Analyse économétrique des migrations résidentielles

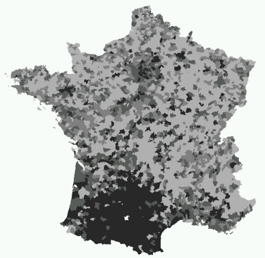
Gaël Guymarc

Séminaire de méthodologie statistique

20 janvier 2015

Introduction

- Stage d'application (ENSAE 2è année), été 2014
- Application de méthodes économétriques innovantes (LeSage & Pace 2014, Thomas-Agnan 2015) sur une base de données interne à l'INSEE non encore exploitée par ce champ de l'économétrie (flux migratoires tirés du RSL)
- Analyse critique des méthodes spatiales et de leur implémentation



Analogie avec la physique

Par analogie avec la physique (interaction gravitationnelle), modélisation des flux migratoires selon la population des deux points d'intérêt et la distance qui les sépare : modèle gravitaire

$$\phi_{ij} = k \frac{P_i^\alpha P_j^\beta}{D_{ij}\gamma}$$

Enrichissement et linéarisation

- Modèle enrichi :

$$\phi_{ij} = k \frac{\prod_{q=1}^{Q} O_{iq}^{\alpha_q} \prod_{p=1}^{P} D_{jp}^{\beta_p}}{D_{ij}^{\gamma}}$$

avec $\{O_{iq}\}_{q\in\{1,Q\}}$ et $\{D_{jp}\}_{p\in\{1,P\}}$ ensembles de variables caractérisant respectivement l'origine i et la destination j du flux (population, taux de chômage, pourcentage de jeunes...)

- Linéarisation par passage au logarithme :

$$log(\phi_{ij}) = log(k) + \sum_{q=1}^{Q} \alpha_q log(O_{iq}) + \sum_{p=1}^{P} \beta_p log(D_{jp}) - \gamma log(D_{ij}) + \epsilon_{ij}$$

Origine, destination, ou origine-destination

- **Origine** : destination fixée. Flux des cantons métropolitains i vers une agglomération fixe j_0 (une observation = un canton)

$$\varphi_{ij_0} = c + X_i\beta_o + \epsilon_i$$

- **Destination** : origine fixée. Flux d'une agglomération fixe i_0 vers les cantons métropolitains j (une observation = un canton)

$$\varphi_{i_0j} = c + X_j\beta_d + \epsilon_j$$

- **Origine-destination** : ni origine, ni destination fixée. Flux inter-agglomérations (les 50 plus grandes) (une observation = un couple d'agglomérations)

$$\varphi_{ij} = c + X_i\beta_o + X_j\beta_d + \epsilon_{ij}$$

Variables et sources

Selon l'approche choisie :

- φ = flux migratoires agrégés, selon le cas, par canton (approche origine ou destination) ou par agglomération (approche origine-destination). Source : **RSL**
- X = distance, population, taux de chômage, taux de jeunes, taux de seniors, revenu fiscal de référence, indicatrice de chef-lieu d'arrondissement, de département ou de région. Sources : **référentiels géographiques, recensement et EDL**.

Insuffisance

Mais modèle gravitaire insuffisant en présence de données présentant une structure spatiale. Conséquences possibles :

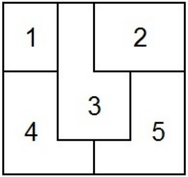
- Estimation biaisée des coefficients (erreur de spécification du modèle)
- Estimation biaisée de la précision (termes d'erreur spatialement corrélés)

⇒Nécessité d'intégrer au modèle gravitaire de base une structure spatiale (LeSage & Pace, 2008, 2014)

Notion de voisinage

- La matrice de voisinage est une notion centrale en économétrie spatiale
- Associée au maillage territorial des unités statistiques
- Permet une description du voisinage d'un individu par l'expérimentateur

Exemple



$$W = \begin{array}{c|ccccc} & 1 & 2 & 3 & 4 & 5 \\ \hline 1 & & & 1/2 & 1/2 & \\ 2 & & & 1/2 & & 1/2 \\ 3 & 1/4 & 1/4 & & 1/4 & 1/4 \\ 4 & 1/3 & & 1/3 & & 1/3 \\ 5 & & 1/3 & 1/3 & 1/3 & \end{array}$$

$$Wy = \begin{bmatrix} & & 1/2 & 1/2 & \\ & & 1/2 & & 1/2 \\ 1/4 & 1/4 & & 1/4 & 1/4 \\ 1/3 & & 1/3 & & 1/3 \\ & 1/3 & 1/3 & 1/3 & \end{bmatrix} . \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{bmatrix} = \begin{bmatrix} 1/2\,(y_3 + y_4) \\ 1/2\,(y_3 + y_5) \\ 1/4\,(y_1 + y_2 + y_4 + y_5) \\ 1/3\,(y_1 + y_3 + y_5) \\ 1/3\,(y_2 + y_3 + y_4) \end{bmatrix}$$

Statistique de Moran

- Pour une définition donnée du voisinage (i.e. une matrice de voisinage fixée), la statistique de Moran permet de tester la structure spatiale des données :

$$I = \frac{\frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}}}{\frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n}}$$

- Sous H_0 = indépendance des observations, I suit une loi normale (conduite de tests d'indépendance)

