aide de l'échantillon s'étendent sans biais à la population totale, mais que celles-ci sont plus précises quand la taille de l'échantillon augmente et quand le plan d'échantillonnage est adapté à la variable estimée. Pour approfondir les questions de sondage, la lecture de ARDILLY 1994, TILLÉ 2001 ou COCHRAN 2007 est conseillée.

Dans la suite de cette partie, nous présentons quelques techniques de sondage classiques et leur application dans le cadre des NUTS3 européens. Nous pouvons cependant déjà faire quelques hypothèses et remarques générales, en suivant les idées développées dans GOULARD et al. 2013 concernant le nouveau recensement de la population. D'une part, l'effet ne devrait évidemment pas être le même selon la taille n de l'échantillon retenu. Avec une petite dizaine de zones, la structure spatiale initiale ne pourra pas être reconstituée, tandis qu'échantillonner 95 % voire 99 % des zones devrait permettre de la retrouver facilement. D'autre part, la question de la méthode d'échantillonnage va également se poser : la dimension spatiale est-elle prise en compte dans le cadre de la méthode ? On pourra se reporter au chapitre 10 : "échantillonnage spatial" pour approfondir ces questions.

Sondage aléatoire simple

Le sondage aléatoire simple consiste à tirer indépendamment et sans remise n boules au sein d'une urne de taille N. Les individus ont alors tous la même chance d'être sélectionnés dans l'échantillon. La sélection d'un individu dans un échantillon diminue la probabilité qu'ont les autres d'être également inclus. Dans notre cas, on sélectionne n zones complètement au hasard. La figure 11.2 présente un exemple d'échantillon.

Sondage poissonnien

Le sondage poissonnien (ou bernoullien) consiste à tirer à pile ou face pour chaque individu de la population son appartenance à l'échantillon. Dans ce cas, les individus ont toujours la même chance d'être sélectionnés dans l'échantillon. La sélection d'un individu dans un échantillon n'influe pas sur la probabilité qu'ont les autres d'être également inclus, mais la taille de l'échantillon n'est pas fixée *a priori*. Dans notre cas, chaque zone a une probabilité p d'être retenue dans l'échantillon : la taille de l'échantillon obtenu est alors pN en espérance.

Sondage par grappes

Le sondage par grappes (ou aréolaire) consiste à sélectionner ensemble des groupes d'individus. Les individus ont toujours la même chance d'être sélectionnés dans l'échantillon. Cependant, la sélection d'un individu au sein d'un échantillon influe fortement sur la probabilité qu'ont les autres d'être également inclus, car les individus d'une même grappe sont toujours sélectionnés ensemble.

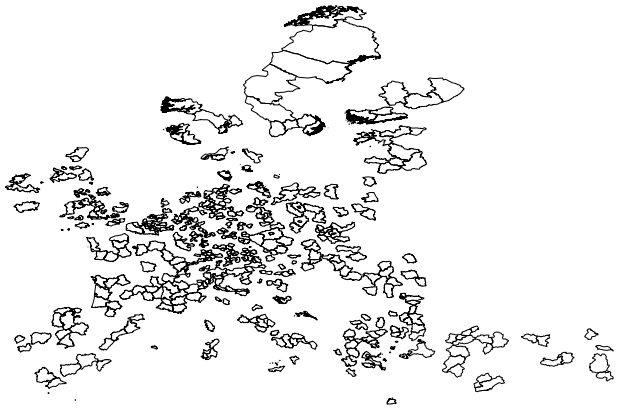

FIGURE 11.2 – Un échantillon obtenu par sondage aléatoire simple ($n = 500$)
Copyright : EuroGeographics pour les limites administratives

Ici, il s'agit de rassembler les zones NUTS3 en différentes grappes et ensuite de réaliser une sélection aléatoire de certaines de ces grappes. L'intérêt principal est de limiter les coûts de collecte, au prix d'une perte en précision liée à l'homogénéité intra-grappe.

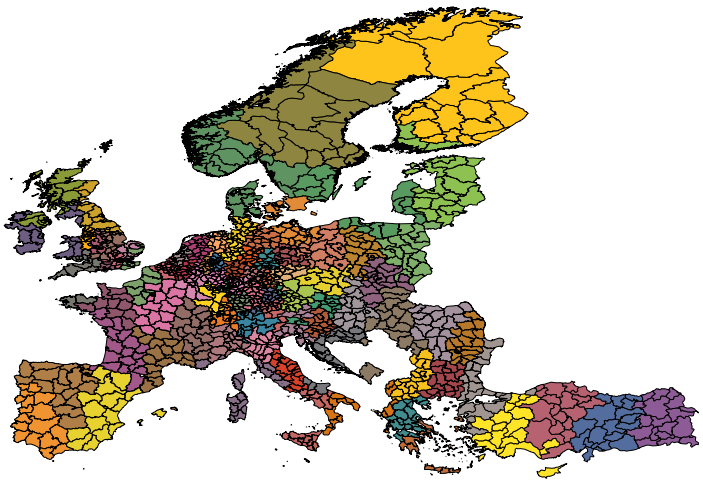

FIGURE 11.3 – Division de l'Europe en grappes
Copyright : EuroGeographics pour les limites administratives

Il serait envisageable d'utiliser les différents niveaux NUTS1 ou NUTS2 comme grappes. Cependant, ils sont de taille importante et ne comportent pas tous le même nombre de NUTS3. Or des grappes de taille trop importante vont limiter le nombre de simulations possibles. Au contraire, des grappes comportant des nombres de zones différents introduisent soit une problématique

de poids de sondage différents entre les individus, que nous ne souhaitons pas traiter ici (voir DAVEZIES et al. 2009 pour un débat sur l'usage des poids de sondage en économétrie), soit une problématique de taille d'échantillon variable ce qui peut avoir des effets complexes à analyser. Nous séparons donc les zones en grappes de même taille tout en maintenant une certaine cohérence géographique. Comme la matrice de pondération est basée sur la distance géographique, nous privilégions les grappes les moins étendues possible.

Afin d'obtenir des grappes de taille identique, il est nécessaire que le nombre de grappes soit un diviseur du nombre de zones NUTS3. En vue de limiter la taille des grappes, nous rassemblons les 1445 zones NUTS3 en 85 grappes de 17 zones chacune. Pour cela, nous utilisons un algorithme de construction des grappes : partant de la zone la plus éloignée du centre de la carte, nous agrégeons les zones les plus proches de celle-ci jusqu'à en obtenir 17. Comme les grappes sont construites une à une, les NUTS3 les plus éloignés seront déjà affectés pour la construction des grappes précédentes, et l'algorithme se poursuit avec des zones plus centrales. Les grappes obtenues sont représentées sur la figure 11.3.

Sondage stratifié

Le sondage stratifié correspond à un tirage de n boules, mais en tirant n_1 boules dans une première urne, n_2 dans une deuxième, jusqu'à n_H dans une H-ième, où $n = n_1 + n_2 + \ldots + n_H$. Pour réaliser un tirage stratifié, il convient de bien définir les H strates d'une part, et de bien choisir l'allocation $(n_1, \ldots, n_H)$ d'autre part. Une allocation classique est l'allocation de Neyman, qui a pour propriété de minimiser la variance de l'estimateur du total d'une variable d'intérêt (voir par exemple TILLÉ 2001). La formule est la suivante :

$$n_h = n \frac{N_h S_h}{\sum_{i=1}^{H} N_i S_i} \tag{11.2}$$

avec n la taille de l'échantillon total, N_h la taille de la strate h et S_h la dispersion de la variable d'intérêt au sein de la strate h. Dans certains cas, lorsque les comportements vis-à-vis de la variable d'intérêt sont hétérogènes, cette formule peut conduire à enquêter exhaustivement certaines strates, c'est-à-dire à leur appliquer un taux de sondage de 100 %.

11.1.3 Résultats et interprétation

Afin d'estimer l'effet de l'échantillonnage d'une partie des NUTS3 européens, nous suivons la méthode de Monte-Carlo. Nous réalisons ainsi 100 simulations de Y selon un modèle SAR puis nous tirons 100 échantillons pour chacune d'entre elles. Le modèle SAR est estimé sur chaque échantillon et l'on récupère les paramètres d'intérêt. Enfin, les paramètres présentés en résultat sont les moyennes des ρ et β sur les 10 000 échantillons et leurs écart-types sont calculés sur ces 10 000 valeurs.

Pour chacun des 10 000 tirages, sont ainsi conservées les valeurs de X et de Y des zones échantillonées. On reconstruit alors une matrice de pondération spatiale $\mathbf{W}_{\text{echantillon}}$ fondée sur la distance, tel que précédemment, mais limitée aux unités présentes dans l'échantillon. Différentes tailles d'échantillon et différentes méthodes d'échantillonnage sont considérées.

