Région et Développement

n° 62-2025

www.regionetdeveloppement.org

Econométrie des données spatiales : enjeux d'identification et perspectives

Nicolas DEBARSY[*]
Julie LE GALLO[**]

Résumé - Cet article propose un point d'étape sur l'économétrie des données spatiales en mettant l'accent sur les enjeux d'identification des interactions spatiales et les perspectives ouvertes par le développement des données massives géolocalisées. Ce champ de l'économétrie a pour objectif l'analyse des phénomènes où la proximité géographique, les interdépendances spatiales et les hétérogénéités territoriales jouent un rôle structurant. Si les modèles habituels, tels que le SAR ou le SDM, permettent de formaliser les interactions spatiales, leur mise en œuvre empirique soulève des difficultés d'identification majeures. Identifier rigoureusement ces interactions suppose de clarifier la nature des interactions, de traiter les problèmes d'endogénéité et de contrôler les sources d'hétérogénéité spatiale. L'article discute ensuite les défis que doit relever les méthodes de l'économétrie des données spatiales tant dans une perspective méthodologique structurelle que dans le cadre des modèles d'inférence causale, tout en intégrant les méthodes d'inférence développées dans d'autres branches de l'économétrie. Notamment, il souligne que l'essor du *big spatial data* et des algorithmes de *spatial machine learning* constituent une opportunité décisive pour dépasser certaines limites des approches traditionnelles.

Classification JEL
C21, R15, C55

Mots-clés
Econométrie des données spatiales
Identification
Interactions spatiales
Big spatial data

[*] Université de Lille, CNRS, IESEG School of Management, UMR 9221, Lille Économie Management, France ; nicolas.debarsy@cnrs.fr
[**] Institut Agro, INRAE, UMR CESAER, France ; julie.le-gallo@institut-agro.fr

INTRODUCTION

"I invoke the first law of geography: everything is related to everything else, but near things are more related than distant things." (Tobler, 1970). Cette citation constitue le point de départ de nombreuses recherches empiriques portant sur l'économétrie des données spatiales, et souligne l'importance de la prise en compte des *spillovers* spatiaux (sous la forme d'externalités ou d'interactions) et des effets de diffusion spatiale dans l'explication de comportements et/ou résultats. Au-delà de ces *spillovers*, les phénomènes économiques se caractérisent souvent par des structures géographiques distinctes, comme des organisations centre-périphérie ou des *clusters* locaux, qui renvoient à la présence d'hétérogénéité spatiale.

L'économétrie des données spatiales est un champ méthodologique qui vise à adapter et enrichir les outils économétriques classiques pour prendre en compte la dimension géographique des données. Il s'agit, d'une part, d'offrir un cadre pour modéliser explicitement les différents mécanismes sous-tendant les *spillovers* spatiaux, afin d'améliorer la compréhension de phénomènes économiques caractérisés par des interactions spatiales, comme les questions relatives à la concurrence fiscale, et/ou par des externalités territoriales, qu'il s'agisse de la dynamique immobilière, de la mobilité résidentielle ou encore des enjeux environnementaux. D'autre part, ces outils visent à intégrer les différentes formes d'hétérogénéité spatiale.

Le développement récent de ce champ ne peut être dissocié de l'évolution des données disponibles. L'essor de la géolocalisation et la généralisation d'informations à haute résolution spatiale – issues des registres administratifs, des capteurs, de l'imagerie satellitaire ou encore des traces numériques laissées par les individus – offrent aujourd'hui des données empiriques géolocalisées d'une ampleur sans précédent. Ce que l'on qualifie désormais de *big spatial data* ouvre la voie à une nouvelle génération de matériau empirique, combinant des bases massives, des résolutions fines et des outils de *machine learning*, mais posant aussi de nouveaux défis en matière de modélisation et d'identification.

Cet article ne vise pas à fournir une revue exhaustive de ce champ qui dispose déjà de nombreuses synthèses[1] mais il a pour objectif de mettre en évidence quelques axes qui pourraient constituer des développements importants du champ. La première section présente le modèle général de modélisation des effets spatiaux en coupe transversale et discute des enjeux d'identification des *spillovers géographiques*. La deuxième section suivante sélectionne des perspectives actuelles du champ : les pistes qui s'ouvrent pour les méthodes structurelles et causales d'identification, d'une part, et le dialogue avec l'émergence des données massives et les méthodes de *machine learning*, d'autre part. La dernière section conclut brièvement.

