JMS Journées de méthodologie statistique de l'Insee 2018

Économétrie spatiale sur données d'enquête

Thomas MERLY-ALPA(*), Raphaël LARDEUX(**)

(*) Insee, Direction de la Méthodologie et de la Coordination Internationale
(**) Insee, Direction des Études et des Synthèses Économiques

thomas.merly-alpa@insee.fr, raphael.lardeux-schutz@insee.fr

Mots-clés. Économétrie spatiale, autocorrélation spatiale, données d'enquête, imputation.

Résumé

L'économétrie spatiale requiert des données exhaustives sur un territoire, ce qui interdit en principe l'utilisation de données d'enquête. Cet article présente les écueils relatifs à l'estimation d'un modèle spatial autorégressif (SAR) sur données échantillonnées et évalue quelques méthodes classiques de correction. Nous identifions deux sources de biais : *(i)* un « effet taille » résultant de la restriction à un sous-ensemble localement complet d'un territoire et *(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. Ces deux effets amènent à sous-estimer la corrélation spatiale, mais moins fortement dans le cas d'un sondage par grappes ou lorsque l'échantillon est suffisamment grand. Deux solutions sont évaluées : le passage à l'échelle supérieure par agrégation et l'imputation des valeurs manquantes (par régression linéaire ou hot deck). Elles ne fonctionnent que sous des hypothèses très restrictives, la difficulté étant de reconstituer une information complexe avec peu d'observations. La dernière partie illustre cette problématique par l'estimation d'externalités de production entre les industries du département français des Bouches-du-Rhône.

Abstract

This paper presents several issues resulting from the estimation of spatial econometric models in social sciences. Taking into account spatial interactions requires exhaustive data on a population, as observations are not independent from each other. However, socioeconomic statistics are generally based on survey data, gathering a very rich information on a sample of the population. We focus here on the possibility to detect spatial autocorrelation when only a sample of the population is available. We simulate data from a SAR generating process and run Monte-Carlo estimations on samples from this population of counties drawn from different sampling methods. Except with specific sampling, we find little spatial autocorrelation because such data are locally sparse. Some solutions (data aggregation, imputation) are explored in this paper. Finally, the last part of the paper illustrates the problem with the estimation of a production function on French industries in the Bouches-du-Rhône.

Remerciements

Cet article est très largement inspiré du chapitre du même nom du Manuel de Statistique Spatiale, dont la publication a été organisée par le Département des Méthodes Statistiques de l'Insee. Les auteurs remercient Marie-Pierre de Bellefon et Vincent Loonis pour la coordination du manuel, ainsi que Ronan le Saout pour sa relecture attentive.

Introduction

Les développement récents de la géolocalisation et de l'économétrie spatiale favorisent l'analyse de phénomènes spatiaux à des échelles très locales. Cependant, l'application de ces méthodes requiert des données exhaustives, qui sont loin d'être toujours accessibles (non-réponse, temps de collecte trop important, ...) et ne peuvent pas être aisément 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 finement prendre en compte les corrélations spatiales. Pinkse et Slade, 2010 vont jusqu'à qualifier ces perspectives de « futur de l'économétrie spatiale ».

Dans cet article, nous discutons les développements récents relatifs à l'application des méthodes d'estimation spatiale lorsqu'une partie des observations est manquante, en particulier dans le cas de données d'enquête. Nous montrons que l'application de modèles d'économétrie spatiale à des données échantillonnées amène à sous-estimer l'ampleur des corrélations spatiales, tout particulièrement dans le cas d'un sondage aléatoire simple et lorsque l'échantillon est plus petit. Si ignorer les observations manquantes n'est jamais souhaitable, les autres solutions envisageables telles que l'imputation des valeurs manquantes ou le passage à l'échelle supérieure par agrégation ne fonctionnent que sous des hypothèses très restrictives. Enfin, dans cet article, nous ne traitons ni la possibilité d'un sondage spatialisé, complexe dans le cas des données sociales, 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, en économétrie spatiale, les observations sont considérées comme corrélées entre elles : chaque unité est influencée par ses voisins et réciproquement. 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 le statut qu'elle confère aux 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 [1]. L'économétrie spatiale a été développée dans le cadre des modèles de Cliff et Ord, 1972, caractérisé par une information exhaustive et parfaite sur les unités spatiales et par l'absence de données manquantes (Arbia, Espa et Giuliani, 2016). En pratique, ces conditions ne sont quasiment jamais réunies et appliquer directement des techniques d'estimation spatiale peut fortement altérer les résultats.

1. L'analyse spatiale se rapproche en cela des séries temporelles, où le jeu de données observé est issu d'un processus stochastique.

L'application de méthodes spatiales à des données échantillonnées pose plusieurs problèmes. Premièrement, les estimations sont perturbées par un « effet taille ». L'existence de m données manquantes parmi une population de taille n amène à considérer une matrice de pondération spatiale de taille $(n-m)\times(n-m)$ au lieu de la vraie matrice de pondération de taille $n\times n$. Ce changement de taille affecte en soi l'estimation du paramètre de corrélation spatiale (Arbia, Espa et Giuliani, 2016). Nous montrons ainsi que l'estimation d'un modèle SAR sur un sous-ensemble localement complet d'un territoire ne permet pas de retrouver le paramètre de corrélation spatiale ayant servi à simuler les observations à l'échelle de ce territoire. Deuxièmement, le tirage aléatoire des unités donne lieu à un échantillon plus ou moins dispersé sur le territoire (selon le processus de sondage retenu), ce qui engendre une erreur de mesure sur l'effet du voisinage (régresseur $\mathcal{W}Y$) et donc un biais dans l'estimation du paramètre de corrélation spatiale. En procédant par simulation, nous montrons qu'au-delà de l'« effet taille », cet effet lié à la répartition spatiale des observations a des conséquences importantes. En corollaire, à nombre d'observations donné, il est toujours préférable que ces observations soient regroupées entre elles en vue d'estimer des corrélations spatiales.

Différentes corrections ont été proposées, sans qu'aucune ne s'impose radicalement [2]. Lorsque la localisation des individus est connue, les solutions par imputation sont généralement privilégiées (D. B. Rubin, 1976 ; Little, 1988 ; Little et D. B. Rubin, 2002). Cependant, une imputation naive, par exemple par un modèle linéaire, ne permet pas de corriger les biais (Belotti, Hughes et Mortari, 2017). Pour contourner ce problème, Kelejian et Prusha, 2010 développent des estimateurs lorsque seul un sous-ensemble incomplet d'une population est disponible. Wang et Lee, 2013 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, J. LeSage et Pace, 2004 recourent à l'algorithme EM (Dempster, Laird et D. Rubin, 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, Schilling et Hays, 2015 étendent cette procédure au cas d'observations manquantes (variables dépendante et indépendantes inconnues).

