Techniques d'enquête, décembre 1990
Vol. 16, n° 2, pp. 229–249
Statistique Canada

Estimation robuste pour petits domaines par la combinaison de données chronologiques et transversales

D. PFEFFERMANN et L. BURCK[1]

RÉSUMÉ

Dans l'estimation pour petits domaines (ou petites régions), on cherche le plus souvent à exploiter le caractère transversal des données de manière que l'information relative à une petite région serve pour d'autres petites régions. Par ailleurs, dans le cas des enquêtes à passages répétés, il est possible d'accroître l'efficacité de l'estimation en modélisant aussi les propriétés temporelles des données. Nous expliquons notre pensée en considérant des modèles de régression à coefficients corrélés de façon transversale et qui varient de façon longitudinale. L'emploi de données de périodes antérieures pour estimer des moyennes de périodes courantes nous amène à nous interroger sur la façon de prévenir les défaillances de modèle. Dans cet article, nous proposons des modifications pour que les prédicteurs de moyennes globales de petites régions (fondés sur un modèle) concordent avec les estimateurs d'enquête correspondants et nous examinons les propriétés statistiques de ces modifications. Nous appliquons ensuite la nouvelle méthode à des données sur le prix de vente des maisons, qui servent au calcul des indices des prix du logement.

MOTS CLÉS: Filtre de Kalman; contraintes linéaires; modèles d'espace d'états.

1. INTRODUCTION

Les organismes nationaux de statistique sont souvent appelés à construire des estimateurs fiables pour les moyennes de petites régions mais la taille des échantillons à l'intérieur de ces régions est habituellement trop faible pour que l'on puisse utiliser des estimateurs directs. C'est pourquoi on a proposé ces dernières années de nouveaux estimateurs qui combinent les données d'enquête de toutes les petites régions avec de l'information supplémentaire (tirée d'un recensement ou de dossiers administratifs). Ces estimateurs ont tous pour caractéristique de pouvoir être construits, en règle générale, comme une combinaison linéaire de deux éléments: un "estimateur synthétique" de forme $\bar{X}_i'\hat{\beta}$, où $\bar{X}_i$ représente l'information supplémentaire moyenne pour une petite région et $\hat{\beta}$ est un vecteur de coefficients de régression estimés, et un "facteur de correction" de forme $(\bar{y}_i - \bar{x}_i\hat{\beta})$, où $\bar{y}_i$ et $\bar{x}_i$ sont les moyennes observées de la variable étudiée et de la variable auxiliaire respectivement. Les facteurs de correction servent à expliquer la variation des moyennes de petites régions que n'expliquent pas les variables auxiliaires. Les divers estimateurs se distinguent principalement les uns des autres par la méthode de calcul des poids attribués aux deux éléments de la combinaison linéaire; ces méthodes vont de l'"approche fondée sur un plan" (Särndal et Hidiroglou 1989) à la "méthode empirique de Bayes" (Fay et Herriot 1979), en passant par les "modèles linéaires composés" (Battese, Harter et Fuller 1989; Pfeffermann et Barnard 1991).

Il existe très peu d'ouvrages en statistique où l'on examine la possibilité d'exploiter l'aspect chronologique des données pour accroître davantage l'efficience des estimateurs pour petites régions et ce malgré le fait que bon nombre de ces estimateurs viennent d'enquêtes à passages répétés comme les enquêtes sur la population active. La littérature économétrique renferme un très grand nombre d'études portant sur la modélisation combinée de données chronologiques et de données transversales; voir, par exemple, Rosenberg (1973b), Johnson (1977, 1980), Maddala (1977, chapitre 7), Dielman (1983) et Pfeffermann et Smith (1985) pour des analyses.

[1] D. Pfeffermann, Department of Statistics, Hebrew University, Jerusalem 91905. L. Burck, Unit for Statistical Analysis, Central Bureau of Statistics, Jerusalem 91130.