1. MODÈLES PRINCIPAUX ET LEURS ENJEUX

1.1. Modéliser les interactions spatiales

Un des premiers objectifs de l'économétrie des données spatiales est la volonté de formaliser et d'identifier les interactions spatiales entre les unités, qu'elles soient

[1] Parmi les contributions les plus récentes, les lecteurs pourront par exemple se référer à Bellefon et Loonis (2018) et les sections consacrées à l'analyse de données spatiales dans Fisher et Nijkamp (2021) et Nijkamp et al. (2025).

agrégées (régions, pays) ou microéconomiques (individus, entreprises). L'approche structurelle consiste à expliciter un modèle théorique, s'appuyant par exemple sur une fonction de production ou d'utilité, puis à en dériver un modèle économétrique dont les paramètres ont une interprétation économique.

Ainsi, dans la littérature sur la concurrence fiscale par exemple, des modèles tels que le *Spatial Autoregressive Model* (SAR), le *Spatial Durbin Model* (*SDM*) sont utilisés pour estimer l'existence et l'ampleur d'interactions stratégiques entre juridictions locales (Agrawal et al., 2022).

Manski (1993) propose une typologie en trois catégories pour classer les différents effets du voisinage :
- Les effets endogènes : quantifient la dépendance du comportement d'une unité aux comportements de ses voisins.
- Les effets contextuels : mesurent la dépendance aux caractéristiques des voisins.
- Les effets corrélés : captent la corrélation induite par des facteurs communs ou des chocs non observés.

L'enjeu est alors de distinguer les vrais effets d'influence des corrélations liées à des caractéristiques partagées ou des chocs collectifs.[2] Gibbons et al. (2015) proposent un modèle général qui englobe ces différents effets :

$$y_i = x_i\beta + \lambda \sum_{j=1}^{n} w_{ij}^{y}\, y_j + \sum_{j=1}^{n} w_{ij}^{x}\, x_j\gamma + \sum_{j=1}^{n} w_{ij}^{z}\, z_j\theta + \sum_{j=1}^{n} w_{ij}^{v}\, v_j\kappa + \varepsilon_i \qquad i = 1 \dots n$$

où y_i est la variable à expliquer pour l'unité i ; x_i désigne un vecteur regroupant les caractéristiques propres à l'unité ; la première somme correspond aux effets endogènes, la deuxième aux effets contextuels tandis que les deux dernières sommes captent l'effet des effets corrélés, qui se manifestent respectivement sous la forme de facteurs communs (observés) et de chocs non observés. $w_{ij}^{\kappa}, \kappa = y, x, z, v,$ est l'élément de la matrice de connectivité W^{κ},qui modélise le lien entre les observations i et j et qui peut varier selon le type d'effet considéré. On parlera de modèle *spatial* lorsque cette matrice de connectivité est basée sur une fonction de la proximité géographique (inverse de la distance, plus proche voisin, contiguïté, etc.). Finalement, ε_i représente le terme d'erreurs idiosyncratique.

Le SDM, qui impose l'absence d'effets corrélés dans le modèle général ci-dessus ainsi que l'égalité des différentes matrices d'interactions, est habituellement considéré comme la spécification la plus générale dans les papiers empiriques mobilisant les outils de l'économétrie des interactions spatiales. En effet, le SDM englobe les spécifications habituellement utilisées dans les études empiriques : SAR (si $\gamma = 0$), *le modèle avec caractéristiques des voisins* (SLX) (si $\lambda = 0$), et le modèle avec autocorrélation des erreurs (SEM) (si $\gamma = -\lambda\beta$). De plus, les papiers empiriques veulent également souvent identifier la spécification la plus pertinente pour modéliser le phénomène d'intérêt. Cette sélection, basée sur des critères statistiques et d'ajustement aux données, mobilise deux approches complémentaires. Dans l'approche *général-vers-spécifique*, le modèle SDM est initialement estimé et un

