Groupe de lecture "Econométrie des données d'enquête"

Compte-rendu de la 7ième réunion, 15 juin 2015
Extensions

Suivi par Marine Guillerm et Ronan Le Saout

Cette dernière session s'articulait sur deux présentations qui avaient pour but d'étendre la réflexion sur les liens entre l'économétrie et les sondages. La première, par Pascal Ardillay, dressait un panorama des méthodes d'estimation sur petits domaines. La deuxième, par Thomas Merly Alpa et Raphaël Lardeux, étudiait l'applicabilité de l'économétrie spatiale sur des données d'enquête.

La réunion s'est conclu sur les perspectives futures de ce groupe de lecture. L'historique des réunions (présentations et comptes-rendus) est mis en ligne sur l'Intranet de la DMCSI. Un travail de recherche sur le calcul de variance des modèles économétriques est envisagé par l'ENSAI, d'autres travaux en partenariat pourraient être envisagés. Enfin, un document de travail de synthèse pourrait être rédigé au sein de la DMCSI dans les prochains mois.

1 Les méthodes d'estimation sur petits domaines

Pascal Ardilly a présenté les méthodes d'estimation sur petits domaines. On se place dans le cadre où on souhaite estimer un total ou une moyenne sur un domaine noté a, i.e. sur une sous population de la population totale. On dispose d'un échantillon s de taille n. Parmi ces individus sélectionnés, n_a (aléatoire) individus appartiennent au domaine a. Il est possible d'estimer sans biais les paramètres sur le domaine mais lorsque n_a est faible, la variance de l'estimateur est importante. C'est ce problème que les estimations sur petits domaines cherchent à résoudre.

Trois grandes catégories de méthodes ont été présentées.

1. Le calage au niveau local. Il s'agit de décliner la méthode du calage au niveau du domaine: on utilise l'échantillon du domaine et les marges locales. Il s'agit d'une estimation directe au sens où on n'utilise pas d'information en dehors du domaine.

2. Modèle impliquant des paramètres non aléatoires, approche entièrement descriptive.
   Le principe de base de l'estimation consiste à faire une hypothèse descriptive sur le domaine, du type "paramètre sur a = paramètre sur U". L'exemple de l'estimateur par la régression a été développé. On suppose que le coefficient de la régression de Y sur X (variables auxiliaires) dans le domaine est égal à celui sur l'ensemble de la population, soit $\hat{Y}_a = X_a^T \hat{B}$ avec $\hat{B} = \left(\sum_{i \in s} \frac{X_i X_i^T}{P_i}\right)^{-1} \left(\sum_{i \in s} \frac{X_i Y_i^T}{P_i}\right)$, calculé sur l'échantillon complet.
   Il s'agit d'une estimation indirecte au sens où on utilise de l'information en dehors du domaine. La spécificité du domaine n'intervient que par l'intermédiaire des variables auxiliaires.
   Cette méthode pose problème car il n'est pas possible d'estimer de manière stable les Erreurs Quadratiques Moyennes (EQM).

3. Modèle expliquant Y par des variables auxiliaires X, version stochastique.
   Comme la méthode précédente, il s'agit d'une estimation indirecte, au sens où l'échantillon complet s est utilisé (on utilise donc de l'information en dehors du domaine). La spécificité est que Y est aléatoire.

   On spécifie un modèle valable sur l'ensemble de la population. Il va servir à prédire $Y_a = \sum_{i=1}^{N_a} Y_i$. Les paramètres du modèle sont estimés à partir de l'ensemble des observations de l'échantillon s (pas sur les seuls observations du domaine d'intérêt).

Un exemple a été développé. On suppose la variable d'intérêt Y quantitative et continue. Le domaine est pris en compte en incluant dans le modèle un effet aléatoire v_a. Le modèle prend ainsi la forme d'un modèle linéaire mixte:

$$Y_{a,i} = X_{a,i}^t \beta + v_a + e_{a,i}$$

v_a et $e_{a,i}$ font partie du résidu. Des hypothèses classiques sont faites sur ces résidus (indépendance, homoscédasticité, normalité).

X est une variable auxiliaire, connue pour l'ensemble des individus du domaine et explicative de la variable d'intérêt.

Le modèle spécifié sur la population complète doit rester valide sur l'échantillon. Cela revient à faire l'hypothèse que l'échantillonnage est non informatif. Le prédicteur BLUP (Best Linear Unbiased Prediction) a été présenté. Le paramètre β est estimé à partir de tous les individus échantillonnés. Il est donc très stable.