Des travaux empiriques illustrent l'importance de ces corrections. Dans un modèle de prix hédoniques, J. LeSage et Pace, 2004 appliquent l'algorithme EM pour prédire la valeur des logements non-vendus. Dans un modèle de réseaux avec autocorrélation spatiale, Liu, Patacchini et Rainone, 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 Tovmo, 2007), ce qui peut amener à sous-estimer les corrélations à la bordure de l'espace considéré (Kelejian et Prusha, 2010). Enfin, la plupart des applications sont réalisées sur données agrégées pour bénéficier de données exhaustives sur une échelle plus

2. 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. D. B. Rubin, 1976 ; Huisman, 2014.

large, mais cette solution peut provoquer des erreurs positionnelles (Arbia, Espa et Giuliani, 2016) ou des biais écologiques (Anselin, 2002). 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 Lin, 1976, Ferreiro, 1987). Jones, 1980 ; Harvey et Pierse, 1984 recommandent l'utilisation d'un filtre de Kalman pour simultanément estimer un modèle et imputer des valeurs. L'analyse géostatistique corrige des 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). 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 fondementaux 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 Richards, 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, Robins et Pattison, 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 Handcock, 2010).

Le présent article se concentre sur deux questions : quels sont les biais engendrés par l'application de méthodes d'estimation spatiale à des données d'enquête ? quelles sont les conséquences des diverses corrections classiques (suppression des données, imputation, agrégation) ? Ces questions sont abordées par Arbia, Espa et Giuliani, 2016, qui notent 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 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 2 discute les conséquences de quelques solutions usuelles : le passage à l'échelle supérieure par agrégation et l'imputation des valeurs manquantes. La section 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.

1 Première approche par simulations

Cette première partie met en évidence les limites liées à l'estimation d'un modèle autorégressif spatial (SAR) sur des données échantillonnées. 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 de ce paramètre.

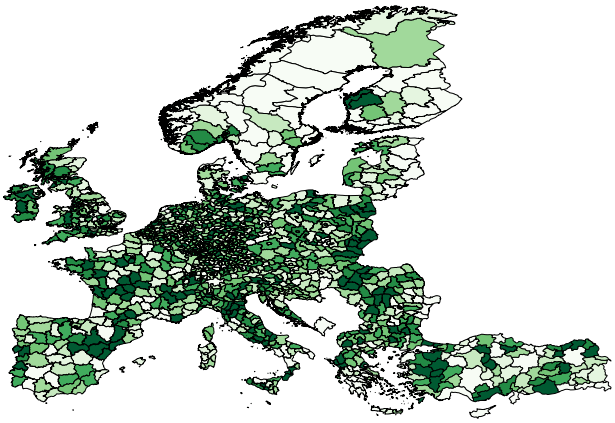
1.1 Simulation d'un SAR

L'espace géographique retenu est une carte de l'Europe [3], 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 un espace géographique homogène et compact. À partir du *shapefile* de l'Europe, nous construisons une matrice de voisinage $\mathcal{W}$ 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 : $\varepsilon \hookrightarrow \mathcal{N}(0,1)$ et $X \hookrightarrow \mathcal{N}(5,2)$. Enfin, la variable dépendante Y suit un modèle SAR :

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

avec $\beta = 1$ et $\rho = 0,5$. L'enjeu est d'évaluer la capacité des modèles économétriques spatiaux à estimer la valeur de ces paramètres de référence à partir d'un échantillon de zones. Les valeurs de la variable simulée Y sont représentées sur la figure 1. La présence de zones colorées concentrées est caractéristique de l'autocorrélation spatiale positive résultant du processus générateur des données.

FIGURE 1 – Variable dépendante simulée selon un modèle SAR



Note : Chaque zone NUTS3 est colorée selon la valeur de Y_i, en vert d'autant plus foncé que cette valeur est élevée.
©EuroGeographics pour les limites administratives

La Table 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 $\hat{\beta}$ et $\hat{\rho}$ estimés sont très proches des valeurs calibrées initialement.

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

TABLE 1 – Paramètres estimés par SAR sur l'ensemble des zones

$\hat{\beta}$	$\hat{\rho}$	Direct	Indirect	Total
0.99	0.49	1.04	0.86	1.90

1.2 Tirage des échantillons et estimations

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.

Dans cette perspective, nous évaluons 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, LAURENT et THOMAS AGNAN, 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 ?

FIGURE 2 – Un échantillon obtenu par SAS ($n = 500$)

©EuroGeographics pour les limites administratives.

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 la variable dépendante 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 $\hat{\rho}$ et $\hat{\beta}$ 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 $\mathcal{W}_{\text{echantillon}}$ fondée sur la distance, tel que précédemment, mais limitée aux unités présentes dans l'échantillon. Deux méthodes d'échantillonnage sont considérées : *(i)* le sondage aléatoire simple (SAS) et *(ii)* le sondage par grappes.

1.2.1 Sondage aléatoire simple

La Table 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. Quelle que soit la taille de l'échantillon, le paramètre estimé $\hat{\beta}$ n'est pas significativement différent de sa vraie valeur ($\beta = 1$), tandis que le paramètre estimé $\hat{\rho}$ est quant à lui 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.

Ne vaudrait-il pas mieux estimer le modèle par les moindres carrés ordinaires (MCO) sur données d'enquête, même si le "vrai" modèle sous-jacent aux données est spatial ? De manière générale, les MCO fournissent une valeur biaisée du paramètre β. En effet, l'estimateur $\hat{\beta}^{MCO}$ obtenu à partir d'un modèle sous-jacent spatial vaut :

$$(X'X)^{-1}X'\left[(\mathbb{I} - \rho\mathcal{W})^{-1}X\beta + u\right]$$

avec u qui suit une loi normale centrée. Ainsi :

$$E[\hat{\beta}^{MCO}] = \left(X'X\right)^{-1}X'(\mathbb{I} - \rho\mathcal{W})^{-1}X\beta \neq \beta$$

L'estimateur des MCO ne sera sans biais qu'en l'absence de corrélations spatiales ($\rho = 0$).