Sondage aléatoire simple

La table 11.2 présente les estimations obtenues dans le cas d'un sondage aléatoire simple pour des tailles d'échantillon n variant de 50 à 250 zones. Ces estimations mettent en évidence une autocorrélation spatiale significative à partir d'un échantillon de taille $n = 150$, ce qui correspond approximativement, dans notre cas, à un taux de sondage de $1/10$. Le paramètre β est estimé sans biais quelle que soit la taille de l'échantillon, mais le paramètre estimé $\hat{\rho}$ est nettement inférieur

à sa vraie valeur $\rho = 0.5$. Par conséquent, pour des échantillons de petite taille, l'effet indirect n'est pas significativement différent de zéro et reste bien inférieur à celui observé sur la population entière. L'autocorrélation spatiale est largement sous-estimée.

n	$\hat{\rho}$	$\hat{\beta}$	Direct	Indirect	Total
50	0.043	1.055***	1.056***	0.016	1.072***
	(0.043)	(0.125)	(0.125)	(0.017)	(0.128)
100	0.058*	1.050***	1.052***	0.032*	1.083***
	(0.031)	(0.087)	(0.087)	(0.019)	(0.091)
150	0.072**	1.049***	1.051***	0.048**	1.099***
	(0.028)	(0.068)	(0.068)	(0.020)	(0.073)
250	0.101***	1.051***	1.054***	0.080***	1.135***
	(0.026)	(0.051)	(0.052)	(0.023)	(0.060)

TABLE 11.2 – Estimation d'un modèle SAR sur des échantillons tirés par sondage aléatoire simple
Note : *** désigne une significativité à 1 %, ** une significativité à 5 % et * une significativité à 10 %. Les écart-types sont entre parenthèses. n : nombre d'observation dans l'échantillon. Ces estimations proviennent de 10 000 simulations.

Sondage par grappes

L'échantillonnage par grappes permet de conserver une structure géographique forte localement, ce qui dans notre cas semble bénéfique pour la détection d'effets spatiaux, en particulier pour des petites valeurs de n. À partir des grappes présentées dans la partie 11.1.2, nous réalisons des tirages de nombres différents de grappes, allant de 3 à 15 grappes, soit 51 à 255 zones. La table 11.3 montre les résultats obtenus pour des valeurs de $n = 17p$, la taille de l'échantillon composé de p grappes.

n	p	$\hat{\rho}$	$\hat{\beta}$	Direct	Indirect	Total
51	3	0.309*	1.015***	1.051***	0.441*	1.492***
		(0.237)	(0.091)	(0.097)	(0.262)	(0.310)
102	6	0.348***	1.017***	1.054***	0.493***	1.546***
		(0.100)	(0.063)	(0.066)	(0.188)	(0.215)
153	9	0.363***	1.017***	1.054***	0.516***	1.571***
		(0.078)	(0.052)	(0.055)	(0.152)	(0.176)
255	15	0.377***	1.014***	1.052***	0.541***	1.593***
		(0.058)	(0.038)	(0.040)	(0.119)	(0.136)

TABLE 11.3 – Estimation d'un modèle SAR sur des échantillons tirés par grappes
Note : *** désigne une significativité à 1 %, ** une significativité à 5 % et * une significativité à 10 %. Les écart-types sont entre parenthèses. n : nombre d'observations dans l'échantillon. p : nombre de grappes dans l'échantillon. Ces estimations proviennent de 10 000 simulations.

Avec un sondage par grappes, le paramètre $\hat{\rho}$ est plus proche de sa vraie valeur et l'inclut dans son intervalle de confiance. La précision de l'estimation s'améliore nettement lorsque n augmente, mais l'estimateur reste biaisé. Ainsi, contrairement au cas du sondage aléatoire simple, il est possible de capter les interactions spatiales même avec un taux de sondage très faible de l'ordre de 3 %. En effet, les unités enquêtées sont fortement concentrées dans l'espace et donc très représentatives de la corrélation spatiale. En revanche, si le nombre d'unités tirées est faible, il en va de même pour la précision de l'estimation de la corrélation spatiale. Dès lors, l'effet indirect

est bien détecté même pour des échantillons de petite taille et sa valeur est plus proche de celle obtenue sur la population totale. L'estimation d'effets géographiques semble ainsi raisonnable dans le cadre d'un tel type de sondage.

Deux questions subsistent : tout d'abord, est-ce que cet échantillonnage par grappes n'aurait pas tendance à favoriser la détection d'un modèle autocorrélé spatialement, même si la tendance n'est pas majeure sur la totalité de la population ? Comme l'on dispose de peu de valeurs de X et Y, le terme $\mathbf{W}Y$ est paradoxalement assez bien connu, ce qui pourrait amener à favoriser cette piste. D'autre part, et cela sera developpé dans la partie 11.1.4, on peut s'étonner de l'écart observé entre le $\hat{\rho}$ estimé et la vraie valeur utilisée pour la génération du SAR, alors même qu'on détecte bien les effets spatiaux.

11.1.4 Un "effet taille"

Les résultats obtenus par simulation peuvent étonner les économètres. En effet, dans le cadre d'un sondage aléatoire simple, assimilable au modèle de superpopulation utilisé en économétrie [7], l'estimation d'un paramètre d'une population ou d'un modèle est usuellement sans biais, tant que le plan d'échantillonnage est correctement spécifié. Il apparaît alors que le paramètre d'autocorrélation spatiale ρ ne suit pas cette "loi" classique de la théorie des sondages [8].

Nonobstant la question de la sélection aléatoire des zones sur lesquelles on récupère l'information relative aux Y, se restreindre à un nombre de zones inférieur à celui de la population entière induit une modification de la structure spatiale sous-jacente.

La question des biais écologiques, c'est-à-dire des erreurs d'estimation de modèles économétriques spatiaux qui proviennent de la mauvaise spécification spatiale, que cela soit au niveau du grain (de la résolution) des données ou des problèmes frontaliers, est proche de ce sujet. Ainsi, il est tout à fait possible que lorsque que l'on se restreint à n zones, avec $n < N$, on ne puisse jamais obtenir un effet spatial aussi fort que sur la population entière.

Pour illustrer ce point, nous estimons plusieurs centaines de modèles SAR, générés sur la population entière, à partir d'une restriction de la population aux n NUTS3 les plus centraux : l'idée étant de se limiter à une sous-partie de l'Europe sans qu'elle ne soit choisie aléatoirement ni de façon morcelée comme c'était le cas pour les échantillons obtenus précédemment (figure 11.2). La figure 11.4 compare les valeurs de $\hat{\rho}$ obtenues en suivant trois protocoles pour différents pourcentages P % de la population totale : la sélection des P % NUTS3 les plus centraux, un sondage par grappes où chaque grappe de zones a P % de chances d'être sélectionnée, et un sondage poissonnien classique où chaque zone a indépendamment P % de chances d'être sélectionnée.

De même que dans la partie 11.1.3, le sondage poissonnien (proche du sondage aléatoire simple) donne des valeurs estimées de $\hat{\rho}$ bien plus faibles que le sondage par grappes. L'apport principal de cette figure est dans la courbe rouge, qui repose sur une sélection non aléatoire d'une partie des zones. Elle converge plus rapidement que les autres vers 0.5, la vraie valeur de ρ. Ce constat semble confirmer l'hypothèse d'un biais lié à la déformation de la structure spatiale ou "effet taille", résultant d'une restriction à un sous-ensemble de la population totale.

7. Ce terme est lié à la différence entre approche *sous le plan* et *sous le modèle* en sondages. Si on raisonne sous le plan, on suppose que la population a des valeurs de Y déterministes - approche usuelle. Sous le modèle, on suppose qu'il y a un modèle dit de *superpopulation* dont dérivent les Y de la population. Ici on doit suivre cette approche pour pouvoir estimer nos modèles SAR

8. On notera que plusieurs paramètres ne respectent pas cette loi : on peut par exemple penser au maximum d'une variable Y sur une population, qui n'est pas possible à estimer sans biais à partir d'un échantillon. Par ailleurs, dans notre cas de sondage aléatoire simple ou par grappes, il n'y a pas de problèmes de sous-couverture, c'est-à-dire d'unités de la population qui ne peuvent pas appartenir à l'échantillon pour des raisons souvent liées à la qualité des registres. Cette piste ne peut pas expliquer le biais sur $\hat{\rho}$.