Cependant, aucune de ces études n'a pour objet l'estimation (la prédiction) de moyennes de petites régions à partir de données d'enquête. Certains ouvrages portant sur l'estimation de moyennes de population ont traité de l'ajustement de modèles chronologiques à des données d'enquête; voir à ce propos les articles de synthèse de Smith (1979) et de Binder et Hidiroglou (1988) ainsi que les articles plus récents de Binder et Dick (1989), de Tiller (1989) et de Pfeffermann (1991). Pourtant, ces méthodes ne sont pas très répandues parce que les estimateurs classiques de moyennes globales sont souvent presque aussi efficaces lorsque les modèles sont valides ou plus robustes lorsque les modèles s'avèrent non fondés.

La situation est tout à fait différente lorsqu'il s'agit d'estimation pour petites régions et nous sommes d'avis que l'utilisation de modèles chronologiques peut être très avantageuse dans ce cas. Bien que la nature exacte du modèle à utiliser pour une application particulière dépende évidemment des données, la classe de modèles que nous examinons dans la section suivante est suffisamment large pour nous permettre de résoudre bon nombre, et même la plupart, des problèmes ayant trait à l'estimation pour petites régions qui surviennent en pratique. En outre, l'estimation de ces modèles a l'avantage d'être relativement simple (voir section 3).

Lorsqu'on utilise un modèle, on est toujours amené à se demander comment se prémunir contre sa défaillance et cette question devient d'autant plus délicate lorsqu'il s'agit d'un modèle destiné à la production de statistiques officielles. Dans la section 4, nous nous penchons sur le problème et proposons de modifier les prédicteurs (fondés sur un modèle) de moyennes globales de petites régions de telle sorte qu'ils concordent avec les estimateurs d'enquête correspondants, pourvu que ceux-ci soient fiables. Nous examinons ensuite les propriétés statistiques des nouveaux prédicteurs. Enfin, dans la section 5, nous présentons des résultats empiriques de l'application du modèle selon deux scénarios: avec prédicteur modifié et sans prédicteur modifié. Les données utilisées dans les circonstances sont les prix de vente des maisons enregistrés à Jérusalem entre septembre 1985 et juin 1989. Le Bureau central de la statistique, en Israël, se sert régulièrement de ces données pour calculer les indices des prix du logement.

2. RÉGRESSION AVEC DES COEFFICIENTS QUI VARIENT DE FAÇON TRANSVERSALE ET LONGITUDINALE

2.1 Une classe générale de modèles

Pour les besoins de notre exposé, nous allons désigner par $\underset{\sim}{Y}_{tk}$ le vecteur $n_{tk} \times 1$ des valeurs observées de la variable expliquée Y pour la région k à la période t, $k = 1, \ldots, K$, $t = 1, 2, \ldots$. Pour des raisons de commodité, nous supposons que $n_{tk} \geq 1$ mais nous verrons plus loin que le modèle prévoit l'absence d'observations pour des régions à certaines périodes. Posons X_{tk} comme la matrice de plan $n_{tk} \times (p + 1)$ des variables auxiliaires, dont la première colonne est formée de uns. Dans de nombreuses applications, le même vecteur ligne $\underset{\sim}{x}'_{tk}$ de valeurs auxiliaires s'applique à toutes les valeurs observées de Y pour une période donnée, auquel cas $X_{tk} = \underset{\sim}{1}_{n_{tk}} \underset{\sim}{x}'_{tk}$, où $\underset{\sim}{1}_{n_{tk}}$ est un vecteur colonne formé de uns et de dimension n_{tk}. Cela se produit lorsque les estimateurs d'enquête pour petites régions sont les seules données dont nous disposions. À cause des exigences du secret statistique et des frais de traitement, on peut difficilement utiliser les micro-données provenant d'enquêtes menées auprès de particuliers. La théorie que nous exposons ici n'est pas contrainte à la disponibilité de micro-données (voir l'exemple à la section 2.2) mais le fait qu'il existe ou non des données a un effet indéniable sur la spécification des modèles et la précision de l'estimation.

