Econométrie spatiale : une introduction pratique

Jean-Michel Floch (INSEE), Ronan Le Saout (INSEE)

SMS, 5 Juillet 2016

Motivations

- Modélisation nécessaire pour dépasser l'analyse spatiale exploratoire (Floch 2012) ;
- De multiples modèles pour tenir compte des interactions spatiales et de la dépendance spatiale ;
- De nombreux cas d'interactions entre agents dans la théorie économique : des modèles aux applications plus larges que la géographie ;
- Des liens à effectuer avec la modélisation des phénomènes différenciés spatialement (hétérogénéité spatiale) ;
- Nécessité d'un guide détaillant pas à pas la mise en place de ces modèles et leur interprétation ;
- Détailler la programmation informatique à l'instar d'autres documents du DMS (régression quantile, pseudo-panels).

Quelques utilisations au sein de l'INSEE

- Le chômage et le marché de l'emploi : Blanc et Hild (2009), puis le PSAR de Toulouse (2014) ;
- La non-réponse à l'enquête emploi : Loonis (2012) ;
- La relation entre les prix immobiliers et les risques industriels : Grislain-Letrémy et Katossky (2013) ;
- Les migrations résidentielles : Guymarc (2015) ;
- Une prise en compte de l'espace mais des modèles différents. Pourquoi ?

Plan de la présentation

- Distinguer l'analyse exploratoire de la modélisation
- Choix de modèle et interprétation des résultats
- Limites et difficultés économétriques
- Illustration à partir des taux de chômage locaux

Distinguer l'analyse exploratoire de la modélisation

- Des données autocorrélées spatialement ne nécessitent pas forcément un modèle spatial ;
- Supposons un modèle linéaire $Y = X + \varepsilon$ avec X variable explicative autocorrélée et ε i.i.d. ;
- Des tests de Moran concluront au caractère autocorrélé des variables X et Y ;
- Mais il n'y a pas d'interaction spatiale dans ce modèle, le modèle linéaire est adapté ;
- Nécessité d'analyse exploratoire (Floch 2012) mais ce n'est pas suffisant pour le choix du modèle ;
- Un test de Moran adapté peut être conduit sur les résidus.

Les différentes interactions

3 types d'interaction spatiale :

- Une interaction endogène, i.e. que la décision économique d'un agent va dépendre de la décision de ses voisins ;
- Une interaction exogène, i.e. que la décision économique d'un agent va dépendre des caractéristiques observables de ses voisins ;
- Une corrélation spatiale des effets liée à de mêmes caractéristiques inobservées.

Ne pas tenir compte de ces interactions engendre un biais des estimateurs d'un modèle linéaire et/ou de leur précision.

Le modèle de Manski

- Le modèle de Manski (1993) intègre ces 3 effets :