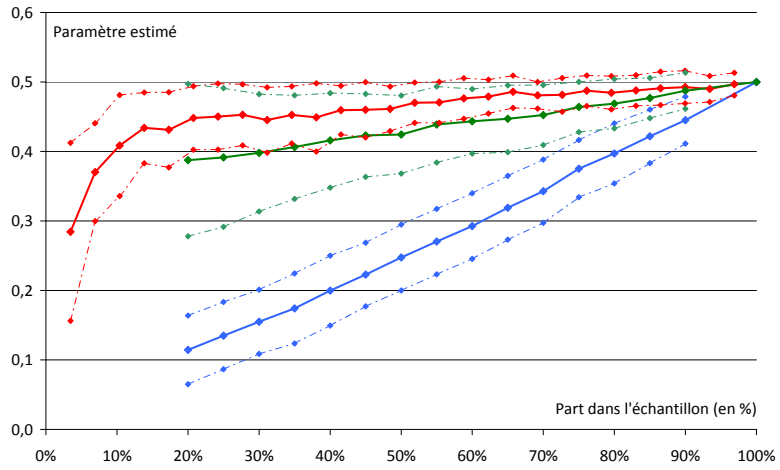

FIGURE 11.4 – L' "effet taille"
Note : Chaque point d'une courbe en trait plein représente une estimation du paramètre $\hat{\rho}$ pour une taille d'échantillon exprimée en pourcentage de la population exhaustive. Lorsque l'estimation est réalisée sur données exhaustives, on retrouve $\hat{\rho} = 0.5$. La courbe bleue correspond à un sondage aléatoire simple, la courbe verte à un tirage par grappes. La courbe rouge représente un sélection déterministe des régions, en partant d'un point initial puis en s'en éloignant progressivement. Le courbes en pointillés représentent les intervalles de confiance à 95 %

11.1.5 Robustesse des résultats

En conclusion de cette section, notons que la spécification retenue pour le modèle spatial n'affecte les résultats obtenus que de façon marginale. Ces derniers restent inchangés lorsque le seuil maximum de distance varie ou lorsque la notion de distance retenue est fondée sur les plus proches voisins (table 11.10 en annexe 11.3.6). Enfin la vraie valeur du paramètre ρ n'affecte pas les résultats des estimations. La figure 11.5 montre que, à taux de sondage donné, une estimation sur un échantillon tiré par sondage aléatoire simple ne permet presque jamais de retrouver la vraie valeur du paramètre ρ. Dans le cas d'un sondage par grappe, cette valeur peut être incluse dans l'intervalle de confiance du paramètre estimé, mais le biais lié à l'estimation ne disparaît pas lorsque son ampleur ou son signe changent. Dans tous les cas, le biais atténue l'ampleur de la corrélation spatiale estimée.

Enfin, considérer un modèle de type SEM (*Spatial Error Model*) : $Y_2 = X\beta + (1 - \lambda \mathbf{W})^{-1}\varepsilon$ n'affecte pas radicalement les résultats (table 11.11 en annexe 11.3.6).

11.2 Pistes de résolution

Une première position face à un problème de données manquantes est d'ignorer, consciemment ou non, ces données et d'appliquer directement le modèle spatial aux unités observées. Ce choix atténue le paramètre de corrélation spatiale relativement à sa vraie valeur, de par un "effet taille" et un effet d'échantillonnage.

Le premier effet provient de différences entre le modèle théorique et le modèle estimé concernant la dimension de la matrice de pondération spatiale. Pour le supprimer, il faut comparer données exhaustives et données échantillonnées selon une même structure géographique, et donc sur un même nombre d'unités. Pour compenser le second, il faut être en mesure de reconstituer les corrélations spatiales entre unités observées et manquantes. Dans le cas présent, la localisation des unités

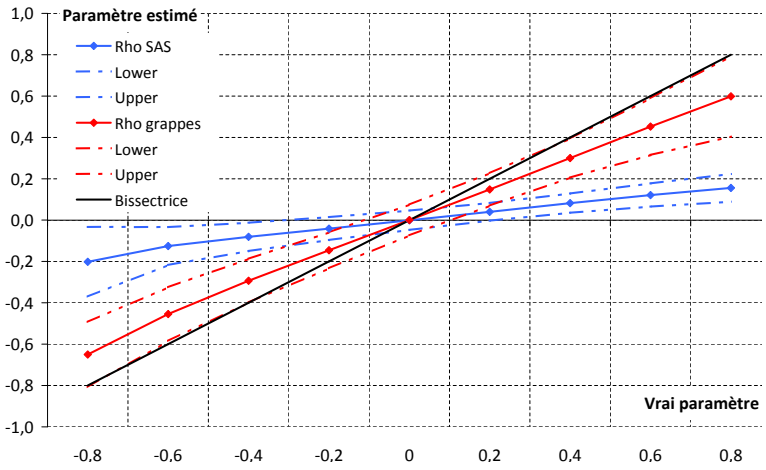

FIGURE 11.5 – Estimations de $\hat{\rho}$ pour diverses valeurs de ρ
Note : Les courbes en trait plein représentent la valeur estimée $\hat{\rho}$ en fonction de la valeur ρ fixée pour la simulation des données. La courbe bleue correspond au cas de données tirées par sondage aléatoire simple et la courbe rouge au cas d'un sondage par grappes. Les courbes en pointillés représentent les intervalles de confiance à 95 % de l'estimateur $\hat{\rho}$

est toujours supposée connue [9].

Dans cette partie, nous discutons l'incidence de deux solutions généralement appliquées dans les travaux empiriques : le passage à une échelle supérieure par agrégation des données et l'imputation des données manquantes. Ces méthodes conservent la structure géographique mais sont plus ou moins efficaces pour reconstituer les corrélations spatiales.

11.2.1 Passer à l'échelle supérieure par agrégation

Problématique

En l'absence de données individuelles exhaustives, de nombreux travaux sont réalisés à une échelle agrégée de régions, de départements, de zones d'emploi. Ce choix dépend de façon cruciale de l'échelle d'analyse de la problématique, requiert de disposer d'un bon estimateur de la moyenne locale et peut mener à des biais écologiques (voir ANSELIN 2002b pour plus de précisions). Les corrélations intra-zone sont alors omises, au profit des corrélations entre zones.

Pour évaluer cette solution, nous simulons 6 000 points selon une loi uniforme sur un espace carré et leur affectons, comme dans la partie 11.1, des valeurs de X et de Y correspondant à un SAR de paramètre d'autocorrélation spatiale $\rho = 0.5$. Ces points sont représentés sur la surface de gauche de la figure 11.6. Puis cet espace carré est découpé selon une grille de taille $G \times G$ pour différentes valeurs de G, et à chaque centroïde de chaque case est affectée la moyenne des points situés à l'intérieur de cette case. Les panneaux du centre et de droite de la figure 11.6 représentent cette configuration pour $G = 50$ et $G = 20$ respectivement.

L'estimation d'un modèle SAR sur données exhaustives agrégées avec $G = 50$ permet d'estimer un paramètre de corrélation spatiale $\hat{\rho} = 0,47$ d'écart-type $\hat{\sigma}_\rho = 0,068$. Ce paramètre est significativement positif et l'estimation inclut $0,5$ dans son intervalle de confiance. L'agrégation de données sur des parcelles limiterait la perte d'interactions spatiales et minimiserait le biais dans l'estimation du paramètre de corrélation spatiale. La figure 11.7 montre que les paramètres ρ et β sont précisément estimés et sans biais dès lors que la grille sur laquelle les données sont agrégées

9. Le manque d'information concernant la localisation de certaines unités est un autre enjeu des recherches actuelles en économétrie spatiale (ARBIA et al. 2016) qui dépasse cependant le cadre du présent chapitre.

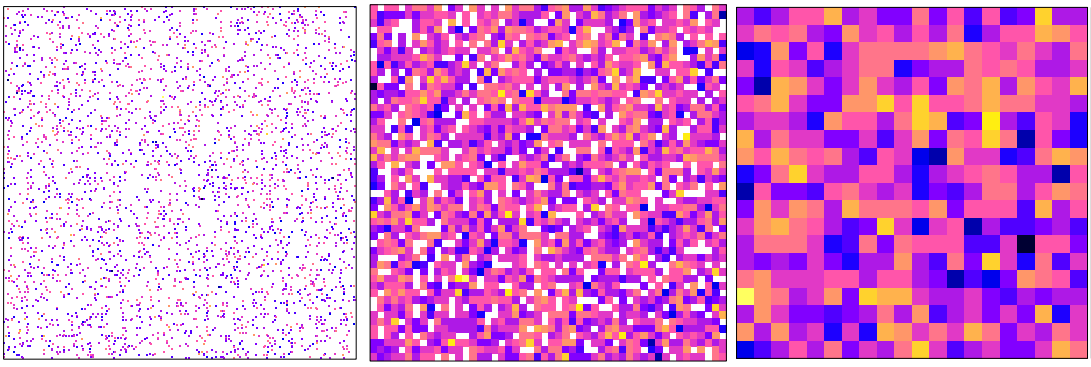

FIGURE 11.6 – Agrégation de données spatiales
Note : *à gauche :* 6 000 points simulés selon une loi uniforme. *au centre :* données agrégées selon 50×50 parcelles. *à droite :* données agrégées selon 20×20 parcelles

est relativement fine. Ainsi, pour des données exhaustives, plus le quadrillage est fin, plus on se rapproche de la structure spatiale des données ponctuelles et, par conséquent, plus la corrélation spatiale estimée est proche de sa vraie valeur.

