L'Actualité économique, Revue d'analyse économique, vol. 69, n° 3, septembre 1993

MODÈLE D'EXPLICATION DE FLUX À COMPOSANTES D'ERREURS SPATIALEMENT CORRÉLÉES*

Denis BOLDUC
Département d'économique
Université Laval
Richard LAFERRIÈRE
Centre de recherche sur les transports
Université de Montréal
Gino SANTAROSSA
Département d'économique
Université Laval

RÉSUMÉ — Dans cette étude, nous proposons une généralisation de la formulation à composantes d'erreurs qui permet de représenter différents effets explicatifs de la présence de corrélation dans les erreurs de modèles de régression avec données de flux. Selon la formulation proposée, le terme d'erreur se décompose en une somme d'une erreur relative à la zone d'origine, une erreur relative à la zone de destination et une erreur associée au flux. Chaque composante d'erreur est issue d'un processus générateur autorégressif spatial d'ordre 1. L'estimation des paramètres du modèle est basée sur la méthode du maximum de vraisemblance. La méthodologie proposée a l'avantage de demeurer applicable même dans le contexte d'échantillons de grande taille.

ABSTRACT — *Spatial Autoregressive Components in Travel Flow Models.* In this study, we propose a generalization of the error component formulation to model the correlation among the errors of a regression based on travel flow data. The error term is broken down into a sum of one component related to the origin zone, one component related to the destination zone and a remainder. Each component is assumed to result from a first-order spatial autoregressive generating process. An efficient estimation approach based on maximum likelihood which addresses the practical implementation of such a model with a large sample size is suggested.

* Article présenté dans le cadre du 30e Congrès annuel de la Société canadienne de science économique, du 15 au 17 mai, Ste-Foy, Québec, 1991. Nous tenons à remercier Pierre Fortin ainsi qu'un arbitre anonyme pour ses commentaires judicieux. Le premier auteur désire remercier le Fonds FCAR nouveaux chercheurs ainsi que le Conseil de recherche en sciences humaines du Canada pour leur suppport financier.

INTRODUCTION

Cet article introduit une formulation générale qui permet de tenir compte du problème d'autocorrélation spatiale des erreurs résiduelles d'un modèle de régression basé sur des données de flux. La formulation provient d'un prolongement du modèle conventionnel à composante d'erreur où chaque composante admet une structure autorégressive spatiale d'ordre 1 qui lui est propre. Le terme d'erreur se scinde en une somme de trois composantes : la première associée à la zone d'origine du flux, la seconde à la zone de destination et la dernière au flux lui-même. La formulation complète du modèle est désignée par l'acronyme EC-SAR(1).

Brandsma et Ketellapper (1979) (BK) et Bolduc, Dagenais et Gaudry (1989) (BDG) constituent deux tentatives intéressantes pour résoudre le problème de corrélation spatiale dans les erreurs de flux. L'approche BK est fondamentalement correcte mais ignore un groupe de facteurs quelquefois importants pour expliquer la corrélation. L'approche BDG, pour sa part, peut être vue comme une généralisation de la formulation BK qui permet d'introduire ces derniers facteurs dans le modèle. L'importance de l'influence entre les erreurs du modèle est caractérisée par des fonctions estimables appelées pondérations. Toutefois, sauf pour des structures de dépendance très simples, la spécification des pondérations peut s'avérer compliquée. L'approche EC-SAR(1) diffère de ces méthodologies car elle traite distinctement les différentes sources d'erreurs de spécifications qui causent la corrélation spatiale des erreurs. De plus, le modèle EC-SAR(1) inclut comme cas particuliers les formulations BDG et BK.

À la section suivante, nous présentons les principaux concepts généralement utilisés dans le champ d'étude des données de flux. La section 2 introduit le modèle à composantes d'erreurs que l'on propose. Par la suite, nous présentons les formulations de BK et de BDG. Dans la dernière section, nous développons un algorithme d'estimation du maximum de vraisemblance qui permet des applications de la méthodologie même lorsque la taille de l'échantillon est grande. Par souci de généralité, nous utilisons une notation qui permet de traiter tant le cas de données équilibrées que celui de données non équilibrées.