[2] Manski a également défini le *Reflection problem*, qui indique que même en l'absence d'effets corrélés, il est impossible de distinguer les effets endogènes des effets exogènes. Pour une discussion approfondie de ce problème, le lecteur peut consulter Bramoullé et al. (2020).

ensemble d'hypothèses sur les différents paramètres sont ensuite testées afin de vérifier s'il peut être simplifié. Dans l'approche *spécifique-vers-général*, un modèle linéaire sans aucun terme de *spillovers* constitue le point de départ et est progressivement enrichi en fonction des résultats de tests du multiplicateur de Lagrange ou du ratio de vraisemblance[3]. Enfin, après avoir sélectionné la « meilleure » spécification et obtenu une estimation des différents paramètres, la dernière étape classique consiste à calculer les effets marginaux (« direct, indirect, total ») de différentes variables explicatives, qui prennent en compte les effets de rétroaction impliqués par la structure du modèle et la matrice de connectivité utilisée (LeSage et Pace, 2009).

1.2. Les défis de l'identification

L'identification du paramètre des effets endogènes doit constituer un enjeu central lorsque l'objectif principal de l'analyse est d'estimer des interactions spatiales. Cependant, plusieurs sources de confusion rendent difficile la distinction entre véritables interactions et simples corrélations géographiques (Gibbons et Overman, 2012 ; Debarsy et Le Gallo, 2025).

La première difficulté tient au choix de la matrice de connectivité. L'avantage d'utiliser une fonction de la proximité géographique pour modéliser les liens entre observations est que l'espace physique est exogène et non manipulable. Cependant, en plus d'être discutable dans certains cas (notamment lorsque les observations sont mobiles), cet argument impose deux grandes limites à l'analyse quantitative. D'une part, les interactions réelles entre observations peuvent être de nature économique, sociale ou institutionnelle et utiliser une approximation géographique peut être réductrice et générer des problèmes de mauvaise spécification. D'autre part, d'un point de vue explicatif, sans justification théorique solide, l'utilisation d'une matrice basée sur des critères géographiques ne permet pas d'identifier les canaux économiques par lesquels les interactions s'opèrent.

Un second défi concerne le décalage possible entre l'unité statistique utilisée et l'unité de décision pertinente. Les données disponibles sont souvent agrégées à l'échelle des communes, régions ou quartiers, tandis que les comportements et interactions peuvent se jouer à d'autres niveaux. Ce décalage introduit des biais d'agrégation (*ecological fallacy*) ou des problèmes de changement d'échelle (*change of support*), qui compromettent l'interprétation claire et précise des paramètres estimés.

Troisièmement, il est nécessaire de distinguer soigneusement les *spillovers* spatiaux de l'hétérogénéité spatiale. La présence de facteurs communs (observés ou non), de régimes spatiaux ou de coefficients variables dans l'espace peut mener à des configurations similaires à la présence de véritables *spillovers* sans pour autant que ceux-ci soient présents. Ainsi, une mauvaise spécification ou l'omission de variables pertinentes peut conduire à confondre autocorrélation spatiale et hétérogénéité structurelle. En outre, des mécanismes comme le *spatial sorting*, c'est-à-dire la tendance des individus à se regrouper en fonction de caractéristiques observables ou non, accentue ce risque en rendant la matrice de connectivité spatiale potentiellement endogène. L'incapacité à contrôler l'hétérogénéité spatiale obère souvent l'identification des paramètres d'interaction et est peu prise en compte dans la littérature empirique.

[3] Nous développons dans les sections suivantes pourquoi cette approche ne permet pas de traiter correctement les défis de l'identification.

Enfin, la définition même du paramètre d'intérêt joue un rôle crucial. Dans certains cas, comme dans les modèles de concurrence fiscale, il s'agit d'identifier un véritable effet d'interaction spatiale, ce qui exige une modélisation structurelle et une construction rigoureuse de la matrice de connectivité. Dans d'autres, l'objectif est l'identification de relations économiques entre variables, qui peut justifier la prise en compte de spillovers et d'hétérogénéité spatiale. La confusion entre ces deux objectifs peut conduire à surinterpréter les coefficients spatiaux ou à leur attribuer une portée causale injustifiée.