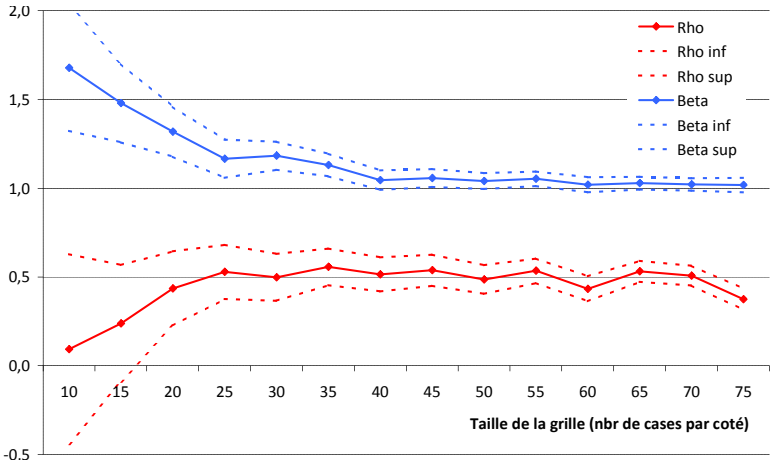

FIGURE 11.7 – Effet de la finesse de la grille sur les paramètres estimés
Note : Les résultats sont obtenus à partir de la simulation de 6 000 points selon une loi uniforme

Application à un échantillon

Cette procédure est répliquée sur des données échantillonnées par sondage aléatoire simple. La finesse de la grille répond à un arbitrage biais-variance : des maillons fins sont plus fidèles aux distances entre observations mais mènent à des estimations de moyennes locales moins précises pour chaque variable. Sous réserve d'assigner des poids nuls ainsi que des valeurs nulles des variables expliquée et explicatives aux mailles sans observation, il est possible de retrouver l'effet spatial simulé.

La table 11.4 présente les résultats de cette procédure pour différentes tailles d'échantillon et diverses grilles spatiales. Dans la majorité des cas, la vraie valeur de ρ se situe bien dans l'intervalle de confiance du paramètre estimé. Pour un petit échantillon, une grille trop grossière écrase les effets spatiaux tandis qu'une grille trop fine fournit une mauvaise estimation des variables individuelles. Comme précédemment, plus l'échantillon est grand, plus l'estimation est précise.

Ces simulations tendent à valider statistiquement l'approche par agrégation, sous réserve que

	$\hat{\rho}$				$\hat{\beta}$			
n \ G	10	30	50	60	10	30	50	60
100	0.487***	0.494***	0.483***	0.478***	1.016***	1.007***	1.027***	1.030***
	(0.070)	(0.060)	(0.068)	(0.072)	(0.134)	(0.115)	(0.129)	(0.134)
200	0.482***	0.499***	0.495***	0.489***	1.020***	0.998***	1.006***	1.011***
	(0.064)	(0.045)	(0.046)	(0.050)	(0.126)	(0.084)	(0.088)	(0.095)
500	-0.093	0.488***	0.483***	0.489***	1.035***	1.022***	1.031***	1.021***
	(0.701)	(0.032)	(0.030)	(0.032)	(0.121)	(0.060)	(0.055)	(0.059)
1000	-0.982	0.487***	0.485***	0.491***	1.048***	1.024***	1.028***	1.019***
	(0.159)	(0.024)	(0.020)	(0.021)	(0.119)	(0.045)	(0.038)	(0.040)

TABLE 11.4 – Estimation d'un SAR sur échantillons agrégés par parcelle
Note : Chaque ligne correspond à une taille d'échantillon n tiré parmi les 6 000 points simulés et chaque colonne correspond à la finesse de la grille en termes de nombre de carreaux (une grille de taille 30 découpe le carré initial en 900 cases) *** désigne une significativité à 1 %.

l'interprétation ne soit pas effectuée directement à l'échelle individuelle, mais reposent cependant sur des hypothèses fortes (coordonnées des unités déterminées par une loi uniforme, processus SAR homogène), rarement vérifiées en pratique.

11.2.2 Imputer les données manquantes

Pour rester à l'échelle des données disponibles, la solution est d'imputer des valeurs aux observations manquantes. C'est là encore une manière de faire abstraction de l' "effet taille" en assurant une cohérence entre la structure spatiale des données d'enquête et des données administratives. Face à des valeurs manquantes dans le cadre d'une enquête ou d'un recensement, attribuer des valeurs "plausibles" à ces unités permet de disposer d'un échantillon voire d'une population complète.

Méthodes d'imputation

Cette partie recense quelques méthodes classiques d'imputation. Le lecteur intéressé pourra se reporter à un bon livre de théorie des sondages, par exemple ARDILLY 1994 ou TILLÉ 2001, pour plus d'informations, de contexte théorique ainsi que pour d'autres méthodes plus avancées. Dans la cas d'une imputation par le ratio ou par hot deck, les variables explicatives X sont supposées connues de façon exhaustive.

Imputation par la moyenne. La méthode d'imputation par la moyenne (ou par la médiane, ou par la classe dominante dans le cas de variables qualitatives) est une méthode usuelle qui consiste à remplacer toutes les valeurs manquantes par la moyenne des valeurs observées. Cette méthode ne respecte pas une éventuelle structure économétrique entre différentes variables de l'enquête et peut conduire à des résultats faux dans le cadre d'estimations de tels modèles.

Imputation par le ratio. La méthode d'imputation par le ratio consiste à mobiliser l'information auxiliaire X disponible sur la totalité de la population, y compris les unités pour lesquelles l'information d'intérêt Y est manquante, afin d'imputer des valeurs de Y plausibles. Pour cela, on postule l'existence d'un modèle linéaire de la forme $Y = \beta X + \varepsilon$. $\hat{\beta}$ est estimé par les moindres carrés ordinaires, puis la valeur $Y_{\text{ratio}} = \hat{\beta} X$ est imputée pour les Y manquants. Le ratio des Y sur les X, dans le cas de données quantitatives, est le même entre les unités observées et les unités pour lesquelles on ne dispose pas d'information. Cette méthode peut être affinée en rajoutant des contraintes sur les unités pour lesquelles on calcule l'estimation du β, par exemple sur un domaine ou sur une strate précise.

Imputation par hot deck. La méthode de hot deck associe un donneur à une valeur manquante

de façon aléatoire, par opposition au cold deck qui établit ce lien de manière déterministe. Un donneur est ici un individu statistiquement "proche" de l'individu manquant (il partage des valeurs proches des X auxiliaires, appartient à la même strate, au même domaine, ou encore se situe à la même position spatiale). La mise en pratique d'un hot deck repose sur la définition d'un critère de distance, à partir duquel sont déterminés k voisins de l'individu dépourvu de valeur Y. Un individu parmi ces k voisins est choisi au hasard, uniformément ou non, pour donner sa valeur pour le nouveau Y_{hotdeck}. Il est possible d'introduire des variantes, par exemple en limitant le nombre de fois où un même individu peut être donneur, ou en réalisant le hot deck de façon séquentielle.

Pour aller plus loin

Les méthodes d'imputation peuvent plus ou moins directement altérer les estimations effectuées sur les données imputées. Le lien entre Y et X sur lequel repose l'imputation peut se retrouver exacerbé dans l'estimation du modèle sur Y et X (voir CHARREAUX et al. 2016 pour une discussion de ce point). De façon similaire, voire même amplifiée, l'utilisation de méthodes d'économétrie spatiale sur de telles données requiert une extrême précaution. En effet, la méthode d'imputation peut faire émerger une structure spatiale *ex-nihilo* ou au contraire briser les corrélations spatiales qu'elle ne prend pas en compte. Des exemples d'application de ces méthodes sont présentés en partie 11.3.6.

Enfin, tel que mentionné en introduction, des méthodes plus raffinées d'imputation au moyen de l'algorithme EM ont été développées (LESAGE et al. 2004 ; WANG et al. 2013a). Elles sont cependant complexes, très spécifiques au type d'information manquante et restent encore peu appliquées.