1. MODÈLES DE FLUX

La forme générale du modèle de régression linéaire défini avec des données de flux s'écrit comme suit :

$$y_{ij} = \beta_0 + N_{ij}\beta_1 + S_i\beta_2 + S_j\beta_3 + \varepsilon_{ij}, \quad i = 1,\ldots,N, \quad j = 1,\ldots,T, \tag{1}$$

où y_{ij} désigne la variable dépendante qui dans le cas présent représente un flux de l'origine i à la destination j. Le nombre d'origines N et de destinations T peut être égal ou différent selon que le type de données concerne respectivement des flux intra-urbains ou interurbains. Les trois groupes de variables explicatives incorporés dans la spécification sont : N_{ij}, les variables de réseau (ex. : temps de

déplacement, coût, fréquence, etc.), S_i, les variables socio-économiques associées à l'origine i (ex. : population, revenu, niveau d'emploi, etc.), et S_j les variables socio-économiques associées à la destination j. β_0 est un paramètre affecté à la constante du modèle β_1, β_2 et β_3 sont des vecteurs de coefficients fixes.

L'explication fournie à y_{ij} n'implique que l'information propre au marché ij. En effet ni les variables de réseau qui caractérisent les marchés compétitifs is (même origine mais destination différente) ou rj (même destination mais origine différente) ni les variables socio-économiques de zones autres que i et j n'apparaissent à l'équation (1). Il est raisonnable de supposer que les variables de réseau et socio-économiques de zones adjacentes au marché ij puissent influencer y_{ij}, mais l'inclusion de ces variables dans le modèle (1) conduirait à un nombre trop élevé de variables explicatives. Selon Cliff et Ord (1981), l'omission de variables explicatives reliées à la structure physique et économique de la région (distance entre les zones, taille des zones, la longueur de la frontière commune entre les zones adjacentes, etc...) est une source d'autocorrélation spatiale des erreurs résiduelles. Dès lors, une formulation adéquate pour traiter l'autocorrélation spatiale des erreurs résiduelles doit incorporer trois composantes : une reliée aux zones d'origine, une reliée aux zones de destination et une autre associée aux facteurs spécifiques du réseau.

2. LA FORMULATION À COMPOSANTES D'ERREUR AUTORÉGRESSIVES SPATIALE D'ORDRE 1

Nous proposons de définir une structure d'erreur qui tient compte des diverses sources de corrélation mentionnées :

$$\varepsilon_{ij} = \alpha_i + \lambda_j + u_{ij}. \tag{2}$$

Cette décomposition du flux inexpliqué entre les zones i et j est compatible avec la structure de la partie déterministe du modèle de régression. Le terme d'erreur α_i permet de capter les effets non mesurés associés à l'origine i (il contient les variables manquantes du vecteur S_i), λ_j contient les effets manquants reliés à la destination j (il contient les variables manquantes du vecteur S_j) et finalement u_{ij} incorpore les variables de réseau omises de la spécification du modèle. Nous supposons que tous ces termes sont de moyennes nulles et mutuellement non corrélés,

$$E(\alpha_i\lambda_j) = E(\alpha_i u_{ij}) = E(\lambda_i u_{ij}) = 0 \quad \forall i, j. \tag{3}$$

Chaque composante admet un processus autorégressif spatial normalisé d'ordre 1 (SAR(1)).

$$\alpha_i = \rho_1 \sum_{r=1}^{N} w_{1,ir}\alpha_r + \xi_i, \tag{4}$$

INTRODUCTION

Cet article introduit une formulation générale qui permet de tenir compte du problème d'autocorrélation spatiale des erreurs résiduelles d'un modèle de régression basé sur des données de flux. La formulation provient d'un prolongement du modèle conventionnel à composante d'erreur où chaque composante admet une structure autorégressive spatiale d'ordre 1 qui lui est propre. Le terme d'erreur se scinde en une somme de trois composantes : la première associée à la zone d'origine du flux, la seconde à la zone de destination et la dernière au flux lui-même. La formulation complète du modèle est désignée par l'acronyme EC-SAR(1).