$$Y = \rho \cdot WY + X \cdot \beta + WX \cdot \theta + u$$

$$u = \lambda \cdot Wu + \varepsilon$$

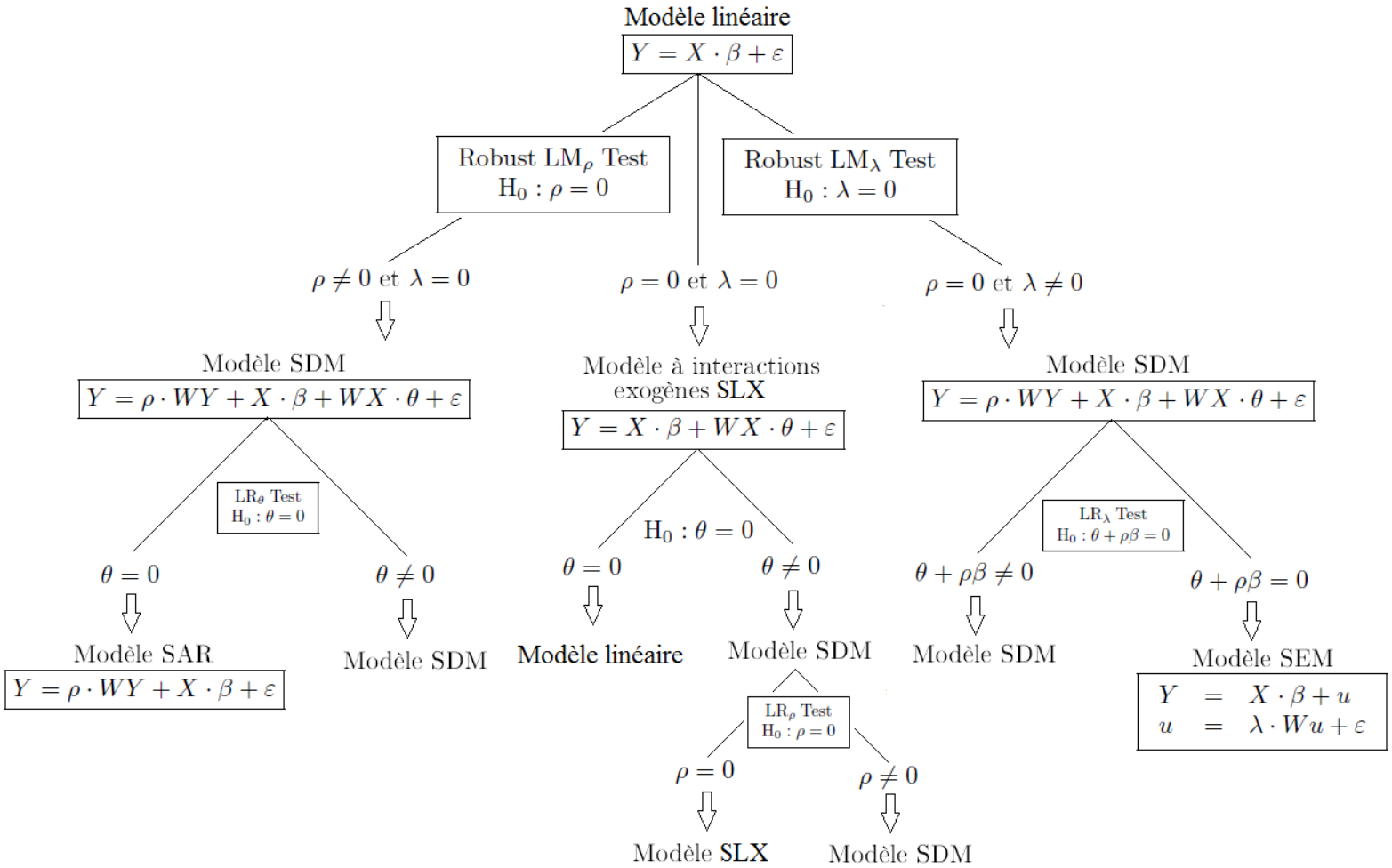
Avec ρ l'interaction endogène, θ l'interaction exogène, λ l'interaction résiduelle et W une matrice de voisinage.

- Problème : ce modèle n'est pas identifiable dans le cas général. Intuition avec les effets de pairs.
- Contraintes d'identification : contraindre un ou deux paramètres à 0.

La galaxie des modèles linéaires d'économétrie spatiale

- Deux modèles classiques :
  - Le modèle SAR (Spatial AutoRegression) $Y = \rho \cdot WY + X \cdot \beta + \varepsilon$, i.e. $\lambda = \theta = 0$
  - Le modèle SEM (Spatial Error Model) $Y = X \cdot \beta + u$ et $u = \lambda \cdot Wu + \varepsilon$, i.e. $\rho = \theta = 0$
- Le modèle actuellement plébiscité (LeSage et Pace 2009), le Spatial Durbin Model (SDM), $Y = \rho \cdot WY + X \cdot \beta + WX \cdot \theta + \varepsilon$ i.e. $\lambda = 0$
  - Intègre les cas particuliers des modèles SAR et SEM ;
  - Robuste à la présence d'autocorrélation résiduelle.
- De nombreux autres modèles.
- Estimation par maximum de vraisemblance des modèles dans le cas de variables exogènes.