TABLE 2 – Estimation d'un modèle SAR sur des échantillons tirés par SAS

n	$\hat{\rho}$	$\hat{\beta}$	Direct	Indirect	Total
50	0.04	1.06***	1.06***	0.02	1.07***
	(0.04)	(0.13)	(0.13)	(0.02)	(0.13)
100	0.06*	1.05***	1.05***	0.03*	1.08***
	(0.03)	(0.09)	(0.09)	(0.02)	(0.09)
150	0.07**	1.05***	1.05***	0.05**	1.10***
	(0.03)	(0.07)	(0.07)	(0.02)	(0.07)
250	0.10***	1.05***	1.05***	0.08***	1.14***
	(0.03)	(0.05)	(0.05)	(0.02)	(0.06)

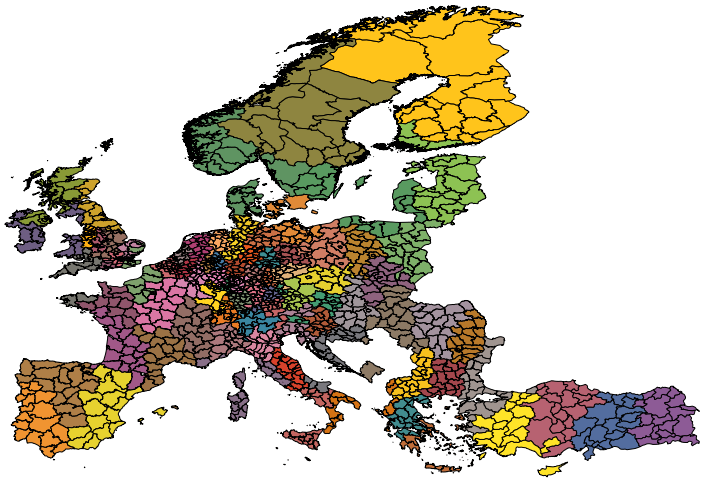
*Note : Ces estimations proviennent de 10 000 simulations. Le paramètre $\hat{\beta}$ du modèle SAR estimé par Monte Carlo pour un échantillon de taille 100 est de 1.050, avec un écart-type empirique de 0.087. *** désigne une significativité à 1%, ** à 5% et * à 10%.*

1.2.2 Sondage par grappes

Nous choisissons de constituer des grappes regroupant des zones NUTS3. En effet, les niveaux plus agrégés (NUTS1 ou NUTS2) sont de taille importante et ne comportent pas tous le même

nombre de NUTS3. Considérer des grappes de taille trop importante limite le nombre de simulations possibles et constituer des grappes comportant des nombres de zones trop différents introduit soit une problématique de poids de sondage différents entre les individus, que nous ne souhaitons pas traiter ici (voir Davezies et D'Haultfoeuille, 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. Ainsi, nous séparons les zones en grappes de même taille tout en maintenant une certaine cohérence géographique. Enfin, étant donné que la matrice de pondération spatiale est fondée sur la distance géographique, nous privilégions les grappes les moins étendues possible.

Figure 3 – Regroupement par grappes des régions NUTS 3 Européennes



Note : Les NUTS3 proches de même couleur forment une grappe de 17 zones. ©EuroGeographics pour les limites administratives.

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 en Figure 3.

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. Nous procédons à des tirages de nombres différents de grappes, allant de 3 à 15 grappes (51 à 255 zones). La Table 3 montre les résultats obtenus pour des valeurs de $n = 17p$, la taille de l'échantillon composé de p grappes.

Avec un sondage par grappes, le paramètre $\hat{\rho}$ est plus proche de sa vraie valeur, l'inclut dans son intervalle de confiance mais continue de la sous-estimer. Il en va de même pour l'effet indirect. La précision de l'estimation s'améliore nettement lorsque n augmente. Ainsi, contrairement au cas du SAS, 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

TABLE 3 – Estimation d'un modèle SAR sur des échantillons tirés par grappes

n	p	$\hat{\rho}$	$\hat{\beta}$	Direct	Indirect	Total
51	3	0.31*	1.02***	1.05***	0.44*	1.49***
		(0.24)	(0.09)	(0.10)	(0.26)	(0.31)
102	6	0.35***	1.02***	1.05***	0.49***	1.55***
		(0.10)	(0.06)	(0.07)	(0.19)	(0.22)
153	9	0.36***	1.02***	1.05***	0.52***	1.57***
		(0.08)	(0.05)	(0.06)	(0.15)	(0.18)
255	15	0.38***	1.01***	1.05***	0.54***	1.59***
		(0.06)	(0.04)	(0.04)	(0.12)	(0.14)

*Note : Ces estimations proviennent de 10 000 simulations. Le paramètre $\hat{\beta}$ du modèle SAR estimé par Monte Carlo pour un échantillon de 6 grappes est de 1.017, avec un écart-type empirique de 0.063. *** désigne une significativité à 1%, ** à 5% et * à 10%.*

et donc très représentatives de la corrélation spatiale. Par contre, si le nombre d'unités tirées est faible, la corrélation spatiale sera peu précisément estimée. 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 ? Étant donné que l'on dispose de peu de valeurs de X et Y, le terme WY est paradoxalement assez bien connu, ce qui pourrait amener à favoriser cette piste. D'autre part, et cela sera developpé dans la partie 1.3, 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.

1.3 L' « Effet taille »

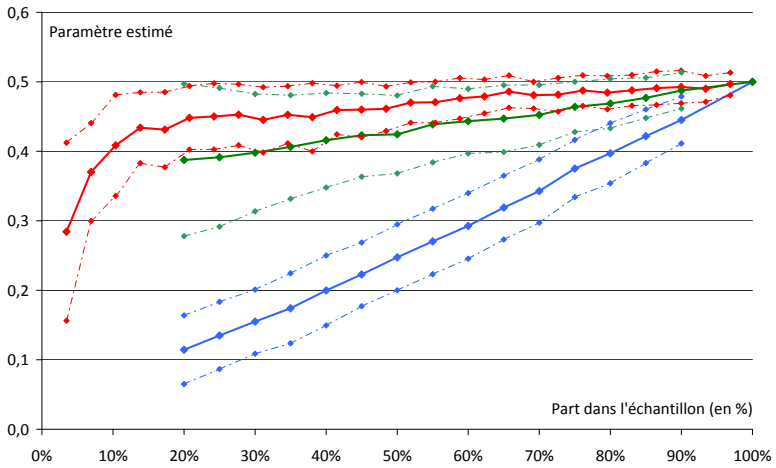
Les résultats obtenus par simulation peuvent étonner les économètres. En effet, dans le cadre d'un sondage aléatoire simple, assimilable au modèle de superpopulation utilisé en économétrie, l'estimation d'un paramètre d'une population ou d'un modèle est usuellement sans biais, tant que le plan d'échantillonnage est correctement spécifié. Il apparaît alors que le paramètre d'autocorrélation spatiale ρ ne suit pas cette « loi » classique de la théorie des sondages [4].