Brandsma et Ketellapper (1979) (BK) et Bolduc, Dagenais et Gaudry (1989) (BDG) constituent deux tentatives intéressantes pour résoudre le problème de corrélation spatiale dans les erreurs de flux. L'approche BK est fondamentalement correcte mais ignore un groupe de facteurs quelquefois importants pour expliquer la corrélation. L'approche BDG, pour sa part, peut être vue comme une généralisation de la formulation BK qui permet d'introduire ces derniers facteurs dans le modèle. L'importance de l'influence entre les erreurs du modèle est caractérisée par des fonctions estimables appelées pondérations. Toutefois, sauf pour des structures de dépendance très simples, la spécification des pondérations peut s'avérer compliquée. L'approche EC-SAR(1) diffère de ces méthodologies car elle traite distinctement les différentes sources d'erreurs de spécifications qui causent la corrélation spatiale des erreurs. De plus, le modèle EC-SAR(1) inclut comme cas particuliers les formulations BDG et BK.

À la section suivante, nous présentons les principaux concepts généralement utilisés dans le champ d'étude des données de flux. La section 2 introduit le modèle à composantes d'erreurs que l'on propose. Par la suite, nous présentons les formulations de BK et de BDG. Dans la dernière section, nous développons un algorithme d'estimation du maximum de vraisemblance qui permet des applications de la méthodologie même lorsque la taille de l'échantillon est grande. Par souci de généralité, nous utilisons une notation qui permet de traiter tant le cas de données équilibrées que celui de données non équilibrées.

1. MODÈLES DE FLUX

La forme générale du modèle de régression linéaire défini avec des données de flux s'écrit comme suit :

$$y_{ij} = \beta_0 + N_{ij}\beta_1 + S_i\beta_2 + S_j\beta_3 + \varepsilon_{ij}, \quad i = 1,\ldots,N, \quad j = 1,\ldots,T, \tag{1}$$

où y_{ij} désigne la variable dépendante qui dans le cas présent représente un flux de l'origine i à la destination j. Le nombre d'origines N et de destinations T peut être égal ou différent selon que le type de données concerne respectivement des flux intra-urbains ou interurbains. Les trois groupes de variables explicatives incorporés dans la spécification sont : N_{ij}, les variables de réseau (ex. : temps de

déplacement, coût, fréquence, etc.), S_i, les variables socio-économiques associées à l'origine i (ex. : population, revenu, niveau d'emploi, etc.), et S_j les variables socio-économiques associées à la destination j. β_0 est un paramètre affecté à la constante du modèle β_1, β_2 et β_3 sont des vecteurs de coefficients fixes.

L'explication fournie à y_{ij} n'implique que l'information propre au marché ij. En effet ni les variables de réseau qui caractérisent les marchés compétitifs is (même origine mais destination différente) ou rj (même destination mais origine différente) ni les variables socio-économiques de zones autres que i et j n'apparaissent à l'équation (1). Il est raisonnable de supposer que les variables de réseau et socio-économiques de zones adjacentes au marché ij puissent influencer y_{ij}, mais l'inclusion de ces variables dans le modèle (1) conduirait à un nombre trop élevé de variables explicatives. Selon Cliff et Ord (1981), l'omission de variables explicatives reliées à la structure physique et économique de la région (distance entre les zones, taille des zones, la longueur de la frontière commune entre les zones adjacentes, etc...) est une source d'autocorrélation spatiale des erreurs résiduelles. Dès lors, une formulation adéquate pour traiter l'autocorrélation spatiale des erreurs résiduelles doit incorporer trois composantes : une reliée aux zones d'origine, une reliée aux zones de destination et une autre associée aux facteurs spécifiques du réseau.