11.3 Application empirique : la production industrielle dans les Bouches-du-Rhône

Afin d'illustrer les enjeux relatifs à l'estimation de modèles spatiaux sur données d'enquête, nous procédons dans cette section à l'estimation d'une fonction de production sur des établissements issus du répertoire SIRUS. L'approche spatiale permet de mesurer l'incidence des interactions entre les processus de production des diverses entreprises sur la production de chacune d'entre elles. De tels *spillovers* entre entreprises ont déjà été mis en évidence par une importante littérature sur les économies d'agglomération, voir notamment LESAGE et al. 2007, ERTUR et al. 2007, LÓPEZ-BAZO et al. 2004, EGGER et al. 2006.

11.3.1 Données

Le répertoire SIRUS (Système d'Identification au Répertoire des Unités Statistiques) est le répertoire référent en termes de champ de la statistique d'entreprises française. Il est composé des entreprises, des groupes et de leurs établissements, contenus dans SIRENE (Système Informatisé du RÉpertoire National des Entreprises et des établissements), le répertoire administratif qui permet l'enregistrement des unités légales. Sont enregistrés pour chaque entreprise son chiffre d'affaires, son activité principale (disponible *via* le code APE, suivant la nomenclature française), son total de bilan, ses exportations, son effectif (tant administratif qu'en équivalent temps plein), son adresse physique ainsi que la liste des établissements qui la composent.

Les informations géographiques disponibles sur les établissements ont permis, grâce à un travail réalisé par la Division des Méthodes et Référentiels Géographiques de l'Insee, de géolocaliser aux coordonnées (x,y) chacun d'entre eux. Pour cela, différentes données ont été utilisées : du plus précis au moins précis, la référence cadastrale, la voie puis le centre de la commune pour les cas pour lesquels on dispose de trop peu d'information. Ces données géographiques, associées aux données économiques disponibles dans le répertoire SIRUS, permettent la modélisation de relations économétriques en prenant en compte la structure spatiale.

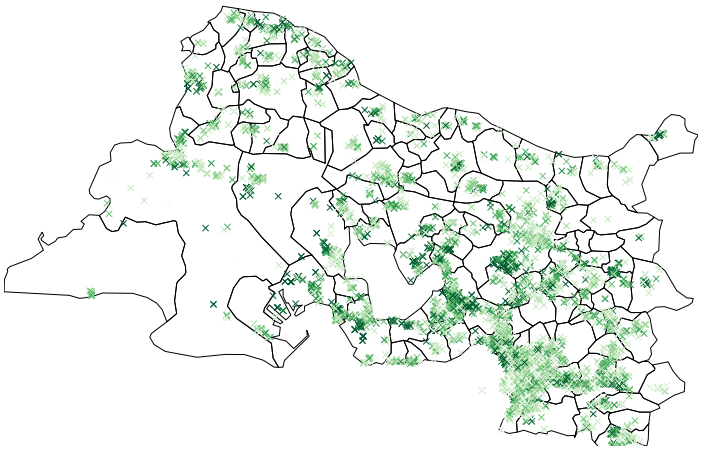

FIGURE 11.8 – Établissements industriels dans les Bouches-du-Rhône
Champ : établissements du secteur de l'industrie, dans le département des Bouches-du-Rhône
Source : *répertoire SIRUS, 2013*
Note : La couleur est plus foncée lorsque le chiffre d'affaires est important

11.3.2 Modèle

Une entreprise peut être influencée dans son processus de production par la proximité géographique qu'elle entretient avec des entreprises voisines. Ces interactions sont regroupées sous le concept d' "externalités" qui peuvent être positives lorsque le voisinage a un impact favorable sur la production (complémentarités entre secteurs, intégration des chaînes de production, relation avec des fournisseurs, transport, partage de connaissance,..) ou négatives lorsqu'elles nuisent à la production (concurrence, pollution, embouteillages, etc.).

La production Y_i d'une entreprise i peut s'exprimer selon une loi de type Cobb-Douglas : $Y_i = AL_i^{\beta_L} K_i^{\beta_K}$, en fonction de son effectif moyen L_i, de son capital K_i et de la productivité générale des facteurs A. Les paramètres β_L et β_K représentent respectivement la part des revenus du travail et du capital dans la production [10]. Traditionnellement, le terme A désigne l'ensemble des mécanismes qui influencent la production (capital humain, progrès technologique, complémentarités...) sans pouvoir être directement mesurés. Il peut aussi être conçu comme représentant les externalités positives liées à la production et s'écrire : $A = \exp(\beta_0) \prod_{j \in v_i} Y_j^{\rho \omega_{ij}}$, où v_i désigne le voisinage de l'établissement i, Y_j le niveau de production d'une unité voisine de i. Le terme $\rho \omega_{ij}$ désigne l'élasticité de la production de l'entreprise i par rapport à celle de l'entreprise j : lorsqu'une entreprise j voisine de i augmente sa production de 1 %, la production de l'entreprise i augmente de $\rho \omega_{ij}$ %. Le paramètre ρ capte les complémentarités communes à toutes les unités tandis que ω_j capte les complémentarités spécifiques, résultant de l'impact de l'activité de j sur la production de i. En composant par la fonction logarithme, l'équation estimée peut se réécrire :

$$\log(Y_i) = \beta_0 + \rho \sum_{j \in v_i} \omega_{ij} \log(Y_j) + \beta_L \log(L_i) + \beta_K \log(K_i) + \varepsilon_i \tag{11.3}$$

On voit ainsi apparaître la forme d'une équation caractéristique d'un modèle spatial autorégressif (SAR), où la variable expliquée de l'observation i est régressée sur la somme pondérée des valeurs de cette variable chez les observations voisines de i. ρ peut alors être interprété comme le paramètre de corrélation spatiale. ω_{ij} représente la force de l'interaction entre les unités i et j : c'est le poids associé à ces unités dans le matrice de pondération spatiale.

10. Ces paramètres peuvent également être interprétés respectivement comme les élasticités de la production au travail et au capital.

11.3.3 Estimation

L'équation 11.3 est estimée sur 6 306 établissements géolocalisés dans les Bouches-du-Rhône, appartenant au secteur de l'industrie [11]. Ce secteur est particulièrement approprié à une estimation spatiale, car il ne fait pas directement intervenir la localisation géographique dans la production (contrairement au commerce, aux transports ou à l'agriculture), n'est pas trop concentré (comme les hautes technologies) et ne fait pas particulièrement intervenir des logiques de réseau autres que spatiales (comme en finance ou dans les communications).

La production Y_i d'un établissement est donnée par le chiffre d'affaires. Le total du bilan de l'entreprise, qui est une mesure de son patrimoine, sert de proxy pour le capital de l'établissement K_i. Ces deux variables, uniquement disponibles à l'échelle de l'entreprise, sont divisées par le nombre d'établissements au sein de l'entreprise. Enfin, l'effectif L_i est disponible au niveau de l'établissement dans SIRUS.

La figure 11.8 représente par des croix la localisation de ces établissements. L'intensité de la couleur verte de ces croix matérialise la taille de leur chiffre d'affaires : plus la couleur est foncée, plus le chiffre d'affaires est important. Des cliques d'établissements avec des forts chiffres d'affaires semblent se former, par exemple vers Aix-en-Provence ou autour de Fos-sur-Mer. De même que dans les simulations de la section 11.1, le voisinage des établissements est représenté par une matrice de poids fondée sur la distance. Selon notre définition, chaque établissement a en moyenne 109 voisins et 76 établissements n'ont pas de voisins [12].

β_0	β_L	β_K	ρ
0.422	0.535	0.769	0.051
(0.050)	(0.015)	(0.009)	(0.009)

TABLE 11.5 – Estimation du modèle SAR : ensemble des établissements
Champ : ensemble des établissements du secteur de l'industrie, dans le département des Bouches-du-Rhône, dont le chiffre d'affaires et le total de bilan sont strictement positifs
Source : *répertoire SIRUS, 2015*
Note : Les paramètres estimés sont tous significatifs au seuil de 1 %.

La table 11.5 présente les résultats du modèle SAR estimé sur données exhaustives à l'échelle du département des Bouches-du-Rhône. Les parts des revenus du travail et du capital dans la production sont proches de celles généralement estimées (de l'ordre d'un demi à deux tiers pour la première, un tiers à deux tiers pour la seconde, le fort rendement marginal du capital pouvant ici s'expliquer par le choix du secteur industriel). Il existe bien une corrélation spatiale positive et significative : lorsque le chiffre d'affaires moyen des établissements voisins de i augmente de 1 %, le chiffre d'affaires de i augmente de 0,05 %.

11.3.4 Estimations spatiales sur des échantillons

Plans de sondage

De même que dans la section 11.1, nous répliquons l'estimation du modèle 11.3 sur un échantillon d'établissements. La sélection par sondage aléatoire simple sert de référence, mais n'est pas courante dans le cadre d'enquêtes auprès des entreprises. Les sondages stratifiés sont plus

11. Le secteur de l'industrie regroupe les établissements dont l'activité principale appartient aux divisions 10 à 33 de la NAF rév 2. 2008.

