Méthodologie

Justification de l'économétrie spatiale

L'économétrie spatiale est justifiée par des raisons économiques et économétriques. D'un point de vue économique, la proximité spatiale joue un rôle clé dans les décisions des agents économiques. Les entreprises, par exemple, ajustent leurs stratégies en fonction de la concurrence locale, tandis que la diffusion des innovations et les effets d'agglomération influencent la productivité régionale. Les externalités spatiales, telles que l'effet de pair et les interactions entre industries voisines, ont également un impact direct sur les marchés. Ainsi, la prise en compte de la dimension spatiale est essentielle pour comprendre les dynamiques économiques locales et globales.

Sur le plan économétrique, l'omission des effets spatiaux peut introduire des biais dans les estimations, rendant les modèles classiques inefficaces. L'autocorrélation spatiale des résidus (tout comme l'autocorrélation temporelle des résidus) est un problème récurrent, pouvant fausser l'inférence statistique si elle n'est pas correctement modélisée. De plus, l'hypothèse d'indépendance des observations, souvent supposée dans les modèles classiques, est rarement vérifiée lorsque des interactions spatiales existent. En intégrant des structures de dépendance spatiale, les modèles économétriques spatiaux permettent d'améliorer la précision des estimations et de mieux comprendre les relations entre unités géographiques, évitant ainsi les erreurs d'interprétation liées à des phénomènes locaux ou régionaux.

Concepts fondamentaux en statistique spatiale

Autocorrélation et hétérogénéité spatiales

L'autocorrélation spatiale désigne la dépendance statistique entre des observations géographiquement proches. En d'autres termes, les valeurs prises par une variable en un lieu donné sont influencées par les valeurs observées dans les localisations voisines. Cette dépendance peut être positive, lorsque des valeurs similaires se regroupent, ou négative, lorsqu'une valeur élevée en un point est associée à une valeur faible dans les zones environnantes.

L'hétérogénéité spatiale, quant à elle, fait référence à la variabilité des relations économiques en fonction de la localisation. Une même variable explicative peut avoir des effets différents selon les régions. Cette non-stationnarité spatiale implique qu'il est essentiel d'examiner l'existence de régimes spatiaux distincts et d'adapter les modèles en conséquence.

Matrice des poids spatiaux

Pour quantifier la proximité spatiale entre unités géographiques, on utilise une matrice de poids spatiaux notée W. Cette matrice représente les relations de voisinage et permet d'introduire la structure spatiale dans les modèles économétriques. Il existe plusieurs méthodes pour la définir :

- **Matrice de contiguïté** : Deux régions sont considérées comme voisines si elles partagent une frontière commune.
- **Matrice de distance** : La pondération est inversement proportionnelle à la distance entre deux régions.
- **Matrice des K plus proches voisins** : Chaque observation est associée aux K unités les plus proches.

Une matrice de contiguïté binaire est définie comme suit :

$$W_{ij} = \begin{cases} 1, & \text{si } i \text{ et } j \text{ sont voisins} \\ 0, & \text{sinon} \end{cases}$$

Indices de corrélation spatiale

L'un des indicateurs les plus couramment utilisés est l'indice de Moran, défini par :

$$I = \frac{N}{\sum_i \sum_j W_{ij}} \times \frac{\sum_i \sum_j W_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_i (y_i - \bar{y})^2}$$

où y_i est la valeur de la variable d'intérêt en un point i, $\bar{y}$ est la moyenne de cette variable et W_{ij} représente l'élément (i, j) de la matrice de poids.

D'autres indices existent, comme la **statistique de Geary**, qui est moins sensible aux valeurs extrêmes, et les **indicateurs locaux d'autocorrélation spatiale (LISA)**, qui permettent d'identifier des clusters spatiaux spécifiques.

Modélisation en économétrie spatiale

Voici un rappel des différents éléments utilisés dans l'ensemble des modèles d'économétrie spatiale :