2. LA FORMULATION À COMPOSANTES D'ERREUR AUTORÉGRESSIVES SPATIALE D'ORDRE 1

Nous proposons de définir une structure d'erreur qui tient compte des diverses sources de corrélation mentionnées :

$$\varepsilon_{ij} = \alpha_i + \lambda_j + u_{ij}. \tag{2}$$

Cette décomposition du flux inexpliqué entre les zones i et j est compatible avec la structure de la partie déterministe du modèle de régression. Le terme d'erreur α_i permet de capter les effets non mesurés associés à l'origine i (il contient les variables manquantes du vecteur S_i), λ_j contient les effets manquants reliés à la destination j (il contient les variables manquantes du vecteur S_j) et finalement u_{ij} incorpore les variables de réseau omises de la spécification du modèle. Nous supposons que tous ces termes sont de moyennes nulles et mutuellement non corrélés,

$$E(\alpha_i\lambda_j) = E(\alpha_i u_{ij}) = E(\lambda_i u_{ij}) = 0 \quad \forall i, j. \tag{3}$$

Chaque composante admet un processus autorégressif spatial normalisé d'ordre 1 (SAR(1)).

$$\alpha_i = \rho_1 \sum_{r=1}^{N} w_{1,ir}\alpha_r + \xi_i, \tag{4}$$

$$\lambda_j = \rho_2 \sum_{s=1}^{T} w_{2,js}\lambda_s + \zeta_j, \tag{5}$$

$$u_t = \rho_3 \sum_{l=1}^{L} w_{3,tl} u_l + v_t, \tag{6}$$

où t désigne un flux donné $(i \rightarrow j)$ et l tout autre flux $(r \rightarrow s)$ de l'étude. Nous supposons que $\xi_i \sim N(0,\sigma_\xi^2)$, $\zeta_j \sim N(0,\sigma_\zeta^2)$ et $\upsilon_t \sim N(0,\sigma_\upsilon^2)$ $\forall i, \forall j, \text{et} \forall t$, respectivement. Les équations (4)-(6) formalisent les interdépendances discutées précédemment. Notons qu'on obtient le modèle d'erreurs composées si $\rho_1 = \rho_2 = \rho_3 = 0$. Les processus sont normalisés de sorte que $\Sigma_r w_{1,ir} = 1, \forall i, \Sigma_s w_{2,js} = 1, \forall j$ et $\Sigma_l w_{3,tl} = 1, \forall t$. Le paramètre ρ_1 est un coefficient d'autocorrélation spatiale qui mesure le degré de dépendance linéaire entre α_i et chacun des « voisins » α_r en origine. Les paramètres ρ_2 et ρ_3 s'interprètent de façon analogue. Ces coefficients sont contraints à demeurer dans l'intervalle unitaire (-1,1) afin d'assurer la stabilité de chacun des processus. La stabilité du processus SAR(1) permet d'assurer que la fonction de vraisemblance soit continue sur le domaine défini par ρ_1, ρ_2 et ρ_3.

Les pondérations $w_{1,ir}$, $w_{2,js}$ et $w_{3,tl}$ sont des fonctions **estimables** qui doivent être spécifiées par l'analyste afin de décrire l'influence de chaque type d'erreur résiduelle sur les autres erreurs résiduelles du même type. Ces fonctions s'écrivent en général comme suit :

$$w_{a,bc} = \frac{w^*_{a,bc}}{\sum_c w^*_{a,bc}}, \quad w_{a,bb} = 0 \quad \forall b, \tag{7}$$

$$w^*_{a,bc} = g\left(R^1_{a,bc}, R^2_{a,bc}, \ldots, R^{L_1}_{a,bc}, \theta_{a1}, \theta_{a2}, \ldots, \theta_{aH_a}\right), w^*_{a,bb} = 0 \quad \forall b, \tag{8}$$