Une approche itérative de choix de modèle (Elhorst 2010)



Une approche itérative de choix de modèle (Elhorst 2010)

- Pas de règles intangibles : une aide statistique à la décision, nécessité de se baser aussi sur la théorie économique et la littérature ;
- A partir du modèle non spatial (MCO), tester la présence d'autocorrélation spatiale ;
- En cas d'autocorrélation, estimer le modèle SDM et tester la significativité des paramètres d'interactions ;
- En cas d'absence d'autocorrélation, estimer le modèle à interactions exogènes ($\theta \neq 0$) et tester la significativité des paramètres θ.

L'interprétation des résultats : attention aux rétroactions

- Le modèle SEM s'interpète comme le modèle MCO ;
- Pour le modèle SAR $Y = \rho \cdot WY + X\beta + \varepsilon$, le modèle peut également s'écrire

$$\begin{aligned} Y &= (1-\rho W)^{-1} X\beta + (1-\rho W)^{-1} \varepsilon \\ &= \sum_{r=1}^{k} (1-\rho W)^{-1} \beta_r X_r + (1-\rho W)^{-1} \varepsilon \\ &= \sum_{r=1}^{k} S_r(W) X_r + (1-\rho W)^{-1} \varepsilon \end{aligned}$$

- La valeur prédite est donc $\hat{y} = (1-\hat{\rho} W)^{-1} X\hat{\beta}$ et non $X\hat{\beta}$ comme dans un modèle MCO classique.

L'interprétation des résultats : attention aux rétroactions

- L'effet marginal (pour une variable quantitative) d'une modification de la variable X_{ir} (pour l'individu i et la variable r) n'est pas β_r mais la valeur diagonale de rang i de la matrice $(1-\rho W)^{-1}\beta_r$ ou $S_r(W)$.
- Une modification de mon territoire impacte mes voisins, ce qui m'impacte en retour.

Indicateurs pour l'interprétation

- L'effet direct moyen (leSage et Pace 2009) correspond à la moyenne des termes diagonaux de la matrice S_r. C'est cet indicateur qui est le plus proche de l'interprétation des coefficients β calculés par MCO.
- L'effet total moyen correspond à une moyenne de l'ensemble des termes de la matrice S_r.
- L'effet indirect moyen est la différence entre l'effet total moyen et l'effet direct moyen.
- De tels indicateurs peuvent être définis pour le modèle SDM $Y = \rho \cdot WY + X\beta + WX \cdot \theta + \varepsilon$, mais leurs calculs doivent tenir compte des interactions exogènes $WX \cdot \theta$.

Pour aller plus loin

- Les modèles supposent une observation exhaustive de la population, ce qui est peu adapté aux données d'enquête ;
- Le choix de la matrice de poids fait l'objet de controverses :
  - Supposée exogène, i.e. indépendante des variables explicatives X ;
  - Tentatives récentes d'utilisation de matrices endogènes ou estimées de manière non paramétrique ;
  - Choix empirique de définir plusieurs matrices pour tester la robustesse des analyses.
- Le découpage administratif ne correspond pas forcément à la réalité économique des relations entre agents (MAUP "Modifiable Areal Unit problem") ;
- L'interprétation des résultats n'est valable que pour le découpage géographique choisi (risque de régression dite écologique).

Et si le phénomène est hétérogène spatialement ?