2 Sondage et économétrie spatiale

Thomas Merly Alpa et Raphaël Lardeux ont présenté un travail de recherche initié dans le cadre du cours d'économétrie spatiale à l'ENSAE. L'objet de l'économétrie spatiale est d'étudier un phénomène économique en prenant en compte cette dimension spatiale. Les observations sont alors dépendantes car elles interagissent au sein de l'espace. Un individu a une influence sur ses voisins, qui ont eux-mêmes une influence sur cet individu. Pour étudier ce phénomène, des modèles particuliers sont utilisés. Ils s'appuient sur la définition d'une matrice de pondération spatiale W qui définit les relations de voisinage entre les individus. Deux principaux modèles ont été étudiés:

- Le modèle SAR (Spatial Auto-Regressive) $Y = \rho W \cdot Y + X \cdot \beta + \varepsilon$, où l'interaction spatiale (identifiée par le paramètre ρ) est définie pour la variable à expliquer. Des effets directs et indirects sont estimés par ce modèle.
- Le modèle SEM (Spatial Error), $Y = X \cdot \beta + \varepsilon$ et $\varepsilon = \lambda W \cdot \varepsilon + \mu$, où l'interaction spatiale (identifiée par le paramètre λ) est définie pour le terme d'erreur. Il n'y a pas d'effet indirect dans ce modèle.

L'estimation de ces modèles s'appuie sur des données exhaustives. Les données étant dépendantes, une seule réalisation du processus générateur des données est en effet observée. Avec des données d'enquête, les observations de la population ne sont au contraire observées que sur un échantillon. La question posée est donc de savoir si l'autocorrélation spatiale peut être étudiée à l'aide de données d'enquête (un échantillon), et le rôle du plan de sonsage dans ce cadre.

La démarche consiste à étudier différentes stratégies de tirage (et donc de plan de sondage) parmi les 3114 comtés américains, à partir de jeux de données simulés de modèles SAR et SEM. Les matrices de pondération spatiale s'appuient principalement sur la distance entre comtés car c'est le critère le plus stable en présence d'observations manquantes. Pour chaque tirage, le modèle ayant servi à générer les observations (SAR ou SEM) est estimé sur le seul échantillon. En répétant l'observation, on peut étudier pour différents taux de sondage et stratégies de tirage (sondage aléatoire simple, stratifié, par grappes) la valeur moyenne et la dispersion des paramètres $\hat{\rho}$ et $\hat{\lambda}$ ainsi que les effets directs et indirects du modèle SAR.

Avec un sondage aléatoire simple ou stratifié, on ne détecte de l'autocorrélation spatiale que pour un taux de sondage supérieur à 1/10. Les paramètres estimés sont alors très éloignés de leur valeur théorique (l'intervalle de confiance sous hypothèse de normalité n'inclut pas la vraie valeur du paramètre). Avec un sondage par grappes géographiques, les résultats sont améliorés. L'autocorrélation spatiale est détectée dès que le taux de sondage est supérieur à 3/100, avec un paramètre plus proche de sa vraie valeur.

En termes d'implication pour les études économiques, il apparaît illusoire d'estimer des modèles d'économétrie spatiale si les données ont été obtenues par enquête, avec un plan de sondage non défini spatialement et un taux de sondage faible. Pour les enquêtes ménages, en dehors de l'enquête emploi qui est réalisée par grappes (mais pour laquelle on ne connait pas la localisation précise des individus), il conviendrait donc au préalable de calculer des statistiques agrégées (par département, IRIS...) pour estimer un modèle d'économétrie spatiale. Pour les enquêtes entreprises, les taux de sondage peuvent être plus élevés (et la localisation précise connue), notamment pour les grandes entreprises. L'estimation de modèles d'économétrie spatiale pourraient être envisagée sur ces données. Les résultats présentés restent provisoires, le travail étant en cours d'avancement.

La discussion a abordé plusieurs points. 1) Dans tous les cas, les intervalles de confiance asymptotiques ne recouvrent pas la vraie valeur des paramètres (même avec un sondage par grappes), ce qui montrerait la présence d'un biais important des estimateurs. 2) Le choix de la matrice de pondération spatiale W fait débat, bien que les résultats semblent robustes à différents choix. 3) La présence d'effets de bords (frontières) n'est pas prise en compte. 4) Même dans le cas où le sondage est tiré par grappes, la théorie des sondages visera à obtenir des grappes les plus hétérogènes possibles pour réduire la variance des estimateurs. S'il y a une très forte auto-corrélation spatiale dans les grappes tirées, c'est que le plan de sondage n'est pas très efficace. Il y a donc une opposition entre les objectifs de la théorie des sondages et ce que cherche à mesurer l'économétrie spatiale. 5) Par ailleurs, pour des taux de sondage plus élevés (par exemple sur données entreprises), des algorithmes spécifique (EM par exemple) permettent de tenir compte dans l'estimation de la présence de valeurs manquantes.