laquelle comporte H_a paramètres θ_{ah} et L_a variables $R^l_{a,bc}$. La pondération $w_{1,ir}$, en particulier, mesure l'importance que prend α_r dans l'explication de α_i. Plus $w_{1,ir}$ est grand, plus α_i et α_r sont reliés. Le processus est dit généralisé puisque les paramètres θ_{1h} sont estimés. Il faut noter qu'une pratique courante dans les études spatiales consiste à fixer arbitrairement la valeur de ces paramètres. Une définition simple de pondération est $w^*_{1,ir} = (d_{ir})^{-\theta}$ où $\theta > 0$ et où d_{ir} est la distance entre les zones i et r. Un poids plus fort est donné aux zones les plus rapprochées géographiquement. L'interprétation de ρ_1 dépend de la définition de proximité incorporée dans $w^*_{1,ir}$. Pour la spécification discutée précédemment, un $\rho_1 > 0$ implique que les erreurs de même signe sont géographiquement regroupées. Si nous avions défini $w^*_{1,ir} = (P_i P_r)^\theta, \theta > 0$, où P_i et P_r sont des populations des zones i et r, nous aurions trouvé les plus fortes corrélations parmi les paires d'erreurs associées aux zones les plus populeuses. Il est important de remarquer ici que pour les deux exemples donnés, la normalisation (7)

imposée sur les lignes de la matrice de contiguïté W formée par les fonctions à l'équation (8) rend ces fonctions invariantes par rapport aux unités de mesure de distance ou de population. Anselin (1989), Bivand (1984), Blommestein (1983) ou Cliff et Ord (1981) utilisent parfois des processus d'erreurs SAR(1) avec une structure de contiguité définie par l'entremise d'une matrice Booléenne, où $w^*_{1,ir} = 1$ si i et r sont contiguës et $w^*_{1,ir} = 0$ autrement.

2.1 *Quelques sous-modèles*

2.1.1 *La formulation de Brandsma et Ketellapper*

Une solution possible au problème de corrélation spatiale parmi les erreurs de flux est le processus autorégressif spatial d'ordre 2 proposé par Brandsma et Ketellapper (1979), lequel s'écrit comme suit :

$$u_t = \gamma_1 \sum^{L} w^o_{tl} u_l + \gamma_2 \sum^{L} w^d_{tl} u_l + \upsilon_t, \quad \upsilon_t \sim N(0,\sigma_\upsilon^2) \tag{9}$$

où les pondérations non normalisées sont définies comme

w^{o*}_{tl} = 1 si les flux t et l partagent la même origine,
= 0 autrement, et
w^{d*}_{tl} = 1 si les flux t et l partagent la même destination.
= 0 autrement.

Afin de vérifier que cette expression est bel et bien un cas spécial de EC-SAR(1), nous réécrivons l'équation (9) comme suit :

$$u_t = \gamma_1 \sum_{l=1}^{L} \left(w^o_{tl} + \frac{\gamma_2}{\gamma_1} w^d_{tl} \right) u_l + \upsilon_t, \quad \upsilon_t \sim N(0,\sigma_\upsilon^2). \tag{10}$$

Cette forme peut être interprétée comme appartenant à un processus générateur EC-SAR(1) avec $\sigma_\xi^2 = \sigma_\zeta^2 = 0$. Toutefois la spécification des poids associés à cette troisième composante du processus est quelque peu particulière car elle n'incorpore que des effets d'origine et de destination. La formulation BK peut être critiquée en ce sens qu'elle inclue les effets de corrélation spécifiques aux zones dans la mauvaise composante de l'erreur.

