Spatial Statistics in Discrete-Choice Models

DRWESH Emad-Aldeen [1] & DABO Sophie [2] & FONCEL Jêrome [3]

[1] *Doctorant, Université Lille, Laboratoire LEM, Domaine du pont de bois, maison de la recherche, 59653 Villeneuve d'ascq cedex, emad-aldeen.drwesh@etu.univ-lille3.fr*

[2] *Professeure en Mathématique Appliquées, Université Lille, Laboratoire LEM, Domaine du pont de bois, maison de la recherche, 59653 Villeneuve d'ascq cedex, sophie.dabo@univ-lille3.fr*

[3] *Professeur en Economie, Université Lille, Laboratoire LEM, Domaine du pont de bois, maison de la recherche, 59653 Villeneuve d'ascq cedex, jerome.foncel@univ-lille3.fr*

Abstract. Spatial binary models are useful in many areas such as in economic and epidemiology where activities are often located in space. A way that makes the analysis of such spatial activities possible is to find a kind of correlation between some random variables in one location with others at neighboring locations, see for instance Pinkse and Slade (1998), Klier and McMillen (2008) and Wang et al. (2013). We proposed here to describe and analyze the spatial (geographical) variation in disease (cancer) with respect to some risk factors using spatial binary models containing spatial latent choice variable and/or spatial autoregressive disturbances in a context of sampling data. This problem is known as Choice-Based Sampling (CBS) in discrete choice model. Unlike the random sample where all items in the population have the same probability of being chosen, the Choice-Based Sampling (CBS) in discrete choice model is a type of sampling where the classification of the population into subsets to be sampled is based on the choices or outcomes. In this context, the use of standard Maximum likelihood estimation (MLE) procedure in CBS could lead to an inconsistent (asymptotically biased) estimation. Thus, in addition to the conditional maximum likelihood estimator (Manski and McFadden, 1981), we adapt the GMM approach (Imbens, 1992) in our context of spatial Choice-Based Sampling. We also provide a GMM estimator based on the generalized residuals (see Gourieroux et al. (1987)). We present a Monte Carlo experiment to investigate the finite sample performance of these estimation methods. An application to real cancer data in northern France is also provided.

Keywords. Spatial econometrics, Spatial autocorrelation, Choice-Based Sampling, GMM.

Résumé. Les modèles binaires spatiaux sont utiles dans de nombreux domaines tels que l'épidémiologie et l'économie où les activités sont souvent situées dans l'espace. Une manière de rendre l'analyse de ces activités spatiales possible est de tenir compte d'une éventuelle dépendance spatiale (Pinkse et Slade (1998), Klier et McMillen (2008) et Wang et al. (2013)). Nous proposons ici de décrire et d'analyser la variation spatiale (géographique) d'une maladie (cancer) à l'égard de certains facteurs de risque. Pour

ce faire, nous avons utilisé des modèles spatiaux binaires contenant une variable spatiale latente de choix et/ou un processus spatial autorégressif d'erreurs dans le cadre d'un échantillonnage des données. Ce problème est connu sous le nom de "Choice-Based Sampling" (CBS) dans les modèles discrets. Contrairement à l'échantillon aléatoire où tous les éléments de la population ont la même probabilité d'être choisi, l'échantillonnage CBS dans le modèle discret est un type d'échantillonnage dans lequel la classification de la population est faite sous forme de sous-ensembles (strates) basés sur les choix alternatifs. Dans ce contexte, l'utilisation de la procédure d'estimation par maximum de vraisemblance standard (MLE) dans le CBS pourrait mener à des estimations incohérentes (asymptotiquement biaisés). Ainsi, en plus de l'estimateur du maximum de vraisemblance conditionnelle (Manski et McFadden, 1981), nous adaptons l'approche GMM (Imbens, 1992) dans notre contexte de CBS spatiale. Nous fournissons également un estimateur GMM, basé sur les résidus généralisés (Gourieroux et al. (1987)). Nous présentons une simulation de Monte Carlo pour étudier la performance, dans un cadre d'un échantillon fini, de ces méthodes d'estimation. Une application sur des données réelles de cancer dans le Nord de la France est également proposée.

Mots-clés. Econométrie spatiale, Autocorrélation spatiale, Choice-based sampling, GMM.

Choice-Based Spatial Modeling

Suppose we have a sample of n observations collected from points or regions located on an irregularly spaced, countable lattice $\mathcal{I} \subset \mathbb{R}^d$, $d \geq 1$; Let $i = (i_1, ..., i_d) \in \mathbb{R}^d$ denote a site, and assume that all sites in $\mathcal{I}$ are located at distances of at least $d_0 > 0$ for each other; i.e $\forall\ i, j \in \mathcal{I}$: $\|i - j\| \geq d_0$. Consider a sample of $i \in \mathcal{I}_n$ ($\mathcal{I}_n$ is a finite subset of $\mathcal{I}$ of cardinal n) individuals and let X be a vector of k exogenous discrete or continuous random variables. Our spatial model with latent dependent (unobserved scalar) variable can be written as

$$\begin{aligned} y^* &= \rho W y^* + X\beta + u; \\ u &= \lambda M u + \epsilon; \\ \epsilon\ &i.i.d.\ \sim N(0, \sigma_\epsilon^2 I_n), \end{aligned} \tag{1}$$

where $y^{*t} = (y_i^*)_{i \in \mathcal{I}_n}$, $\beta \in \mathbb{R}^k$ is unknown, W and M are $n \times n$ known spatial weight matrices with two scalar autoregressive parameters ρ and λ, indicating the degree of spatial dependence and β is a $k \times 1$ vector of coefficients. Consider the binary model, where for individual located at site i, we observe the model

$$y_i = \begin{cases} 1 \ \texttt{if}\ \ y_i^* \geq 0 \\ 0 \ \texttt{if}\ \ y_i^* < 0. \end{cases} \tag{2}$$

Suppose that the sample that is considered is an endogenously stratified sampling in which the classification of the population into M (here $M = 2$) subsets (or strata) to be sampled is based on the choices or outcomes (for instance $y_i = 0, 1$).