Se restreindre à un nombre de zones inférieur à celui de la population entière induit en soi une modification de la structure spatiale sous-jacente. Intuitivement, l'effet spatial estimé sur un territoire résulte des liens entre toutes les zones qui le composent. Dans le cas où l'effet est uniforme, omettre quelques zones revient a négliger leur contribution - directe et indirecte - à l'effet global et donc à sous-estimer ce dernier. Nous nommons « effet taille » cette première composante qui vient sous-estimer le paramètre de corrélation spatiale. De plus, les données

4. 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 $\hat{\rho}$.

disponibles peuvent être plus ou moins dispersées sur le territoire. Or des observations trop écartées les unes des autres ne seront pas en mesure de rendre compte finement de la structure des corrélations spatiales. Cet effet de répartition amène d'autant plus à sous-estimer le paramètre de corrélation spatiale.

FIGURE 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 du barycentre puis en s'en éloignant progressivement. Le courbes en pointillés représentent les intervalles de confiance à 95%.

Les précédentes estimations résultent conjointement des effets taille et de répartition. Pour isoler l'effet de la taille de l'échantillon disponible, 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 complète 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 2). La Figure 4 compare les valeurs de $\hat{\rho}$ obtenues en suivant trois protocoles différents pour retenir une proportion P de la population totale : *(i)* sélection parmi les NUTS3 de la fraction P des zones les plus centrales, *(ii)* sondage par grappes où chaque grappe de zones a $P\%$ de chances d'être sélectionnée et *(iii)* sondage poissonnien classique où chaque zone a indépendamment $P\%$ de chances d'être sélectionnée.

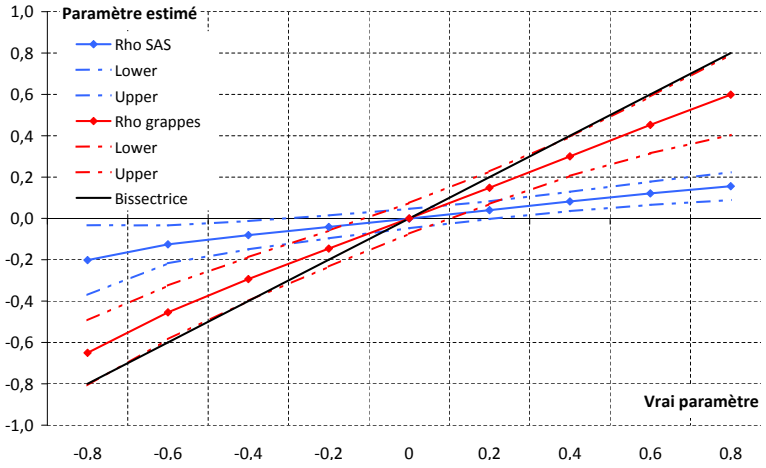
De même que dans la Partie 1.2, le sondage poissonnien (proche du SAS) donne des valeurs estimées de $\hat{\rho}$ bien plus faibles que le sondage par grappes. L'apport principal de cette figure est dans la courbe rouge, qui repose sur une sélection non aléatoire d'une partie des zones. Elle converge plus rapidement que les autres vers la vraie valeur de $\rho = 0,5$. 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.

1.4 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 en Annexe 5.1). Enfin la vraie valeur du paramètre ρ n'affecte pas les

résultats des estimations. La Figure 5 montre que, à taux de sondage donné, une estimation sur un échantillon tiré par SAS 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*) : $Y_2 = X\beta + (1 - \lambda\mathcal{W})^{-1}\,\varepsilon$ n'affecte pas radicalement les résultats (Table 12 en Annexe 5.1).

FIGURE 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 $\hat{\rho}$.

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. Comme montré dans la partie précédente, ce choix atténue le paramètre de corrélation spatiale relativement à sa vraie valeur, par l'effet de la taille de l'échantillon et de la répartition spatiale des données qui le constituent.

Ces effets résultent d'un problème de valeurs manquantes. Pour le pallier, il est nécessaire de comparer données exhaustives et données échantillonnées selon une même structure géographique et un même nombre d'unités. Il faut également être en mesure de reconstituer les corrélations spatiales entre unités observées et manquantes. Dans le cas présent, la localisation des unités est toujours supposée connue [5].

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 la structure des corrélations spatiales.

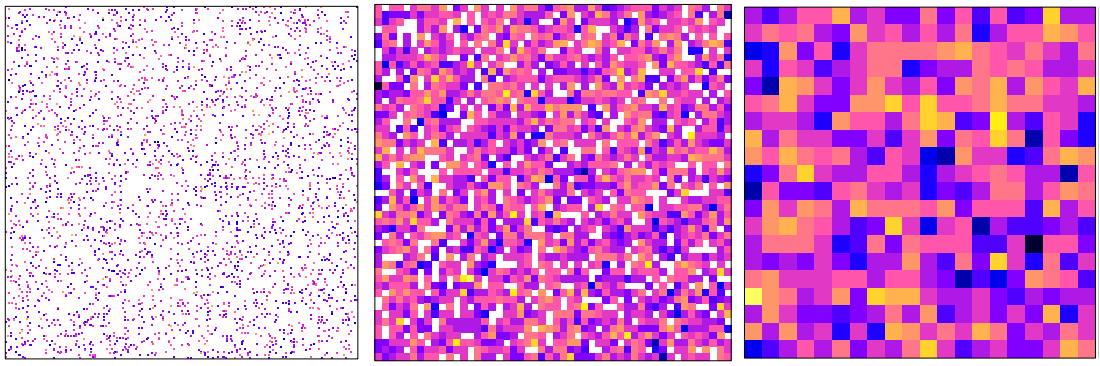
5. Le manque d'information concernant la localisation de certaines unités est un autre enjeu des recherches actuelles en économétrie spatiale (ARBIA, ESPA et GIULIANI, 2016) qui dépasse cependant le cadre du présent article.

2.1 Passer à l'échelle supérieure par agrégation

2.1.1 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 crucialement 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, 2002 pour plus de précisions). Pour y recourir, il faut supposer que les effets mesurés sont invariants selon l'échelle. Les corrélations intra-zone sont alors omises, au profit des corrélations entre zones.