2.1.2 *La formulation de Bolduc, Dagenais et Gaudry*

Cette formulation prolonge celle de BK en incluant des termes supplémentaires permettant de capter des facteurs de réseaux non inclus dans la partie fixe du modèle. Leur structure de corrélation spatiale possède une forme SAR(1)

$$u_t = \rho \sum_{l=1}^{L} w_{tl} u_l + \upsilon_t, \quad \upsilon_t \sim N(0,\sigma_\upsilon^2), \tag{11}$$

où les pondérations non normalisées sont définies à partir de

$$w_{tt}^{*} = (d_{ir} + d_{js})^{-\theta_1} + (d_{is} + d_{rj})^{-\theta_2}, \text{ avec } w_{tt} = 0 \ \forall i. \tag{12}$$

Cette fonction prend en considération les effets directs et les effets croisés de distances entre deux origines et deux destinations. Pour une paire de flux donnée, disons $(i \rightarrow j)$ et $(r \rightarrow s)$, la première des deux composantes de l'équation (12) (l'effet direct) se réfère aux distances entre les deux origines i et r et les deux destinations j et s. Pour une valeur positive donnée de θ_1, les petites distances entre les origines et/ou les destinations impliquent une forte pondération w_{tt}. Ce premier terme est utilisé pour tenir compte d'un agglomérat de relations positives et négatives entre les flux individuels qui originent et qui terminent dans des zones spécifiées. Le second terme du membre droit de l'équation (12) représente les effets croisés de distance entre les origines et les destinations. La formulation implique que pour un $\theta_2 > 0$ donné, toutes choses étant égales par ailleurs, plus les distances sont grandes, plus le facteur de pondération w_{tt} est petit.

Ce processus d'erreur est clairement un cas spécifique de la spécification proposée ici. Même s'il incorpore les trois éléments majeurs pour expliquer l'autocorrélation spatiale, les effets de corrélation ne sont pas inclus dans la composante appropriée. Une façon plus naturelle de faire serait d'incorporer les effets d'origine dans la composante spécifique à la zone et ainsi de suite pour les deux autres effets. C'est ce que la formulation EC-SAR(1) permet de faire.

3. PROCÉDURE D'ESTIMATION

Dans le but de discuter des problèmes relatifs à l'estimation du modèle, nous utilisons l'écriture matricielle suivante.

$$y = X\beta + \varepsilon, \tag{13}$$

$$\varepsilon = C\alpha + D\lambda + u, \tag{14}$$

$$\alpha = \rho_1 W_1 \alpha + \xi, \ \xi \sim N(0, \sigma_\xi^2 I_N), \tag{15}$$

$$\lambda = \rho_2 W_2 \lambda + \zeta, \ \zeta \sim N(0, \sigma_\zeta^2 I_T), \tag{16}$$

$$u = \rho_3 W_3 u + \upsilon, \ \upsilon \sim N(0, \sigma_\upsilon^2 I_L), \tag{17}$$

où L représente le nombre total d'observations de l'échantillon. Les vecteurs y, ε, u et υ sont de dimension L x 1. Les vecteurs α et ξ sont de dimension N x 1 tandis que λ et ζ sont de dimension T x 1. X est une matrice L x K de régresseurs fixes β et est un vecteur K x 1 de paramètres. Les pondérations définies dans les équations (4)-(6) sont incorporées dans les matrices normalisées W_1, W_2 et W_3 de dimension N x N, T x T et L x L, respectivement.

Avec des données équilibrées, c'est-à-dire où à chaque origine correspond le même ensemble de destinations et vice versa, $C = (I_N \otimes e_T)$ et $D = (e_N \otimes I_T)$ où

e_T dénote un vecteur unitaire de dimension T x 1 et I_T une matrice identité de taille T. Avec des données non équilibrées, les matrices C et D doivent être redéfinies afin de tenir compte des flux manquants. Soit T_i le nombre de destinations associées à l'origine i qui est égal à $T - z_i$ où z_i représente le nombre de destinations manquantes à partir de l'origine i. Dénotons e_{T_i} comme un vecteur unitaire T_i x 1 et $I_{(T_i)}$ comme une matrice identité de taille T x T où les rangées associées aux destinations manquantes à partir de l'origine i sont éliminées. Alors nous pouvons vérifier que C et D avec un échantillon non balancé correspondent à

