JMS Journées de méthodologie statistique de l'Insee 2018

ÉCONOMÉTRIE SPATIALE SUR DONNÉES D'ENQUÊTE

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

(*) Insee, Direction des études et des synthèses économiques
(**) Insee, Direction de la méthodologie et de la coordination statistique et internationale

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

Mots-clés : économétrie spatiale, autocorrélation spatiale, données d'enquêtes, imputation.

Résumé

L'économétrie spatiale requiert des données exhaustives sur un territoire, ce qui interdit en principe l'utilisation de données d'enquête. En effet, alors que l'économétrie classique repose sur une hypothèse d'indépendance mutuelle des observations, et que donc 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.

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) X (n-m)* au lieu de la vraie matrice de pondération de taille *n X n*. Ce changement de taille affecte en soi l'estimation du paramètre de corrélation spatiale. 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 *WY*) 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.

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.

Une illustration de cette problématique est proposée par l'estimation d'externalités de production entre les industries du département français des Bouches-du-Rhône. En effet, 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…).

Les caractéristiques des établissements industriels des Bouches-du-Rhône sont connues via le répertoire Sirus, et un travail de géolocalisation a été fait en mobilisant les informations d'adressage présentes dans le répertoire. Les variables disponibles permettent ainsi d'estimer une fonction de type Cobb-Douglas avec présence d'externalités sur le lien entre production et capital et effectif. Des simulations de tirage d'échantillons permettent de conclure sur la possibilité ou non de détecter ces externalités à l'aide d'un sous-ensemble de la population d'intérêt.

Bibliographie

[1] Pinkse, Joris and Slade, Margaret E., « The Future of Spatial Econometrics », *Journal of Regional Science*, vol 50, n° 1, pp 103-117, 2010.
[2] Arbia, Giuseppe and Espa, Giuseppe and Giuliani, Diego, « Dirty spatial econometrics », The Annals of Regional Science, vol 56, n° 1, pp 177-189, 2016.
[3] Huisman, Mark , « Imputation of missing network data », Encyclopedia of Social Network Analysis and Mining, vol 2, pp 707-715, 2014.