2. APPROFONDIR L'IDENTIFICATION : DE LA THÉORIE AUX DONNÉES MASSIVES

2.1. Approches structurelles et approches réduites pour l'identification

Comme l'a montré la section précédente, la modélisation des *spillovers* spatiaux dans une spécification économétrique nécessite des choix conceptuels et méthodologiques importants. A ce titre, les approches itératives spécifique-au-général ou général-au-spécifiques présentées dans la section 1.2 réduisent l'étude des effets de *spillovers* spatiaux à un simple sous-produit issu d'une sélection de spécifications guidée par des critères statistiques et un ajustement aux données, au détriment d'une identification solidement ancrée dans des modèles économiques ou des scénarios de transmission plausibles.

Il convient donc de ne plus recourir à ces approches mais d'enraciner les modèles de *spillovers* spatiaux dans la théorie, qu'elle soit économique, sociologique, ou issue d'autres disciplines (écologie, etc.). Les fondements microéconomiques du comportement des agents économiques, tels que des fonctions de réaction de Nash, les modèles de réseaux sociaux économiques, peuvent par exemple être mobilisés pour expliciter les mécanismes d'interaction, de concurrence et de diffusion (voir notamment Agrawal et al., 2022, pour les modèles de fédéralisme fiscal). Ces modèles permettront par ailleurs de justifier et d'intégrer des matrices de connectivité potentiellement endogènes. L'analyse économétrique doit également modéliser très précisément l'hétérogénéité inobservée et justifier clairement les hypothèses faites en termes d'exogénéité des variables et des éventuels instruments utilisés.

A contrario, si l'objectif est d'identifier des relations causales entre variables géolocalisées, (impact du capital humain sur la croissance économique ou des effets d'une population supplémentaire sur les dépenses de fonctionnement d'une commune par exemple), à l'aide de méthodes d'inférence causale (méthodes de différences-en-différences, régressions sur discontinuité…), il convient d'éviter d'intégrer des termes renvoyant à des effets endogènes puisqu'ils ne peuvent être considérés comme des variables de pré-traitement. La littérature, notamment en bio-statistique, a développé des méthodes d'évaluation causale sous interférence (c'est-à-dire en présence de *spillovers*) qui pourront être mobilisées pour estimer les effets de traitements pertinents lorsque l'hypothèse d'indépendance entre les observations n'est pas satisfaite. Debarsy et Le Gallo (2025) rendent compte de cette littérature.

2.2. Les enjeux liés au big spatial data et au spatial machine learning

Le développement récent des données massives géolocalisées ou à haute résolution bouleverse profondément l'analyse et ouvre des perspectives inédites. Ces *big spatial data* (Wójcik, 2020) ne se distinguent pas seulement par leur volume,

leur variété ou leur vélocité : elles permettent de documenter les comportements individuels et collectifs dans l'espace avec une finesse sans précédent, rendant possible une modélisation plus précise des interactions et des dynamiques spatiales. Dès lors, la mobilisation d'algorithmes de *spatial machine learning* permettront de tirer parti de ces données massives et d'enrichir la modélisation économétrique des données spatiales.

Dans le champ de l'inférence causale, l'essor des méthodes de *machine learning* a déjà produit des avancées considérables (Athey, 2019 ; Brand et al., 2023) : ces outils peuvent être mobilisés pour estimer des effets de traitement hétérogènes, construire des contrefactuels plausibles, ou encore améliorer la robustesse des estimations dans des environnements complexes et riches en données. L'intégration des algorithmes supervisés ou non supervisés à l'évaluation causale a permis de dépasser certaines limites des approches économétriques traditionnelles, en articulant identification et flexibilité prédictive.

Face à ces enjeux, l'économétrie des données spatiales est appelée à dialoguer plus étroitement avec ces développements, qu'il s'agisse d'identifier les interactions spatiales ou d'analyser des relations économiques mobilisant des données géo-localisées. L'intégration de techniques issues de la science des données – algorithmes distribués, méthodes de sélection automatisée des structures de voisinage, modèles semi-paramétriques incorporant des effets non linéaires locaux – ouvre ainsi la voie à une nouvelle génération d'outils. Les approches de *spatial machine learning* (Credit, 2024), en particulier, offrent un potentiel considérable pour détecter des clusters, identifier des hétérogénéités locales, ou encore estimer des interactions spatiales complexes qui échappent aux spécifications usuelles.