Figure 6 – Agrégation de données spatiales



Note : ***A gauche :*** *données simulées.* ***Au centre :*** *données agrégées selon* 50×50 *parcelles.* ***A droite :*** *données agrégées selon* 20×20 *parcelles. Les cases blanches représentent des zones sans points.*

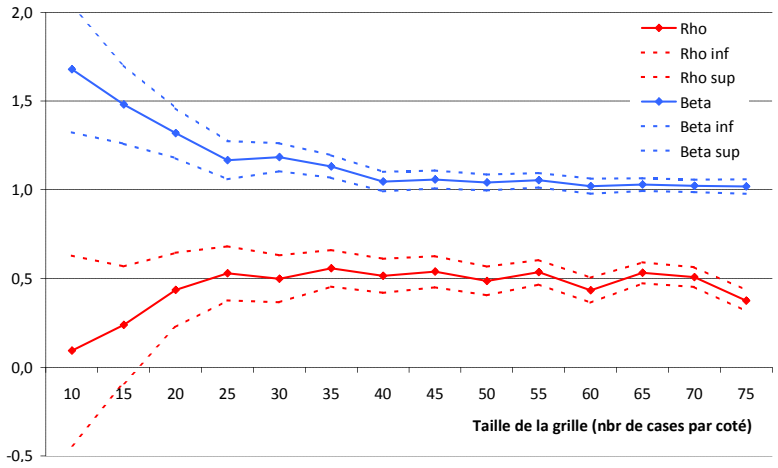
Afin d'évaluer cette solution, nous nous plaçons dans un cas simple où les effets simulés sont homogènes et uniformes, de sorte à éviter tout risque de biais d'agrégation. Nous simulons 6000 points selon une loi uniforme sur un espace carré et leur affectons, comme dans la Partie 1, des valeurs de X 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 6. Puis cet espace carré est découpé selon une grille de taille $G \times G$ 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 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 $\hat{\sigma}_\rho = 0,07$. Ce paramètre est significativement positif et l'estimation inclut $\rho = 0,5$ dans son intervalle de confiance. Dans ce cadre, l'agrégation de données sur des parcelles limiterait la perte d'interactions spatiales. La Figure 7 montre que les estimateurs $\hat{\rho}$ et $\hat{\beta}$ sont précisément estimés et proches de leurs vraies valeurs dès lors que la grille sur laquelle les données sont agrégées est relativement fine. En effet, le cadre de l'estimation se rapproche alors de la structure spatiale des données ponctuelles.

2.1.2 Application à un échantillon

Cette procédure est répliquée sur des données échantillonnées par SAS. 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 moins précises des moyennes locales. 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 4 présente les résultats de cette procédure pour différentes tailles d'échantillon et diverses grilles spatiales.

FIGURE 7 – Paramètres estimés selon la finesse de la grille



Note : Les courbes en trait plein rouge et bleue représentent respectivement les paramètres estimés $\hat{\rho}$ et $\hat{\beta}$ en fonction de la finesse de la grille. Les courbes en pointillés délimitent les intervalles de confiance.

TABLE 4 – Estimation d'un SAR sur échantillons agrégés par parcelle

	$\hat{\rho}$				$\hat{\beta}$			
n \ G	10	30	50	60	10	30	50	60
100	0.49***	0.49***	0.48***	0.48***	1.02***	1.01***	1.03***	1.03***
	(0.07)	(0.06)	(0.07)	(0.07)	(0.13)	(0.12)	(0.13)	(0.13)
200	0.48***	0.50***	0.50***	0.49***	1.02***	1.00***	1.01***	1.01***
	(0.06)	(0.05)	(0.05)	(0.05)	(0.13)	(0.08)	(0.09)	(0.10)
500	-0.09	0.49***	0.48***	0.49***	1.04***	1.02***	1.03***	1.02***
	(0.70)	(0.03)	(0.03)	(0.03)	(0.12)	(0.06)	(0.06)	(0.06)
1000	-0.98	0.49***	0.49***	0.49***	1.05***	1.02***	1.03***	1.02***
	(0.16)	(0.02)	(0.02)	(0.02)	(0.12)	(0.05)	(0.04)	(0.04)

*Note : Chaque ligne correspond à une taille d'échantillon n tiré parmi les 6000 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%, ** à 5% et * à 10%.*

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

2.2 Imputer les données manquantes

2.2.1 Problématique

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 des effets de la taille de l'échantillon et de la répartition des unités qui le composent 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.

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 (cf. Charreaux et al., 2016 pour une discussion de ce point). Dans le cas de modèles spatiaux, ce problème est exacerbé puisque 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.

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

2.2.2 Quelques méthodes d'imputation

Cette partie recense quelques méthodes classiques d'imputation. Le lecteur intéressé pourra se reporter à des livres de théorie des sondages (Ardilly, 1994 ; 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 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 $Y = \beta X + \epsilon$. $\hat{\beta}$ 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 estime 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, en limitant le nombre de fois où un même individu peut être donneur, ou en réalisant le hot deck de façon séquentielle.

2.2.3 Application à un échantillon

Nous illustrons les méthodes proposées à partir d'un exemple simple. Nous simulons la position géographique de $N = 1000$ points auxquels nous assignons des variables X et Y suivant une structure de SAR avec $\beta = 1$ et $\rho = 0.5$. Nous réalisons ensuite le tirage d'échantillons par sondage aléatoire simple pour différentes tailles n. Pour chacun des échantillons, les $N - n$ unités non tirées sont imputées par une des méthodes évoquées plus haut : imputation par le ratio de X, imputation par *hot deck* statistique (les voisins ont des valeurs proches de X) et imputation par *hot deck* géographique (les voisins sont spatialement proches).

TABLE 5 – Méthodes d'imputation

	Direct		**Ratio**		**Hot Deck Statistique**		**Hot Deck Géographique**	
n	ρ	β	ρ	β	ρ	β	ρ	β
100	0.06	1.14***	0.05***	1.13***	0.03	1.06***	0.19***	0.12**
	(0.06)	(0.14)	(0.01)	(0.12)	(0.03)	(0.13)	(0.05)	(0.05)
200	0.09*	1.10***	0.11***	1.11***	0.06**	1.08***	0.22***	0.22***
	(0.04)	(0.08)	(0.02)	(0.08)	(0.02)	(0.09)	(0.04)	(0.05)
500	0.19***	1.08***	0.25***	1.09***	0.19***	1.09***	0.31***	0.55***
	(0.02)	(0.04)	(0.02)	(0.05)	(0.02)	(0.05)	(0.03)	(0.04)

*Note : Le paramètre ρ du modèle SAR estimé par Monte Carlo pour un échantillon de taille 100 après imputation par le ratio est de 0.05, avec un écart-type empirique de 0.01. *** désigne une significativité à 1%, ** à 5% et * à 10%.*

La Table 5 compare les résultats de ces différentes méthodes avec une exploitation directe de l'échantillon. Le choix de la méthode a un impact non négligeable sur les résultats obtenus. L'imputation par le ratio semble bien fonctionner pour les deux paramètres même si elle sous-estime le paramètre ρ pour des petits échantillons. À l'inverse, la méthode de hot deck géographique donne de bons résultats sur le paramètre d'auto-corrélation mais implique un biais très fort sur le paramètre β. Enfin, la méthode de hot deck statistique semble donner des résultats similaires à l'exploitation directe de l'échantillon. Ces résultats illustrent ces méthodes sur un exemple très simple et montrent leur inaptitude à retrouver les paramètres initiaux du modèle.