$$C = \begin{bmatrix} e_{T_1} & 0 \ \dots \ 0 \\ 0 & e_{T_2} \ \dots \ 0 \\ . & . \ \dots \ . \\ . & . \ \dots \ . \\ 0 & 0 \ \dots \ e_{T_N} \end{bmatrix}_{L \times N} \quad \text{et} \quad D = \begin{bmatrix} I_{(T_1)} \\ I_{(T_2)} \\ . \\ . \\ I_{(T_N)} \end{bmatrix}_{L \times T}$$

Les équations (15)-(17) peuvent être réécrites comme :

$$\alpha = (I_N - \rho_1 W_1)^{-1} \xi = P_1^{-1} \xi, \tag{18}$$

$$\lambda = (I_T - \rho_2 W_2)^{-1} \zeta = P_2^{-1} \zeta, \tag{19}$$

$$u = (I_L - \rho_3 W_3)^{-1} \upsilon = P_3^{-1} \upsilon, \tag{20}$$

où nous utilisons $P_1 = I_N - \rho_1 W_1$, $P_2 = I_T - \rho_2 W_2$ et $P_3 = I_L - \rho_3 W_3$.

En substituant les équations (18)-(20) dans l'équation (14), le modèle linéaire de flux avec un processus résiduel EC-SAR(1) peut s'écrire :

$$y = X\beta + \varepsilon, \tag{21}$$

$$\varepsilon = CP_1^{-1}\xi + DP_2^{-1}\zeta + P_3^{-1}\upsilon. \tag{22}$$

Nos hypothèses de distribution impliquent que $\varepsilon \sim N(0, \Phi)$, où

$$\Phi = \sigma_\xi^2 CP_1^{-1}P_1^{-1}C' + \sigma_\zeta^2 DP_2^{-1}P_2^{-1}D' + \sigma_\upsilon^2 P_3^{-1}P_3^{-1}.$$

Dans la discussion qui suit nous considérons que les données sont équilibrées. Pour un simple problème avec 20 origines et 20 destinations, par exemple, Φ et P_3 sont de dimension 400 x 400. L'estimation du modèle (21)-(22) par la méthode du maximum de vraisemblance nécessite l'inversion des matrices Φ et P_3 à chaque itération du processus d'optimisation. Même si le traitement de matrices 400 x 400 est réalisable du point de vue numérique, cela peut en être autrement pour des problèmes de plus grande taille. Dans ce qui suit, nous réécrivons la formulation du modèle sous une écriture équivalente nécessitant l'inversion d'une seule matrice de taille L x L et nous montrons de plus que cet inverse peut être calculé avec l'inversion d'une matrice $(N + T)$ x $(N + T)$. Ceci peut représenter un gain énorme du point de vue du calcul numérique.

Premièrement, nous réécrivons le modèle ainsi :

$$y = X\beta + \varepsilon, \tag{23}$$

$$\varepsilon = Ma + P_3^{-1}\upsilon,\ \upsilon \sim N(0, \sigma_\upsilon^2 I_L), \tag{24}$$

où $M = \left[\sigma_\xi C P_1^{-1} \middle| \sigma_\zeta D P_2^{-1}\right]$ est une matrice de dimension L x $(N + T)$. L'opérateur | désigne la concaténation horizontale de deux matrices. Le vecteur a est de dimension $N + T$ avec comme composantes des éléments indépendants, $a \sim N(0, I_{N+T})$. Définissons $M^* = P_3 M$, alors

$$\varepsilon = P_3^{-1}\left[P_3 Ma + \upsilon\right] = P_3^{-1}\left[M^* a + \upsilon\right] = P_3^{-1}\omega, \text{et}$$