CONCLUSION

Ce court article a pour objectif de fournir un point d'étape sur les enjeux d'identification des interactions spatiales entre agents, territoires et politiques à l'heure des données massives. Si les modèles structurels comme le SDM offrent une formalisation puissante des interactions, ils se heurtent à des défis complexes d'identification, de choix de la matrice de connectivité spatiale et d'intégration de l'hétérogénéité spatiale. Par ailleurs, l'émergence du *big spatial data* et des algorithmes de *spatial machine learning* constituent une opportunité unique d'enrichir à la fois les méthodes et les champs d'application.

REFERENCES

Agrawal D.R., Hoyt W.H., Wilson J.D., 2022, Local Policy Choice: Theory and Empirics, *Journal of Economic Literature*, 60, 4, 1378-1455.

Athey S. (2019) The impact of machine learning on economics, in Agrawal A., Gans J. and Goldfarb A. (Eds.), *The Economics of Artificial Intelligence*, NBER.

Bellefon M.-P., Loonis V., 2018, *Manuel d'analyse spatiale*, n°131, INSEE Méthodes.

Bramoullé Y., Djebarri H., Fortin B. (2020), Peer effects in networks: a survey, *Annual Review of Economics, 12(7),*

Brand, J.E, Xiang Z., Yu X., 2023, Recent Developments in Causal Inference and Machine Learning**,** Annual Review of Sociology**.**

Credit K., 2024, Introduction to the Special Issus on Spatial Machine Learning**,** *Journal of Geographical Systems*, 26, 4, 451-460**.**

Debarsy N., Le Gallo J., 2025, Identification of Spatial Spillovers: Do's and Don'ts, *Journal of Economic Surveys*, 1–22.

Fisher M.M., Nijkamp P. (Eds), 2021, *Handbook of Regional Science*, Springer-Verlag, Berlin.

Gibbons S., Overman H.G., 2012, Mostly Pointless Spatial Econometrics?, *Journal of Regional Science*, 52, 2, 172-191.

Gibbons S., Overman H.G., Patacchini E., 2015, "Spatial Methods", *Handbook of Regional and Urban Economics*, Volume 5, Elsevier, 115-168.

LeSage J., Pace K., 2009, *Introduction to Spatial Econometrics*, Chapman & Hall/CRC.

Manski C.F., 1993, Identification of Endogenous Social Effects: The Reflection Problem, *The Review of Economic Studies*, 60, 3, 531-542

Nijkamp P., Kourtit K., Haynes K., Elburz Z. (Eds), 2025, *Thematic Encyclopedia of Regional Science*, Edward Elgar.

Tobler W., 1970, A computer movie simulating urban growth in the Detroit region, *Economic Geography*, 46, 2, 234-240.

Wójcik P., 2020, Spatial Big Data, in Kopczewska K. (Ed.), *Applied Spatial Statistics and Econometrics*, Routledge.

Spatial data econometrics: identification issues and perspectives

Abstract **-** This article presents a reflection on the field of spatial data econometrics, emphasizing identification issues of spatial interactions and the perspectives opened by the development of geolocated big data. This branch of econometrics enables the analysis of phenomena in which geographic proximity, neighborhood interdependence, and territorial heterogeneity are structuring factors. While standard models — such as SAR and SDM — allow for the formalization of spatial interactions, their empirical implementation raises major identification challenges. Rigorous identification of spatial effects requires clarifying the nature of interactions, addressing endogeneity concerns, and disentangling spatial dependence from spatial heterogeneity. The paper then discusses the challenges that spatial interactions econometrics methods must address, both within a structural methodological approach and in causal inference models, while also considering inference methods developed in other branches of econometrics. Finally, it highlights that the rise of big spatial data and spatial machine learning algorithms represents a decisive opportunity to overcome certain limitations in current modeling.

Key-words
Spatial data econometrics
Identification
Spatial interactions
Big spatial data