Statistique de Moran

TABLE: Test de l'autocorrélation sur les résidus du modèle gravitaire estimé par les MCO (approche univariée origine et destination) :

Sens des flux	Agglomération	Statistique	Espérance (sous H_0)	Variance (sous H_0)	p-value
Entrants	Paris	0,37	$-1{,}27.10^{-3}$	$1{,}04.10^{-4}$	0,00
	Toulouse	0,27	$-1.23.10^{-3}$	$9{,}88.10^{-5}$	0,00
	Grenoble	0,22	$-1.26.10^{-3}$	$9{,}88.10^{-5}$	0,00
	Strasbourg	0,27	$-1.25.10^{-3}$	$9{,}87.10^{-5}$	0,00
	Tours	0,26	$-1.25.10^{-3}$	$9{,}88.10^{-5}$	0,00
	Rennes	0,32	$-1.25.10^{-3}$	$9{,}87.10^{-5}$	0,00
Sortants	Paris	0.,56	$-1{,}27.10^{-3}$	$1{,}04.10^{-4}$	0,00
	Toulouse	0,26	$-1.23.10^{-3}$	$9{,}88.10^{-5}$	0,00
	Grenoble	0,29	$-1.26.10^{-3}$	$9{,}89.10^{-5}$	0,00
	Strasbourg	0,31	$-1.25.10^{-3}$	$9{,}87.10^{-5}$	0,00
	Tours	0,31	$-1.25.10^{-3}$	$9{,}88.10^{-5}$	0,00
	Rennes	0,37	$-1.25.10^{-3}$	$9{,}87.10^{-5}$	0,00

Modèle de Manski

- Introduction du terme d'interaction spatiale ***W*** dans le modèle de base :

$$\begin{cases} y = \rho W_1 \mathrm{y} + \mathrm{X}\beta + W_2 \mathrm{X}\theta + u \\ u = \lambda W_3 \mathrm{u} + \epsilon \end{cases}$$

SAR	SEM	Durbin
$\theta = \lambda = 0$	$\rho = \theta = 0$	$\lambda = 0$
$\mathrm{y} = \rho W y + X\beta + \epsilon$	$\begin{cases} y = X\beta + u \\ u = \lambda W u + \epsilon \end{cases}$	$\mathrm{y} = \rho W_1 \mathrm{y} + \mathrm{X}\beta + W_2 \mathrm{X}\theta + \epsilon$

Identifiabilité Vs. spécification

- Modèle de Manski non identifiable : prise d'hypothèse préalable à la phase d'estimation
- Mais qualité des estimations sensible à la spécification (biais, mesure de précision erronée)
- Outils d'aide à la decision pour le choix de la bonne spécification

Cadre d'estimation

- Voisinage défini au sens des 5 plus proches voisins
- Trois types d'approche conservés :
  - Origine
  - Destination
  - Origine-destination
- Implémentation des modèles spatiaux SAR et SEM

Résultats

			Modèle		
			SAR	SEM	Gravitaire
$\rho(SAR)/\lambda(SEM)$			$0,38^{***}$	$0,81^{***}$	-
Constante			$-19,12^{***}$	$-21,87^{***}$	$-18,70^{***}$
Distance			$-0,94^{***}$	$-1,10^{***}$	$-1,23^{***}$
Population	origine		$0,77^{***}$	$0,89^{***}$	$0,91^{***}$
	destination		$0,80^{***}$	$0,89^{***}$	$0,97^{***}$
Chef-lieu	arrondissement	origine	$0,03$	$0,07$	$0,08$
		destination	$0,30^{***}$	$0,23$	$0,47^{***}$
	département	origine	$0,18^{***}$	$0,35^{*}$	$0,38^{***}$
		destination	$0,42^{***}$	$0,43^{**}$	$0,75^{***}$
	région	origine	$0,27^{**}$	$0,42^{**}$	$0,47^{***}$
		destination	$0,69^{***}$	$0,68^{***}$	$1,07^{***}$
Chômage	origine		$0,31^{**}$	$0,42^{*}$	$0,60^{***}$
	destination		$-0,56^{***}$	$0,07$	$-0,76^{***}$
Jeunes	origine		$0,02$	$-0,16$	$-0,51^{*}$
	destination		$0,11$	$-0,40$	$-0,23$
Seniors	origine		$0,21$	$0,01$	$-0,08$
	destination		$1,25^{***}$	$0,62^{*}$	$1,53^{***}$
RFR	origine		$0,75^{***}$	$1,18^{***}$	$1,14^{***}$
	destination		$-0,31$	$0,87^{**}$	$-0,51^{**}$
$\star\star\star$:significativité à 1% ; $\star\star$:significativité à 5% ; $\star$:significativité à 10%					

Quelques difficultés

- Charge calculatoire accrue
- Divergences ponctuelles entre les modèles (signe, significativité statistique)
- Particularités d'emploi du modèle SAR (comparabilité, interprétation des résultats)

Pour résumer

- Méthodes spatiales justifiées par la structure spatiale (testable) de données
- Deux sources de variabilité dans les estimations :
  - Définition ex ante du voisinage
  - Choix d'une spécification
- Charge calculatoire accrue
- Particularité d'emploi du modèle SAR
- Méthode alternative : théorie des graphes

Merci pour votre attention