- Il peut exister de l'hétérogénéité spatiale, i.e. des déterminants différents par zone plus qu'un phénomène de diffusion ;
- Deux formes d'hétérogénéité existent :
  - Hétéroscédasticité : estimateurs convergents mais statistiques de tests non valides ;
  - Variabilité spatiale des paramètres ou de la forme fonctionnelle du modèle : ajout d'indicatrices géographiques, lissage géographique, régression géographique pondérée.

La régression géographique pondérée

- Pour chaque observation, modèle estimé sur son voisinage en pondérant les observations selon leur distance.
- Modèle linéaire classique cas particulier où les coefficients sont stables dans l'espace.

$$Y_i = \sum_{j=0}^{p} \beta_j (u_i, v_i) X_{ij} + \varepsilon_i$$

Avec (u_i, v_i) le couple désignant les coordonnées du point i et $\beta_j (u_i, v_i)$ les paramètres pour chaque observation de la variable j.

- Estimation proche des MCO avec des poids définis à l'aide d'une fenêtre h (distance autour de chaque observation) à l'instar des méthodes par noyaux.

Et si le phénomène est hétérogène spatialement ?

- Le partage “pédagogique” entre autocorrélation et hétérogénéité ne doit pas faire oublier les interactions entre les deux ;
- Il reste délicat de distinguer hétérogénéité et corrélation spatiales, en l'absence de méthode identifiant de manière distincte ces deux phénomènes ;
- Estimer des modèles concurrents et cartographier les résultats peut permettre de discuter la présence de ces phénomènes ;
- Le choix de modélisation reposera aussi sur la question posée : analyser par exemple la diffusion de phénomènes entre régions ou au contraire analyser les différences structurelles entre territoires.

Illustration : Taux de chômage locaux

- Modélisation du taux de chômage localisé (par zone d'emplois) à l'aide de caractéristiques de
  - la population active (% des peu diplômés et des moins de 30 ans dans la population active) ;
  - la structure économique (% des emplois dans le secteur industriel et dans le secteur public) ;
  - du marché du travail (taux d'activité).
- Modèle descriptif et illustratif ;
- Aucune conclusion causale possible ;
- 6 matrices de voisinage envisagées ;
- Description des commandes du logiciel R.

Matrices de voisinage

Contiguité

Distance Inverse

Flux domicile-travail

Déterminants du taux de chômage par zone d'emploi

	(1) MCO	(2) SEM	(3) SAR	(4) SDM
Taux d'activité	-0.622***	-0.498***	-0.437***	-0.472***
	(0.039)	(0.041)	(0.038)	(0.042)
% Actifs Peu Diplômés	0.186***	0.184***	0.138***	0.182***
	(0.026)	(0.027)	(0.022)	(0.027)
% Jeunes Actifs 15-30 ans	0.138***	0.196***	0.087**	0.209***
	(0.043)	(0.045)	(0.037)	(0.046)
% Emploi Industriel	-0.062***	-0.018	-0.036***	-0.015
	(0.012)	(0.012)	(0.010)	(0.012)
% Emploi Public	-0.068***	-0.044***	-0.063***	-0.042**
	(0.019)	(0.016)	(0.016)	(0.016)
$\hat{\rho}$			0.519***	0.629***
			(0.049)	(0.064)
$\hat{\lambda}$		0.747***		
		(0.051)		
$\hat{\theta}$, Taux d'activité				0.157*
				(0.083)
$\hat{\theta}$, % Actifs Peu Diplômés				-0.135***
				(0.045)
$\hat{\theta}$, % Jeunes Actifs 15-30 ans				-0.140 *
				(0.072)
$\hat{\theta}$, % Emploi Industriel				-0.044**
				(0.020)
$\hat{\theta}$, % Emploi Public				-0.024
				(0.037)
Constante	51.653***	39.729***	34.470***	27.456***
	(3.635)	(3.685)	(3.407)	(6.766)
Observations	297	297	297	297
AIC	1072	967	980	960
Test Facteur Commun				0.004
Test LM residual auto.			0.003	0.572