- Y : Il s'agit du vecteur des observations de la variable dépendante, c'est-à-dire la variable que l'on cherche à expliquer (par exemple, le taux de chômage, les prix immobiliers, etc.).
- X : C'est la matrice des variables explicatives ou indépendantes. Elle regroupe toutes les caractéristiques observées qui sont supposées influencer $\mathbf{Y}$ (comme des variables socio-économiques, démographiques ou structurelles).
- β : Ce vecteur de coefficients mesure l'effet direct des variables $\mathbf{X}$ sur la variable dépendante $\mathbf{Y}$. Chaque coefficient indique l'impact d'une unité de variation dans la variable correspondante sur $\mathbf{Y}$, en l'absence d'effets spatiaux.
- W : La matrice des poids spatiaux définit la structure de voisinage entre les unités géographiques. Chaque élément W_{ij} quantifie l'influence ou la proximité de l'unité j par rapport à l'unité i. Le choix de cette matrice (par contiguïté, distance, ou K plus proches voisins) est crucial car il détermine la manière dont l'information spatiale est intégrée dans le modèle.
- WY Le terme de décalage spatial de $\mathbf{Y}$, obtenu par le produit de la matrice $\mathbf{W}$ par le vecteur $\mathbf{Y}$. Il représente l'influence moyenne pondérée des valeurs de $\mathbf{Y}$ dans les zones voisines et permet de capturer la dépendance spatiale directe de la variable dépendante.
- WX : Il s'agit du terme de décalage spatial des variables explicatives. Concrètement, il représente une version pondérée des variables $\mathbf{X}$ dans les zones voisines, où les pondérations sont définies par la matrice $\mathbf{W}$. Ce terme permet de mesurer l'effet indirect (ou spillover) des caractéristiques des voisins sur $\mathbf{Y}$.
- ε : C'est le terme d'erreur classique, qui capture les influences non observées ou aléatoires sur $\mathbf{Y}$. Il est généralement supposé être indépendant et identiquement distribué (iid).
- ρ : Utilisé dans les modèles qui intègrent directement l'effet des valeurs voisines de $\mathbf{Y}$ (comme dans les modèles SAR et SDM). Ce paramètre mesure la force de l'interaction entre la valeur de $\mathbf{Y}$ d'une unité et les valeurs de $\mathbf{Y}$ des unités voisines. Un ρ positif indique une autocorrélation positive (les zones avec des valeurs élevées de $\mathbf{Y}$ tendent à être entourées de zones à valeurs élevées, et inversement).
- λ : Spécifique au modèle SEM (Spatial Error Model), ce paramètre quantifie la corrélation spatiale présente dans le terme d'erreur. Il mesure l'influence des erreurs des unités voisines sur l'erreur de l'unité considérée, suggérant que des facteurs non observés présentent une structure spatiale.
- θ : Ce vecteur de coefficients est associé au terme $\mathbf{WX}$ et apparaît dans les modèles SDM et SLX. Il mesure l'effet des variables explicatives des zones voisines sur la variable dépendante $\mathbf{Y}$, c'est-à-dire l'impact indirect des caractéristiques locales via leur diffusion spatiale.

Modèles principaux

SAR (Spatial AutoRegressive Model) : Le modèle SAR introduit une dépendance spatiale directement sur la variable dépendante Y. L'idée est que la valeur de Y en un lieu donné dépend des valeurs observées dans les zones voisines. Mathématiquement, il s'écrit :

$$Y = \rho WY + X\beta + \varepsilon$$

Interprétation :

- Si $\rho > 0$, les valeurs de Y ont tendance à être similaires entre voisins (autocorrélation positive).
- Si $\rho < 0$, on observe un effet de dispersion, où les valeurs de Y sont opposées dans les zones voisines (autocorrélation négative).
- Si $\rho = 0$, il n'y a pas de dépendance spatiale, et le modèle classique de régression linéaire est suffisant.

SEM (Spatial Error Model) : Le modèle SEM est utilisé lorsque la dépendance spatiale affecte les erreurs du modèle plutôt que la variable dépendante elle-même. Il est défini par :

$$Y = X\beta + u, \quad u = \lambda Wu + \varepsilon$$

Interprétation :

- Contrairement au modèle SAR, le modèle SEM suppose que la dépendance spatiale est un effet de perturbation, provenant d'omissions de variables pertinentes qui suivent une structure spatiale.
- Il est utilisé lorsque la corrélation spatiale détectée dans un modèle classique provient d'erreurs spatialement autocorrélées, plutôt que d'une interaction directe entre observations.

SDM (Spatial Durbin Model) : Le modèle SDM est une extension du modèle SAR. Il prend en compte non seulement la dépendance de Y aux observations voisines, mais aussi l'effet des variables explicatives des régions voisines. Il est défini par :

$$Y = \rho WY + X\beta + WX\theta + \varepsilon$$

Interprétation :

- Si $\theta = 0$, le modèle SDM devient un SAR classique.
- Si $\rho = 0$, il devient un modèle SLX (voir ci-dessous).
- Il permet de tester si des variables exogènes influencent Y au-delà des frontières administratives.

SLX (Spatial Lag of X Model) : Le modèle SLX est plus simple que SAR et SDM, car il suppose que la variable dépendante Y n'est pas directement influencée par les valeurs voisines, mais uniquement par les variables explicatives des zones voisines. Il est écrit comme suit :

$$Y = X\beta + WX\theta + \varepsilon$$

où WX capture l'effet des variables explicatives des unités voisines.

Interprétation :

- Contrairement aux modèles SAR et SDM, il n'y a pas d'effet direct des valeurs voisines de Y.
- Il mesure uniquement l'effet de "spillover" (d'effet de débordement) des facteurs explicatifs.

Comparaison des modèles

Modèle	Dépendance spatiale sur Y	Effet des X des voisins	Effet des erreurs
SAR	Oui	Non	Non
SEM	Non	Non	Oui
SDM	Oui	Oui	Non
SLX	Non	Oui	Non

Table 1: Comparaison des modèles spatiaux

Limites et difficultés

Le point sur les limites et difficultés en économétrie spatiale se concentre sur plusieurs aspects. D'abord, la présence de données manquantes pose un défi majeur, car les observations spatiales ne sont pas toujours complètes. Pour y remédier, des méthodes comme l'imputation par krigeage (une technique d'interpolation géostatistique) ou l'estimation par maximum de vraisemblance adaptée aux données incomplètes sont utilisées.

Ensuite, d'autres difficultés incluent le choix de la matrice de poids spatiaux, qui influence fortement les résultats des modèles, et l'hétérogénéité spatiale, qui peut nécessiter des techniques avancées comme la régression géographiquement pondérée (GWR). Enfin, l'erreur écologique et le problème du MAUP (Modifiable Areal Unit Problem) compliquent l'interprétation des résultats, car les conclusions peuvent varier selon le niveau d'agrégation des données.