This problem is known as Choice-Based Sampling (CBS) in discrete choice model. CBS is usually used when some choices are rarely chosen, unlike the random sample case where all items in the population have the same probability of being chosen. Let $Q(j)$ be the population share of choice j (1 or 0) (marginal probability of observing an individual choosing alternative j). And let $H(j)$ be the probability according to which we draw the stratum j. Assume that we have knowledge on H and Q, then the conditional density of j given $X = x$ in the sample is

$$g(j|x) = \frac{P(j|x,\theta)H(j)/Q(j)}{\sum_{j'=1}^{M} P(j'|x,\theta)H(j')/Q(j')} \ , \tag{3}$$

where $P(j|x,\theta)$ is the probability that one will choose alternative j given x with a finite set of parameters to be estimated $\theta = \{\beta, \rho, \lambda\}$. In this context, the use of standard Maximum likelihood estimation (MLE) procedure in CBS could lead to an inconsistent (asymptotically biased) estimation. Manski and McFadden (1981), proposed maximizing the conditional density given by equation (3) in a context of independent data. This method is usually referred to as conditional maximum Likelihood estimator (CMLE) using observations $(x_i, y_i = j_i)_{i\in\mathcal{I}_n}$:

$$\begin{aligned} \hat{\theta}_{CMLE} &= \arg\min_{\theta} \sum_{i\in\mathcal{I}_n} \log g(j_i|x_i) \\ &= \arg\min_{\theta} \sum_{i\in\mathcal{I}_n} \log\left(\frac{P(j_i|x_i,\theta)H/Q}{\sum_{l=1}^{M} P(l_i|x_i,\theta)H/Q}\right). \end{aligned} \tag{4}$$

In addition to the CMLE , we adapt the GMM approach of Imbens (1992) in our context of spatial Choice-Based Sampling. Here, The score function in CMLE will be interpreted as a moment function. That is, the GMM estimator can be defined as

$$\begin{aligned} \hat{\theta}_{GMM} &= \arg\min_{\theta} R_{n,C_n}(H, Q, \theta) \\ &= \arg\min_{\theta} \frac{1}{n}\sum_{i\in\mathcal{I}_n} \psi(H, Q, \theta, j_i, x_i)'.C_n.\frac{1}{n}\sum_{i\in\mathcal{I}_n} \psi(H, Q, \theta, j_i, x_i), \end{aligned} \tag{5}$$

where ψ is the vector of derivatives of the conditional density (equation (3)) at θ, and C_n is a positive definite weight matrix estimated as

$$n\left[\sum_{i\in\mathcal{I}_n} \psi(H, Q, \hat{\theta}, j_i, x_i)\psi(H, Q, \hat{\theta}, j_i, x_i)'\right],$$

where $(\hat{\theta})$ is the solution of $\arg\min_\theta R_{n,I}(H, Q, \theta)$.
We also provide a GMM estimator based on the generalized residuals denoted $\tilde{r}_i$, $i \in \mathcal{I}_n$ (Gourieroux et al. (1987)) derived from the first order condition for the CMLE by adapting the non CBS spatial GMM (Pinkse and Slade (1998)). The parameter vector θ is then estimated by

$$\hat{\theta}_{GMM_2} = \arg\min_\theta r'ZTZ'r, \tag{6}$$

where Z is a matrix of some instruments and T is a positive definite matrix.
We present a Monte Carlo experiment to investigate the finite sample performance of these three estimation methods. An application to real cancer data in northern France is also provided.

Bibliography

[1]Gourieroux, C., Monfort, A., Renault, E. and Trognon, A. (1987), *Generalized residuals*, Journal of econometrics, 34(1), 5-32.
[2] Imbens, G. W. (1992), *An efficient method of moments estimator for discrete choice models with choice-based sampling*, Econometrica: Journal of the Econometric Society, 1187-1214.
[3] Klier, T. and D. McMillen (2008), *Clustering of Auto Supplier Plants in the United States*, Journal of Business and Economic Statistics, 26(4), 560-471.
[4] Manski, C. F. and McFadden, D. (Eds.). (1981), *Structural analysis of discrete data with econometric applications*, (pp. 202-4). Cambridge, MA: MIT Press.
[5] Pinkse, J. and Slade M.E. (1998), *Contracting in space: An application of spatial statistics to discrete-choice models*, Journal of Econometrics, 85, 125-154.
[6] Wang, H., Iglesias, E. M., and Wooldridge, J. M. (2013). *Partial maximum likelihood estimation of spatial probit models*. Journal of Econometrics, 172(1), 77-89.