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

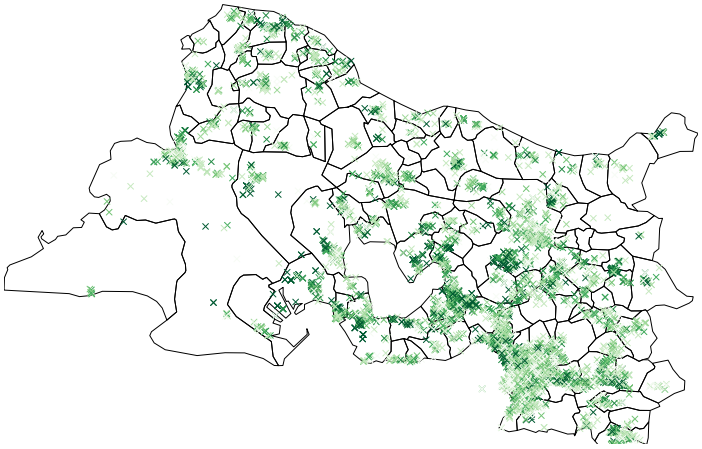
Afin d'illustrer les enjeux relatifs à l'estimation de modèles spatiaux sur données d'enquête, nous procédons dans cette section à l'estimation d'une fonction de production sur des établissements issus du répertoire SIRUS. L'approche spatiale permet de mesurer l'incidence des interactions entre les processus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

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 REpertoire 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 chacun d'entre eux. Pour cela, différentes données ont été utilisées : la référence cadastrale, la voie puis le centre de la commune lorsque l'information disponible est trop faible. 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 8 – Établissements industriels dans les Bouches-du-Rhône



Note : Répertoire SIRUS, 2013. Établissements du secteur de l'industrie, dans le département des Bouches-du-Rhône. La couleur est plus foncée lorsque le chiffre d'affaires est important.

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 connaissances...) ou négatives lorsqu'elles nuisent à la production (concurrence, pollution, embouteillages...).

Afin de mener à bien cette analyse, il faut tout d'abord se restreindre à un champ, un type de liens et un territoire cohérents, de sorte que les unités considérées entretiennent des relations entre elles mais pas (ou peu) avec l'extérieur. Dans ce cas, on pourra considérer que l'on dispose de données exhaustives relativement aux effets spatiaux à l'œuvre sur ce territoire. Le choix d'étudier l'industrie dans les Bouches-du-Rhône est didactique, mais non dénué de sens. La présence du

port de Fos-sur-mer, l'axe de la vallée du Rhône et les nœuds routiers en direction de Toulouse et de l'Italie en font un territoire d'intérêt (la Figure 8 illustre clairement l'implantation des industrie selon les réseaux de transports). Bien sûr, cette estimation ne permettra pas de prendre en compte les relations commerciales nationales voire internationales, mais sera exhaustive concernant des relations locales.

La production Y_i d'une entreprise i peut s'exprimer selon une fonction de type Cobb-Douglas : $Y_i = AL_i^{\beta_L} K_i^{\beta_K}$, qui dépend 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, et peuvent également s'interpréter comme les élasticités de la production au travail et au capital respectivement. 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 également 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 et Y_j le niveau de production d'une unité j voisine de i. En composant par la fonction logarithme, l'équation estimée peut se réécrire :

$$\log(Y_i) = \beta_0 + \rho \sum_{j \in v_i} \omega_{ij} \log(Y_j) + \beta_L \log(L_i) + \beta_K \log(K_i) + \varepsilon_i \tag{2}$$

ou encore :

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

où les notations avec une tilde désignent la version en log de ces variables et $\mathcal{W}$ est la matrice de pondération spatiale, telle que $\mathcal{W}_{i,j} = \omega_{ij}$. On a ainsi la forme d'une équation caractéristique d'un modèle spatial autorégressif (SAR). Le paramètre ρ, qui capte les complémentarités communes à toutes les unités, peut s'interpréter comme le paramètre de corrélation spatiale. Le paramètre ω_{ij} capte les complémentarités spécifiques, résultant de l'impact de l'activité de l'entreprise j sur 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 $\rho\omega_{ij}$ % via des effets directs. En effet, en différenciant l'équation 2 par rapport à $\log(Y_k)$, on a :

$$\frac{d\log(Y_i)}{d\log(Y_j)} = \underbrace{\rho\omega_{ij}}_{\text{Effet direct de j sur i}} + \underbrace{\rho \sum_{k \neq j} \omega_{ik} \frac{d\log(Y_k)}{d\log(Y_j)}}_{\text{Effet indirect via k}}$$

De la même manière, en sommant cette expression sur j, ρ apparaît comme l'élasticité directe de la production de l'entreprise i par rapport à la production de ses voisines :

$$\sum_j \frac{d\log(Y_i)}{d\log(Y_j)} = \rho + \rho \sum_j \sum_{k \neq j} \omega_{ik} \frac{d\log(Y_k)}{d\log(Y_j)}$$

3.3 Estimation

L'équation 2 est estimée sur 6 306 établissements géolocalisés dans les Bouches-du-Rhône, appartenant au secteur de l'industrie [6]. 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

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

que spatiales (comme en finance ou dans les communications). Relativement aux services, les chaînes de production dans l'industrie sont susceptibles d'être à plus petite échelle, ce qui nous permettrait de mieux capter les liens.

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 (cela crée une erreur de mesure, qui sera négligée par la suite). Enfin, l'effectif L_i est disponible au niveau de l'établissement dans SIRUS.

La Figure 8 représente la localisation de ces établissements. L'intensité du vert de ces croix matérialise la taille de leur chiffre d'affaires : plus le vert est foncé, 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 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 [7].

TABLE 6 – Estimation du modèle SAR : ensemble des établissements

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

Note : Répertoire SIRUS, 2015. 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. Les paramètres estimés sont tous significatifs au seuil de 1 %.

La Table 6 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 point, le chiffre d'affaires de i augmente de 0,05 points par effet direct.

3.4 Estimations spatiales sur des échantillons

3.4.1 Plans de sondage

De même que dans la Section 1, nous répliquons l'estimation du modèle 2 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 fréquemment employés dans le cadre d'études identifiant l'effet de l'effectif et du patrimoine sur le chiffre d'affaires.

7. 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.

La stratification est effectuée selon la variable d'effectif, sous l'hypothèse d'une corrélation entre effectif et chiffre d'affaires. La Table 7 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.

TABLE 7 – Définition des strates

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

Note : Répertoire SIRUS, 2013. 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. La strate 3 correspond aux établissements dont l'effectif salarié est entre 10 et 99. Cela correspond à 770 établissements.

3.4.2 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 : $n \in \{250, 500, 1000, 2000\}$.