12. Ces unités sans voisins, aussi appelées "îles", ne participent donc pas à l'estimation du paramètre de corrélation spatiale ρ. Le choix du seuil résulte ainsi d'un arbitrage visant à minimiser à la fois le nombre de voisins et le nombre d'îles.

fréquemment employés dans le cadre d'études identifiant l'effet de l'effectif et du patrimoine sur le chiffre d'affaires. Ces méthodes de sondages ont été présentées en partie 11.1.2.

Dans le répertoire SIRUS, l'effectif est renseigné sur l'ensemble de la population. La stratification est effectuée selon cette variable d'effectif, sous l'hypothèse d'une corrélation entre effectif et chiffre d'affaires. La table 11.6 présente les strates ainsi constituées selon une allocation de Neyman, fondée sur la dispersion des chiffres d'affaires au sein de chacune des strates. La dispersion au sein de la strate 4 est bien supérieure à celle des autres strates, ce qui amène à considérer la strate 4 comme exhaustive, c'est-à-dire à toujours enquêter ces 67 établissements afin de limiter la variance d'estimation.

Numéro de strate	Nombre de salariés	Nombre d'établissements
1	0	3 628
2	1 à 9	2 742
3	10 à 99	770
4	100 et +	67

TABLE 11.6 – Définition des strates
Source : *répertoire SIRUS, 2013*
Champ : ensemble des établissements du secteur de l'industrie, dans le département des Bouches-du-Rhône, dont le chiffre d'affaires et le total de bilan sont strictement positifs

Résultats

Dans cette partie, nous comparons les résultats obtenus avec un plan de sondage aléatoire simple et stratifié, en faisant varier la taille de l'échantillon : $n \in \{250, 500, 1000, 2000\}$.

	Échantillon aléatoire (sondage aléatoire simple)			Échantillon stratifié		
n	ρ	β_L	β_K	ρ	β_L	β_K
250	0.011	0.554^{***}	0.768^{***}	0.015^{*}	0.311^{***}	0.813^{***}
	(0.021)	(0.104)	(0.078)	(0.009)	(0.072)	(0.056)
500	0.017	0.545^{***}	0.773^{***}	0.020^{**}	0.371^{***}	0.796^{***}
	(0.016)	(0.073)	(0.052)	(0.008)	(0.053)	(0.041)
1000	0.024^{**}	0.542^{***}	0.774^{***}	0.024^{***}	0.410^{***}	0.793^{***}
	(0.012)	(0.051)	(0.039)	(0.007)	(0.039)	(0.029)
2000	0.034^{***}	0.541^{***}	0.770^{***}	0.036^{***}	0.457^{***}	0.790^{***}
	(0.010)	(0.031)	(0.023)	(0.007)	(0.028)	(0.022)

TABLE 11.7 – Modèle 11.3 estimé sur échantillon aléatoire (sondage aléatoire simple)
Note : régression non pondérée.
Champ : ensemble des établissements du secteur de l'industrie, dans le département des Bouches-du-Rhône, dont le chiffre d'affaires et le total de bilan sont strictement positifs.
Source : *répertoire SIRUS, 2015*.

La table 11.7 présente les paramètres du modèle SAR estimés à partir de 1 000 tirages d'échantillon par sondage aléatoire simple (*à gauche*) et par sondage stratifié (*à droite*). Dans le cas d'un sondage aléatoire simple, de même que dans la section 11.1, les paramètres classiques de régression β_L et β_K, sont correctement estimés. En revanche, le paramètre de corrélation spatiale ρ n'est

significatif que pour un échantillon de taille supérieure à 1 000 et reste toujours inférieur à la valeur qu'il prend sur données exhaustives.

Le plan de sondage stratifié appliqué aux données non repondérées biaise les estimateurs classiques β_L et β_K lorsque la régression est non pondérée (DAVEZIES et al. 2009). En revanche, le biais sur le paramètre de corrélation spatiale ρ semble moindre. En effet, les grosses entreprises susceptibles d'avoir une influence spatiale importante sont toutes prises en compte dans l'échantillon du fait de ce plan de sondage stratifié.

Le choix de ne pas pondérer la régression est effectué par défaut. En économétrie classique, il est pertinent de pondérer les observations avant d'estimer un modèle économétrique lorsque la structure du plan de sondage est liée aux variables estimées. Cependant, la question de l'utilisation de poids de sondage dans le cadre d'un modèle de type SAR n'a pas été tranchée par la littérature actuelle [13]. En l'état actuel des choses, la régression non pondérée semble le choix le plus sûr et le plus simple à effectuer. Nous n'explorons pas plus avant cette question dans ce chapitre.

11.3.5 Estimation sur données agrégées

Tel qu'évoqué dans la partie 11.2, une approche couramment employée afin de contourner le problème de données manquantes est de passer à une échelle plus large par agrégation des données échantillonnées. Afin de s'abstraire des zonages administratifs, nous découpons le département des Bouches-du-Rhône selon une grille de taille $G \times G$ (figure 11.9).

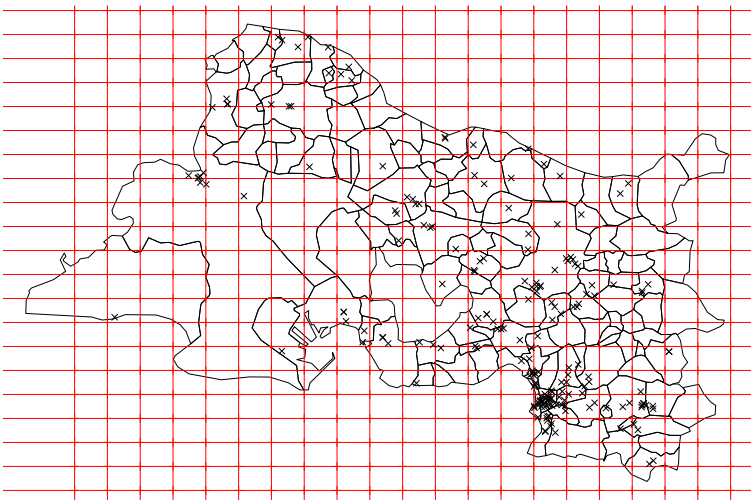

FIGURE 11.9 – Découpage des Bouches-du-Rhône selon une grille 20×20
Source : *répertoire SIRUS, 2013*
Champ : établissements du secteur de l'industrie, dans le département des Bouches-du-Rhône

À partir de ce découpage, les observations d'un échantillon sont moyennées sur chaque cellule de la grille puis l'analyse spatiale est menée à l'échelle de la grille, les distances considérées étant définies entre centroïdes des cellules. Des valeurs nulles sont assignées aux variables et aux poids spatiaux des cellules sans observations, ce qui les exclut de fait de l'estimation sans distordre la taille de la matrice de pondération spatiale. La table 11.8 présente le paramètre ρ estimé pour différentes tailles d'échantillon et diverses tailles de grille.

13. Par exemple, il n'est pas clair s'il est nécessaire ou non de faire intervenir les poids de sondage dans le calcul de la matrice de pondération spatiale $\mathbf{W}$; cela pourrait également induire de l'endogénéité supplémentaire, liée à la structure de l'échantillon.

n \ G	20	30	50	60
100	0.007	0.009	0.014	0.015
	(0.018)	(0.022)	(0.022)	(0.024)
200	0.013	0.007	0.015	0.018
	(0.021)	(0.019)	(0.018)	(0.018)
500	0.024	0.023	0.012	0.013
	(0.031)	(0.023)	(0.014)	(0.013)
1000	0.031	0.057*	0.021*	0.014
	(0.026)	(0.040)	(0.015)	(0.012)

TABLE 11.8 – Paramètre ρ estimé sur données agrégées
Champ : ensemble des établissements du secteur de l'industrie, dans le département des Bouches-du-Rhône, dont le chiffre d'affaires et le total de bilan sont strictement positifs
Source : *répertoire SIRUS, 2015*
Note : Les estimations sont réalisées sur les données d'un échantillon de taille n agrégées à l'échelle d'une grille de taille $G \times G$. Pour des raisons de lisibilité, nous ne représentons que les valeurs du paramètre ρ.

L'estimation de modèles spatiaux sur données agrégées semble permettre de contourner le problème des données manquantes dans un cadre très simple de données simulées de façon uniforme sur un territoire. Cependant, l'application de cette méthode à des données réelles n'est pas immédiate. En particulier, dans le cas présent, le paramètre d'autocorrélation spatiale est toujours sous-estimé et n'est jamais significatif. Cela pourrait être dû à la forte concentration de l'industrie des Bouches-du-Rhône, les distances intra-cellule n'étant par définition pas prises en compte par cette méthode. Les estimations spatiales sur données agrégées requièrent ainsi de s'assurer que le phénomène estimé n'est pas propre à une échelle géographique plus fine.