$$\text{cov}(\varepsilon) = \Phi = P_3^{-1}\Omega P_3^{-1\prime}, \tag{25}$$

où $\omega \sim N(0,\Omega)$ avec $\Omega = \left[M^* M^{*\prime} + \sigma_\upsilon^2 I_L\right]$. L'inverse de Φ est égale à $\Phi^{-1} = P'_3 \Omega^{-1} P_3$, une matrice de dimension L x L. Toutefois,

$$\Omega^{-1} = \sigma_\upsilon^{-2}\left[I_L - M^* V^{-1} M^{*\prime}\right], \tag{26}$$

où $V = [\sigma_\upsilon^2 I_{N+T} + M^{*\prime} M^*]$, est une matrice carrée de dimension $N + T$. Le calcul du déterminant de Ω peut aussi être simplifié car

$$|\Omega| = (\sigma_\upsilon^2)^{L-(N+T)}|V|. \tag{27}$$

La fonction de vraisemblance du modèle (23)-(24) s'écrit :

$$l(\beta, \gamma | y) = (2\pi)^{-L/2}|P_3||\Omega|^{-1/2}\exp\left\{-\frac{(y - X\beta)' P_3' \Omega^{-1} P_3 (y - X\beta)}{2}\right\},$$

$$= (2\pi)^{-L/2}|P_3||\Omega|^{-1/2}\exp\left\{-\frac{\omega' \Omega^{-1}\omega}{2}\right\}, \tag{28}$$

où le vecteur γ représente les paramètres contenus dans la matrice Φ. La fonction de vraisemblance correspondante exprimée sous forme logarithmique s'écrit finalement :

$$L(\beta, \gamma | y) = -\frac{L}{2}\ln(2\prod) + \ln|P_3| - \frac{1}{2}\ln|\Omega| - \frac{\omega' \Omega^{-1}\omega}{2}. \tag{29}$$

Le lecteur peut vérifier que la fonction (29) peut être concentrée par rapport à β. Ceci ramène alors le problème à une maximisation par rapport au vecteur γ seulement.

CONCLUSION

Dans le but de traiter le problème de corrélation spatiale parmi les erreurs d'un modèle de régression de données de flux, nous avons proposé un nouveau

processus de génération d'erreurs basé sur la théorie des composantes d'erreurs et sur les processus autorégressifs spatiaux d'ordre 1. Le terme d'erreur résiduelle se subdivise en une somme d'une composante reliée à la zone d'origine, une composante reliée à la zone de destination et un reste. Chaque composante résulte d'un processus générateur autorégressif spatial d'ordre 1. Nous proposons une approche efficace d'estimation basée sur le maximum de vraisemblance qui permet l'application de la méthodologie même dans le contexte d'échantillons de grande taille.

BIBLIOGRAPHIE

ANSELIN, L. (1989), *Spatial Econometrics: Methods and Models*, Kluwer Academic Publishers.

BIVAND, R.S. (1984), « Regression Modeling with Spatial Dependence : an Application of Some Class Selection and Estimation Methods », *Geographical Analysis*, 16, 1, Janvier : 25-37.

BLOMMESTEIN, H.J. (1983), « Specification and Estimation of Spatial Econometric Models. A Discussion of Alternative Strategies for Spatial Economic Modeling », *Regional Science and Urban Economics*, 13 : 251-270.

BOLDUC, D., M. G. DAGENAIS et M.J. GAUDRY (1989), « Spatially Autocorrelated Errors in Origin-Destination Models : a New Specification Applied to Urban Travel Demand in Winnipeg », *Transportation Research*, B, 23, 5 : 361-372.

BRANDSMA, A.S., et R.H. KETELLAPPER (1979), « A Biparametric Approach to Spatial Autocorrelation », *Environment and Planning*, A, 11 : 51-58.

CLIFF A.D., et J.K. ORD (1981), *Spatial Processes, Models and Application*, Pion, Londres.