Impacts directs et indirects du modèle SDM

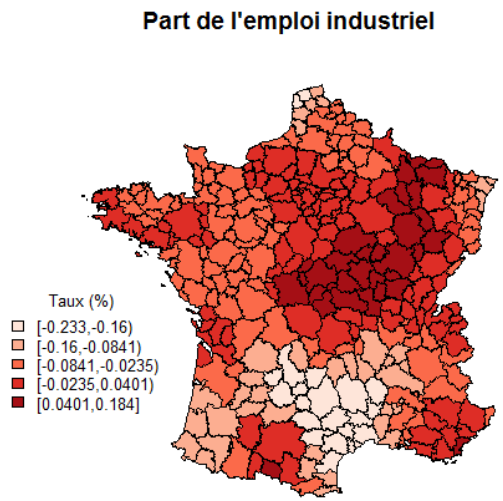
	MCO	SDM Direct	SDM indirect
Taux d'activité	-0.622	-0.490	-0.357
	[-0.700,-0.545]	[-0.574,-0.409]	[-0.658,-0.073]
% Actifs Peu Diplômés	0.186	0.180	-0.053
	[0.136,0.237]	[0.122,0.230]	[-0.254,0.137]
% Jeunes Actifs 15-30 ans	0.138	0.207	-0.023
	[0.054,0.223]	[0.119,0.299]	[-0.352,0.301]
% Emploi Industriel	-0.062	-0.022	-0.136
	[-0.087,-0.038]	[-0.045,0.001]	[-0.229,-0.059]
% Emploi Public	-0.068	-0.049	-0.130
	[-0.106,-0.030]	[-0.081,-0.014]	[-0.335,0.046]

Résultats

- A partir d'une matrice inverse de la distance ;
- Les critères statistiques conduiraient à retenir un modèle SDM ;
- Pour des raisons de parcimonie, le choix d'un modèle SEM pourrait être envisagé ;
- Le choix d'un modèle SAR serait ici déconseillé. Un test montre qu'une autocorrélation spatiale résiduelle reste présente. L'interprétation des résultats peut alors être erronée ;
- L'effet indirect négatif du % d'emploi industriel reste surprenant.

Distribution des paramètres locaux

- A l'aide d'une régression géographique pondérée ;
- Des particularités régionales, permettant de comprendre des résultats surprenants par exemple la relation nulle (ou négative) entre emploi industriel et taux de chômage ;
- Cette analyse devrait nous amener à modifier notre modèle, par l'inclusion d'autres variables ou de paramètres de corrélation spatiale par zones géographiques.



Conclusion et perspectives

- Les modèles d'économétrie spatiale définissent un cadre cohérent (et paramétrique) pour modéliser les interactions entre agents économiques, pas uniquement géographiques ;
- Le choix d'un modèle repose sur des critères statistiques, mais sera aussi fonction de la question et de la théorie économique ;
- Attention, en général, cela ne permet pas d'estimer un effet causal ou de corriger l'endogénéité des variables explicatives ;
- Distinguer hétérogénéité et corrélation spatiales reste néanmoins difficile ;
- Confronter plusieurs modèles spatiaux permet de discuter l'incertitude du processus générateur des données et la robustesse des résultats ;
- Le raffinement méthodologique doit être mis en regard de la complexité des nouveaux modèles, en matière d'interprétation.

Quelques références

- Floch, J.M. (2012) Détection des disparités socio-économiques : l'apport de la statistique spatiale, Document de travail INSEE H2012/04.
- Floch, J.M., et Le Saout R. (2016) Econométrie spatiale : une introduction pratique, Document de travail INSEE à venir.
- Le Gallo, J. (2002) Économétrie spatiale : l'autocorrélation spatiale dans les modèles de régression linéaire, Economie & prévision, 155(4), 139-157.
- LeSage, J., et Pace R. K. (2009) Introduction to Spatial Econometrics, CRC Press, Taylor & Francis Group.