11.3.6 Imputation des données manquantes

Mise en œuvre

La seconde approche, évoquée en section 11.2.2, consiste à imputer les données manquantes, c'est-à-dire à attribuer des valeurs Y_i estimées aux établissements pour lesquels on n'en dispose pas. Nous considérons trois types d'imputation à l'échelle des établissements des Bouches-du-Rhône : *(i) l'imputation par le ratio*, faisant intervenir les variables L et K d'effectif et de capital comme variables explicatives du modèle, *(ii)* l'imputation par *hot deck statistique*, au sens où la distance est calculée en fonction des valeurs de L et de K, c'est à dire que les voisins d'un individu sont les établissements qui partagent des effectifs et des capitaux proches et *(iii)* l'imputation par *hot deck géographique*, où l'on associe à un individu ses voisins au sens géographique.

La mise en œuvre de ces techniques requiert, dans le premier cas, d'estimer un modèle linéaire (fonction `lm` de R) et dans les deux suivants, de définir les voisins (fonction `knn` du package *class* de R) puis de réaliser un tirage aléatoire parmi eux (fonction `sample` de R). Ces trois approches sont testées sur les données de l'industrie dans les Bouches-du-Rhône. 1 000 échantillons de taille n sont tirés selon un sondage aléatoire simple, puis le processus d'imputation assigne des valeurs de Y aux $N - n$ établissements non échantillonnés. Les résultats obtenus sont présentés dans la table 11.9. Pour rappel, les résultats sur la population entière sont en table 11.7.

	Ratio			Hot Deck Statistique			Hot Deck Géographique		
n	ρ	β_L	β_K	ρ	β_L	β_K	ρ	β_L	β_K
250	0.002	0.560***	0.768***	0.043***	0.664***	0.646***	0.419***	0.028	0.104***
	(0.002)	(0.112)	(0.080)	(0.009)	(0.083)	(0.059)	(0.046)	(0.034)	(0.023)
500	0.004	0.548***	0.774***	0.042***	0.613***	0.698***	0.412***	0.061*	0.149***
	(0.003)	(0.077)	(0.058)	(0.008)	(0.061)	(0.044)	(0.035)	(0.034)	(0.022)
1000	0.008**	0.546***	0.774***	0.040***	0.577***	0.734***	0.389***	0.116***	0.217***
	(0.003)	(0.051)	(0.037)	(0.007)	(0.040)	(0.028)	(0.035)	(0.035)	(0.023)
2000	0.017***	0.542***	0.773***	0.040***	0.562***	0.751***	0.333***	0.203***	0.338***
	(0.004)	(0.032)	(0.024)	(0.007)	(0.031)	(0.023)	(0.022)	(0.034)	(0.022)

TABLE 11.9 – Méthodes d'imputation
Source : *répertoire SIRUS, 2015*
Champ : ensemble des établissements du secteur de l'industrie, dans le département des Bouches-du-Rhône, dont le chiffre d'affaires et le total de bilan sont strictement positifs

Résultats

Les résultats sont très variables selon la méthode utilisée. *L'imputation par le ratio* permet de bien conserver la structure linéaire entre chiffre d'affaires, effectif et capital, ce qui se traduit par des estimations sans biais et précises des coefficients β_L et β_K. En revanche, le ρ estimé est très faible, encore plus que dans le cas du sondage aléatoire simple exploité directement (voir table 11.7). En effet, l'imputation ne prend absolument pas en compte la structure spatiale, qui est effacée lors de l'estimation du modèle sur les données complétées. Il n'est donc pas pertinent d'essayer d'appliquer des modèles d'économétrie spatiale sur des données imputées avec cette méthode.

L'imputation par *hot deck statistique* semble plus prometteuse. Les estimateurs sont du bon ordre de grandeur par rapport aux valeurs obtenues sur la population et sont estimés avec précision. Une comparaison avec la table 11.5 révèle un biais lorsque $\hat{\rho}$, $\hat{\beta}_L$ et $\hat{\beta}_K$ sont estimés sur des échantillons de petite taille. Ainsi, l'imputation par hot deck biaise les estimateurs du modèle (CHARREAUX et al. 2016) mais permet de faire ressortir la structure des corrélations spatiales. En effet, le lien entre donneur et receveur semble conserver de façon implicite la structure des interactions spatiales. Il est également possible que la structure spatiale sous-jacente à Y existe aussi pour L et K et soit récupérée par imputation. Ainsi, l'emploi de cette méthode d'imputation à des fins d'analyse économétrique revient à un arbitrage entre biais et variance sur les paramètres β_L et β_K, classique en théorie des sondages. Cependant, dans le cas présent, la méthode présente en outre l'avantage de réduire considérablement le biais préexistant sur ρ. Ces résultats, testés uniquement sur ce jeu de données et sur un plan de sondage simple, sont à utiliser avec précaution. En tout état de cause, ce n'est pas sur la proximité spatiale entre donneur et receveur que repose l'efficacité de cette méthode, comme le montre le dernier exemple.

La méthode d'imputation par *hot deck géographique* conduit à des résultats aberrants. Se fondant directement sur la proximité spatiale entre donneur et receveur, elle entraîne une surestimation très forte de l'effet spatial (ρ très supérieur à la vraie valeur), au détriment de l'effet des autres variables du modèle (β_1 et β_2 très inférieurs aux vraies valeurs). En effet, selon cette méthode, des établissement spatialement proches auront le même chiffre d'affaires Y, ce qui crée *ex-nihilo* une très forte corrélation spatiale positive. L'utilisation de la dimension spatiale pour pallier le problème des données manquantes n'est pas immédiate. La table 11.12 en annexe 11.3.6 présente les résultats obtenus pour une imputation par hot deck géographique en se limitant aux établissements ayant des effectifs proches. Le paramètre ρ est moins surestimé mais les résultats restent très éloignés de

l'estimation sur données exhaustives. Il semblerait possible d'utiliser l'information géographique de façon parcimonieuse pour l'imputation, mais cela demanderait une analyse plus poussée du jeu de données et une bonne connaissance de sa structure spatiale.

Conclusion

Ce chapitre met en évidence les difficultés liées à l'application de modèles d'économétrie spatiale à des données échantillonnées. Deux écueils s'y opposent en particulier : *(i)* un "effet taille" par lequel l'estimation sur un échantillon distord la matrice de pondération spatiale, et *(ii)* un effet résultant de l'omission d'unités spatialement corrélées avec les unités observées. Ces deux effets tendent à sous-estimer l'ampleur de la corrélation spatiale. Néanmoins, ce biais est plus faible dans le cas d'un sondage par grappes et lorsque l'échantillon est plus important.

Les études empiriques résolvent généralement ce problème en ignorant les observations manquantes, en agrégeant les données à une échelle plus large ou en imputant les valeurs manquantes. La première solution n'est jamais souhaitable. Les deux autres sont loin d'être parfaites, la difficulté étant de reconstituer un ensemble d'information complexe à partir de peu d'observations. L'imputation par hot deck statistique est prometteuse, mais nous ne montrons pas sa validité dans un cas général.

Si cette problématique est vouée à se développer avec l'importance des réseaux sociaux et des données géolocalisées, l'estimation de modèles spatiaux sur des données échantillonnées reste rare. En l'état, il reste préférable de considérer des données exhaustives. Le présent chapitre met en garde contre les solutions trop expéditives, telles que l'agrégation des données à une échelle supérieure, les méthodes d'imputation simplistes ou la suppression des données manquantes. Lorsque qu'un échantillon relativement important est disponible, ou issu d'un sondage par grappes, une estimation spatiale pourrait alors être envisagée, en gardant à l'esprit que le paramètre de corrélation spatiale obtenu sera sans doute sous-estimé.

Annexe

Choix du modèle et de la matrice de voisinage

Les tables 11.10 et 11.11 présentent des résultats obtenus en termes d'estimation des paramètres de modèles SAR ou SEM *via* une méthode Monte Carlo selon différentes matrices de voisinage et différentes tailles d'échantillon.

	ρ			β		
$\mathscr{M}$ / n	2 voisins	5 voisins	Distance	2 voisins	5 voisins	Distance
50	0.020	−0.003	0.042	1.107***	1.050***	1.054***
	(0.110)	(0.172)	(0.043)	(0.115)	(0.095)	(0.125)
100	0.063	0.069	0.058*	1.112***	1.056***	1.054***
	(0.076)	(0.111)	(0.031)	(0.079)	(0.065)	(0.086)
150	0.097*	0.115	0.073**	1.107***	1.052***	1.049***
	(0.060)	(0.088)	(0.028)	(0.062)	(0.051)	(0.068)
250	0.150***	0.189**	0.101**	1.105***	1.050***	1.053***
	(0.047)	(0.065)	(0.026)	(0.049)	(0.040)	(0.052)

TABLE 11.10 – Modèle SAR - Estimation par Monte Carlo
Note : Les écarts types sont entre parenthèses.