TABLE 8 – Modèle 2 estimé sur échantillon aléatoire (SAS)

	Echantillon aléatoire (SAS)			Echantillon stratifié		
n	$\hat{\rho}$	$\hat{\beta_L}$	$\hat{\beta_K}$	$\hat{\rho}$	$\hat{\beta_L}$	$\hat{\beta_K}$
250	0.01	0.55***	0.77***	0.02*	0.31***	0.81***
	(0.02)	(0.10)	(0.08)	(0.01)	(0.07)	(0.06)
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)
1000	0.02**	0.54***	0.77***	0.02***	0.41***	0.79***
	(0.01)	(0.05)	(0.04)	(0.01)	(0.04)	(0.03)
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)

*Note : Répertoire SIRUS, 2015. Établissements du secteur de l'industrie, dans le département des Bouches-du-Rhône. Régression non pondérée. Le paramètre ρ du modèle SAR estimé par Monte Carlo pour un échantillon stratifié de 250 établissements est de 0.015, avec un écart-type empirique de 0.009. *** désigne une significativité à 1%, ** à 5% et * à 10%.*

La Table 8 présente les paramètres du modèle SAR estimés à partir de 1000 tirages d'échantillon par sondage aléatoire simple (*à gauche*) et par sondage stratifié (*à droite*). Dans le cas d'un SAS, de même que dans la Section 1, les paramètres classiques de régression β_L et β_K, sont correctement estimés. En revanche, l'estimateur du paramètre de corrélation spatiale $\hat{\rho}$ n'est

significatif que pour un échantillon de taille supérieure à 1000 et reste toujours inférieur à la valeur qu'il prend sur données exhaustives.

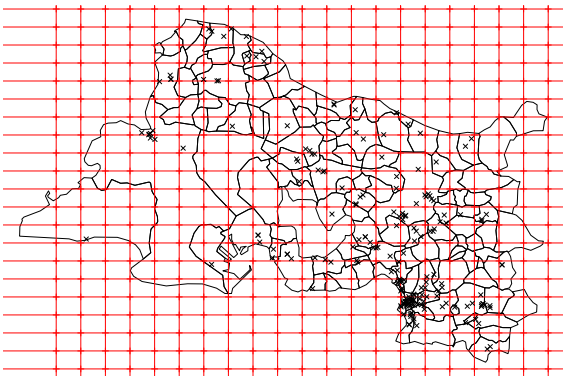
Le plan de sondage stratifié appliqué aux données non-repondérées biaise les estimateurs classiques $\hat{\beta_L}$ et $\hat{\beta_K}$, ce qui est similaire au résultat de Davezies et D'Haultfoeuille, 2009 dans la cas d'une régression linéaire non-pondérée omettant, parmi les explicatives, certaines variables incluses dans le plan de sondage. 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 [8]. En l'état actuel, 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 cet article.

3.5 Estimation sur données agrégées

Afin de contourner le problème de données manquantes, nous évaluons dans un premier temps la possibilité de passer à une échelle plus large par aggrégation 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 9).

Figure 9 – Découpage des Bouches-du-Rhône selon une grille 20×20



Note : Répertoire SIRUS, 2013. Établissements du secteur de l'industrie, dans le département des Bouches-du-Rhône.

A 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. Le Tableau 9 présente le paramètre $\hat{\rho}$ estimé pour différentes tailles d'échantillon et diverses tailles de grille.

8. 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 $\mathcal{W}$; cela pourrait également induire de l'endogénéité supplémentaire, liée à la structure de l'échantillon.

TABLE 9 – Paramètre $\hat{\rho}$ estimé sur données agrégées

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)

*Note : Répertoire SIRUS, 2015. Établissements du secteur de l'industrie, dans le département des Bouches-du-Rhône. Estimations 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 $\hat{\rho}$. Le paramètre estimé $\hat{\rho}$ du modèle SAR estimé par Monte Carlo pour un échantillon de 200 établissements sur une grille 30 x 30 est de 0.023, avec un écart-type empirique de 0.023. *** désigne une significativité à 1%, ** à 5% et * à 10%.*

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. Dans le cas présent, le paramètre d'autocorrélation spatiale est toujours sous-estimé et n'est jamais significatif. Cela pourrait résulter d'une forte concentration de l'industrie des Bouches-du-Rhône, les distances intra-cellule n'étant par définition pas prises en compte par cette méthode, ou encore à une hétérogénéité de l'effet sur le département. 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.

3.6 Imputation des données manquantes

3.6.1 Mise en oeuvre

La seconde approche, évoquée en Section 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, les voisins d'une unité étant 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 oeuvre 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. 1000 échantillons de taille n sont tirés selon un SAS, puis le processus d'imputation assigne des valeurs de la production Y aux $N - n$ établissements non échantillonnés. Les résultats obtenus sont présentées dans la Table 10. Pour rappel, les résultats sur la population entière sont en Table 8.

TABLE 10 – Méthodes d'imputation

	Ratio			Hot Deck Statistique			Hot Deck Géographique		
n	$\hat{\rho}$	$\hat{\beta_L}$	$\hat{\beta_K}$	$\hat{\rho}$	$\hat{\beta_L}$	$\hat{\beta_K}$	$\hat{\rho}$	$\hat{\beta_L}$	$\hat{\beta_K}$
250	0.00	0.56***	0.77***	0.04***	0.66***	0.65***	0.42***	0.03	0.10***
	(0.00)	(0.11)	(0.08)	(0.01)	(0.08)	(0.06)	(0.05)	(0.03)	(0.02)
500	0.00	0.55***	0.77***	0.04***	0.61***	0.70***	0.41***	0.06*	0.15***
	(0.00)	(0.08)	(0.06)	(0.01)	(0.06)	(0.04)	(0.04)	(0.03)	(0.02)
1000	0.01**	0.55***	0.77***	0.04***	0.58***	0.73***	0.39***	0.12***	0.22***
	(0.00)	(0.05)	(0.04)	(0.01)	(0.04)	(0.03)	(0.04)	(0.04)	(0.02)
2000	0.02***	0.54***	0.77***	0.04***	0.56***	0.75***	0.33***	0.20***	0.34***
	(0.00)	(0.03)	(0.02)	(0.01)	(0.03)	(0.02)	(0.02)	(0.03)	(0.02)

*Note : Répertoire SIRUS, 2015. Établissements du secteur de l'industrie, dans le département des Bouches-du-Rhône. Le paramètre $\hat{\rho}$ du modèle SAR estimé par Monte Carlo pour un échantillon de 1000 établissements après imputation par le ratio est de 0.008, avec un écart-type empirique de 0.003. *** désigne une significativité à 1%, ** à 5% et * à 10%.*

3.6.2 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 encore plus faible que dans le cas du sondage aléatoire simple exploité directement (cf Table 10). 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 6 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. Cela expliquerait pourquoi la méthode fonctionne dans notre application alors qu'elle était décevante sur l'exemple très simplifié présenté en partie 2.2, dans lequel les variables explicatives X n'avaient pas de structure spatiale propre. 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 $\hat{\beta}_L$ et $\hat{\beta}_K$, 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 surestime fortement l'effet spatial ($\hat{\rho}$ très supérieur à la vraie valeur ρ), au détriment des autres variables du modèle. 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 13 en Annexe 5.2 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.