	λ			β		
$\mathscr{M}$ / n	2 voisins	5 voisins	Distance	2 voisins	5 voisins	Distance
50	−0.025	−0.110	0.008	1.003***	1.003***	1.002***
	(0.167)	(0.287)	(0.193)	(0.115)	(0.113)	(0.112)
100	0.008	−0.027	0.024	1.003***	1.004***	1.003***
	(0.113)	(0.182)	(0.124)	(0.080)	(0.078)	(0.078)
150	0.023	0.002	0.034	0.998***	0.998***	0.998***
	(0.090)	(0.144)	(0.099)	(0.065)	(0.063)	(0.063)
250	0.047	0.042	0.052	1.000***	1.000***	1.000***
	(0.069)	(0.108)	(0.079)	(0.051)	(0.050)	(0.050)

TABLE 11.11 – Modèle SEM - Estimation par Monte Carlo
Note : Les écarts types sont entre parenthèses

Imputation par hot deck géographique stratifié

La table 11.12 donne les résultats obtenus pour une imputation par hot deck géographique en se limitant aux établissements ayant des effectifs proches, c'est à dire ceux de la même strate (définie dans la table 11.6) que l'établissement ayant une valeur manquante.

	Hot Deck Géographique Stratifié		
n	ρ	β_L	β_K
250	0.137	1.216	0.029
	(0.037)	(0.100)	(0.026)
500	0.148	1.192	0.071
	(0.031)	(0.077)	(0.025)
1000	0.156	1.121	0.149
	(0.026)	(0.061)	(0.025)
2000	0.148	0.542	0.279
	(0.019)	(0.048)	(0.025)

TABLE 11.12 – Une autre méthode d'imputation
Source : *répertoire SIRUS, 2015*
Champ : ensemble des établissements du secteur de l'industrie, dans le département des Bouches-du-Rhône, dont le chiffre d'affaires et le total de bilan sont strictement positifs

Références - Chapitre 11

ANSELIN, Luc (2002b). « Under the hood : Issues in the specification and interpretation of spatial regression models ». *Agricultural Economics* 27.3, p. 247–267.

ARBIA, Giuseppe, Giuseppe ESPA et Diego GIULIANI (2016). « Dirty spatial econometrics ». *The Annals of Regional Science* 56.1, p. 177–189.

ARDILLY, Pascal (1994). *Les techniques de sondage*.

BELOTTI, F., G. HUGHES et A. Piano MORTARI (2017a). « Spatial panel-data models using Stata ». *Stata Journal* 17.1, 139–180(42).

BOEHMKE, Frederick J., Emily U. SCHILLING et Jude C. HAYS (2015). *Missing data in spatial regression*. Rapp. tech. Society for Political Methodology Summer Conference.

BURT, Ronald S. (1987). « A Note on Missing Network Data in the General Social Survey ». *Social Networks* 9, p. 63–73.

CHARREAUX, C et al. (2016). « Econométrie et Données d'Enquête : les effets de l'imputation de la non-réponse partielle sur l'estimation des paramètres d'un modèle économétrique ».

CHOW, Gregory C. et An-Loh LIN (1976). « Best Linear Unbiased Estimation of Missing Observations in an Economic Time Series ». *Journal of the American Statistical Association* 71.355, p. 719–721.

CLIFF, A.D. et J.K. ORD (1972). *Spatial autocorrelation*. Pion, London.

COCHRAN, William G (2007). *Sampling techniques*. John Wiley & Sons.

DAVEZIES, L. et X. D'HAULTFOEUILLE (2009). *To Weight or not to Weight ? The Eternal Question of Econometricians facing Survey Data*. Documents de Travail de la DESE - Working Papers of the DESE g2009-06. INSEE, DESE.

DEMPSTER, A.P., N.M. LAIRD et D.B. RUBIN (1977). « Maximum likelihood from incomplete data via the EM algorithm ». *Journal of the royal statistical society* 39.1, p. 1–38.

EGGER, Peter et Michael PFAFFERMAYR (2006). « Spatial convergence ». *Papers in Regional Science* 85.2, p. 199–215.

ERTUR, Cem et Wilfried KOCH (2007). « Growth, technological interdependence and spatial externalities : theory and evidence ». *Journal of Applied Econometrics* 22.6, p. 1033–1062.

FERREIRO, Osvaldo (1987). « Methodologies for the estimation of missing observations in time series ». *Statistics and Probability Letters* 5.1, p. 65–69.

GILE, Krista J. et Mark S. HANDCOCK (2010). « Respondent-driven sampling : an assessment of current methodology ». *Sociological Methodology* 40.1, p. 285–327.

GOULARD, M., T. LAURENT et C. THOMAS AGNAN (2013). « About predictions in spatial autoregressive models : Optimal and almost optimal strategies ». *Toulouse School of Economics Working Paper* 13, p. 452.

HARVEY, A. C. et R. G. PIERSE (1984). « Estimating Missing Observations in Economic Time Series ». *Journal of the American Statistical Association* 79.385, p. 125–131.

HUISMAN, Mark (2014). *Imputation of missing network data*. Sous la dir. de Reda ALHAJJ et Jon ROKNE. T. 2. Springer, p. 707–715. ISBN : 978-1-4614-6169-2.

JONES, Richard H. (1980). « Maximum Likelihood Fitting of ARMA Models to Time Series with Missing Observations ». *Technometrics* 22.3, p. 389–395.

KELEJIAN, H.H. et I.R. PRUSHA (2010b). « Spatial models with spatially lagged dependent variables and incomplete data ». *Journal of geographical systems*.

KOSKINEN, Johan H., Garry L. ROBINS et Philippa E. PATTISON (2010). « Analysing exponential random graph (p-star) models with missing data using Bayesian data augmentation ». *Statistical Methodology* 7.3", p. 366–384.

KOSSINETS, Gueorgi (2006). « Effects of missing data in social networks ». *Social Networks* 28.3, p. 247–268.

LESAGE, James P., Manfred M. FISCHER et Thomas SCHERNGELL (2007). « Knowledge spillovers across Europe : Evidence from a Poisson spatial interaction model with spatial effects ». *Papers in Regional Science* 86.3, p. 393–421. ISSN : 1435-5957.

LESAGE, J.P. et R.K. PACE (2004). « Models for spatially dependent missing data ». *The journal of real estate finance and economics* 29.2, p. 233–254.

LITTLE, Roderick J. A. (1988). « Missing-Data Adjustments in Large Surveys ». *Journal of Business and Economic Statistics* 6.3, p. 287–296.

LITTLE, Roderick J. A. et Donald B. RUBIN (2002). *Statistical analysis with missing data*. 2nd. Wiley, Hoboken.

LIU, Xiaodong, Eleonora PATACCHINI et Edoardo RAINONE (2017). « Peer effects in bedtime decisions among adolescents : a social network model with sampled data ». *The Econometrics Journal*.

LÓPEZ-BAZO, Enrique, Esther VAYÁ et Manuel ARTÍS (2004). « Regional Externalities And Growth : Evidence From European Regions ». *Journal of Regional Science* 44.1, p. 43–73.

MARDIA, Kanti V. et al. (1998). « The Kriged Kalman filter ». *Test* 7.2, p. 217–282.

PINKSE, Joris et Margaret E. SLADE (2010). « The Future of Spatial Econometrics ». *Journal of Regional Science* 50.1, p. 103–117.

REVELLI, Federico et Per TOVMO (2007). « Revealed yardstick competition : Local government efficiency patterns in Norway ». *Journal of Urban Economics* 62.1, p. 121–134.

RUBIN, Donald B. (1976). « Inference and missing data ». *Biometrika* 63, p. 581–592.

STORK, Diana et William D. RICHARDS (1992). « Nonrespondents in Communication Network Studies ». *Group & Organization Management* 17.2, p. 193–209.

TILLÉ, Y. (2001). *Théorie des sondages : échantillonnage et estimation en populations finies : cours et exercices avec solutions :[2e cycle, écoles d'ingénieurs]*. Dunod.

WANG, W. et L.-F. LEE (2013a). « Estimation of spatial autoregressive models with randomly missing data in the dependent variable ». *The Econometrics Journal* 16.1, p. 73–102.

ZHOU, Jing et al. (2017). « Estimating Spatial Autocorrelation With Sampled Network Data ». *Journal of Business and Economic Statistics* 35.1, p. 130–138.