4 Conclusion

Cet article 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 » résultant de la restriction à un sous-ensemble localement complet d'un territoire et *(ii)* un effet de répartition spatiale des observations, lié à l'omission aléatoire au sein d'un territoire d'unités spatialement corrélées avec les unités observées. Ces deux effets tendent à sous-estimer l'ampleur de la corrélation spatiale, plus particulièrement dans le cas d'un sondage aléatoire simple et lorsque l'échantillon est plus petit.

Les études empiriques passent généralement outre cet écueil, *(i)* soit en ignorant les observations manquantes, *(ii)* soit en agrégeant les données à une échelle plus large, *(iii)* soit 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 article 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.

Références

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

Arbia, Giuseppe, Giuseppe Espa et Diego Giuliani (2016). « Dirty spatial econometrics ». In : *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 (2017). « Spatial panel-data models using Stata ». In : *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 ». In : *Social Networks* 9.

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 ». In :

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

Cliff, A.D. et J.K. Ord (1972). *Spatial autocorrelation*. Pion, London.

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 ». In : *Journal of the royal statistical society* 39.1, p. 1–38.

Egger, Peter et Michael Pfaffermayr (2006). « Spatial convergence* ». In : *Papers in Regional Science* 85.2, p. 199–215.

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

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

Gile, Krista J. et Mark S. Handcock (2010). « RESPONDENT-DRIVEN SAMPLING : AN ASSESSMENT OF CURRENT METHODOLOGY ». In : *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 ». In : *Toulouse School of Economics Working Paper* 13, p. 452.

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

Huisman, Mark (2014). « Imputation of missing network data ». In : *Encyclopedia of Social Network Analysis and Mining*. 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 ». In : *Technometrics* 22.3, p. 389–395.

Kelejian, H.H. et I.R. Prusha (2010). « Spatial models with spatially lagged dependent variables and incomplete data ». In : *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 ». In : *Statistical Methodology* 7.3, p. 366–384.

Kossinets, Gueorgi (2006). « Effects of missing data in social networks ». In : *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 ». In : *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 ». In : *The journal of real estate finance and economics* 29.2, p. 233–254.

Little, Roderick J. A. (1988). « Missing-Data Adjustments in Large Surveys ». In : *Journal of Business & 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 ». In : *The Econometrics Journal*.

López-Bazo, Enrique, Esther Vayá et Manuel Artís (2004). « Regional Externalities And Growth : Evidence From European Regions* ». In : *Journal of Regional Science* 44.1, p. 43–73.

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

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

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

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

Stork, Diana et William D. Richards (1992). « Nonrespondents in Communication Network Studies ». In : *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 (2013). « Estimation of spatial autoregressive models with randomly missing data in the dependent variable ». In : *The Econometrics Journal* 16.1, p. 73–102.

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

5 Annexe

5.1 Choix du modèle et de la matrice de voisinage

Les tables suivantes 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.

TABLE 11 – Modèle SAR - Estimation par Monte Carlo

	$\hat{\rho}$			$\hat{\beta}$		
$\mathcal{M}$ / n	2 voisins	5 voisins	Distance	2 voisins	5 voisins	Distance
50	0.02	0.00	0.04	1.11***	1.05***	1.05***
	(0.11)	(0.17)	(0.04)	(0.12)	(0.10)	(0.13)
100	0.06	0.07	0.06*	1.11***	1.06***	1.05***
	(0.08)	(0.11)	(0.03)	(0.08)	(0.07)	(0.09)
150	0.10*	0.12	0.07**	1.11***	1.05***	1.05***
	(0.06)	(0.09)	(0.03)	(0.06)	(0.05)	(0.07)
250	0.15***	0.19**	0.10**	1.11***	1.05***	1.05***
	(0.05)	(0.07)	(0.03)	(0.05)	(0.04)	(0.05)

*Note : Le paramètre $\hat{\rho}$ du modèle SAR estimé par Monte Carlo pour un échantillon de taille 100 avec une matrice de voisinage basé sur 5 voisins est de 0.069, avec un écart-type empirique de 0.111. *** désigne une significativité à 1%, ** à 5% et * à 10%.*

TABLE 12 – Modèle SEM - Estimation par Monte Carlo

	$\hat{\lambda}$			$\hat{\beta}$		
$\mathcal{M}$ / n	2 voisins	5 voisins	Distance	2 voisins	5 voisins	Distance
50	−0.03	−0.11	0.01	1.00***	1.00***	1.00***
	(0.17)	(0.29)	(0.19)	(0.12)	(0.11)	(0.11)
100	0.01	−0.03	0.02	1.00***	1.00***	1.00***
	(0.11)	(0.18)	(0.12)	(0.08)	(0.08)	(0.08)
150	0.02	0.00	0.03	1.00***	1.00***	100***
	(0.09)	(0.14)	(0.10)	(0.07)	(0.06)	(0.06)
250	0.05	0.04	0.05	1.00***	1.00***	1.00***
	(0.07)	(0.11)	(0.08)	(0.05)	(0.05)	(0.05)

*Note : Le paramètre $\hat{\beta}$ du modèle SEM estimé par Monte Carlo pour un échantillon de taille 100 avec une matrice de voisinage basé sur 5 voisins est de 1.004, avec un écart-type empirique de 0.078. *** désigne une significativité à 1%, ** à 5% et * à 10%.*

5.2 Imputation par hot deck géographique stratifié

La Table 13 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 7) que l'établissement ayant une valeur manquante.

TABLE 13 – Une autre méthode d'imputation

	Hot Deck Géographique Stratifié		
n	$\hat{\rho}$	$\hat{\beta_L}$	$\hat{\beta_K}$
250	0.14**	1.22***	0.03*
	(0.04)	(0.10)	(0.03)
500	0.15***	1.19***	0.07**
	(0.03)	(0.08)	(0.03)
1000	0.16***	1.12***	0.15***
	(0.03)	(0.06)	(0.03)
2000	0.15***	0.54***	0.28***
	(0.02)	(0.05)	(0.03)

*Note : Répertoire SIRUS, 2015. Établissements du secteur de l'industrie, dans le département des Bouches-du-Rhône. Le paramètre $\hat{\rho}$ du modèle SAR estimé par Monte Carlo pour un échantillon de 500 établissements après imputation par hot deck géographique stratifié est de 0.148, avec un écart-type empirique de 0.031. *** désigne une significativité à 1%, ** à 5% et * à 10%.*