Nous définissons comme suit le modèle de régression pour la région k et la période t:

$$\underset{\sim}{Y}_{tk} = X_{tk} \underset{\sim}{\beta}_{tk} + \underset{\sim}{\varepsilon}_{tk};\ E(\underset{\sim}{\varepsilon}_{tk}) = 0,\ E(\underset{\sim}{\varepsilon}_{tk} \underset{\sim}{\varepsilon}'_{tk}) = \sigma_k^2 I_{n_{tk}} \tag{2.1}$$

où $\underset{\sim}{\beta}'_{tk} = (\beta_{tk0}, \beta_{tk1}, \ldots, \beta_{tkp})$.

De même, la moyenne (de superpopulation) des valeurs de la variable étudiée pour la région k à la période t est définie

$$\Theta_{tk} = E(M_{tk} \mid \underset{\sim}{\beta}_{tk}) = \underset{\sim}{\bar{X}}_{tk}\underset{\sim}{\beta}_{tk} \tag{2.2}$$

où

$$M_{tk} = \frac{1}{N_{tk}} \sum_{i=1}^{N_{tk}} Y_{tki} \quad \text{et} \quad \underset{\sim}{\bar{X}}_{tk} = \frac{1}{N_{tk}} \sum_{i=1}^{N_{tk}} \underset{\sim}{x}'_{tki}$$

avec $i = 1, \ldots, N_{tk}$ désignant les unités de la population. Évidemment, lorsque $\underset{\sim}{x}'_{tki} \equiv \underset{\sim}{x}'_{tk}$, alors $\underset{\sim}{\bar{X}}_{tk} = \underset{\sim}{x}'_{tk}$.

Soit $\hat{\underset{\sim}{\beta}}_{tk}$ l'estimateur de $\underset{\sim}{\beta}_{tk}$. Par conséquent, $\hat{\Theta}_{tk} = \underset{\sim}{\bar{X}}_{tk}\hat{\underset{\sim}{\beta}}_{tk}$ et

$$\hat{M}_{tk} = \frac{1}{N_{tk}}\left[\sum_{i=1}^{n_{tk}} Y_{tki} + \sum_{i=n_{tk}+1}^{N_{tk}} \underset{\sim}{x}'_{tki}\hat{\underset{\sim}{\beta}}_{tk}\right] = \hat{\Theta}_{tk} + \frac{1}{N_{tk}}\left(\sum_{i=1}^{n_{tk}} (Y_{tki} - \underset{\sim}{x}'_{tki}\hat{\underset{\sim}{\beta}}_{tk})\right)$$

ce qui signifie que, dans l'hypothèse habituelle de faibles taux de sondage à l'intérieur des régions, $\hat{\Theta}_{tk}$ peut être considéré aussi bien comme un estimateur de la moyenne de population finie M_{tk}. C'est pourquoi nous ne ferons plus désormais la distinction entre la moyenne de la population finie et celle de la superpopulation.

Une caractéristique notable du modèle (2.1) est que les coefficients $\underset{\sim}{\beta}_{tk}$ peuvent varier de façon transversale et longitudinale. Les équations ci-dessous décrivent la variation longitudinale des coefficients:

$$\begin{bmatrix}\beta_{tkj} \\ \beta_{kj}\end{bmatrix} = T_j \begin{bmatrix}\beta_{t-1,kj} \\ \beta_{kj}\end{bmatrix} + \begin{bmatrix}1 \\ 0\end{bmatrix}\eta_{tkj}, \; j = 0, \ldots, p \tag{2.3}$$

dans ces équations, β_{kj}, $j = 0, 1, \ldots, p$, représente des coefficients fixes, que nous interprétons plus bas, et T_j représente des matrices fixes de dimension (2×2) et où les résidus $\{\eta_{tkj}\}$ satisfont les équations

$$E(\eta_{tkj}) = 0, \; E(\eta_{tkj}\eta_{tk\ell}) = \delta_{j\ell}, \; E(\eta_{tkj}\eta_{t-d,k\ell}) = 0 \quad \text{pour} \quad d > 0. \tag{2.4}$$

L'expression (2.4) implique que les résidus de coefficients différents se rapportant à la même période t peuvent être corrélés mais que l'autocorrélation et l'intercorrélation sériale sont supposées nulles.

Considérons maintenant quelques exemples d'utilisation de (2.3) assez simples:

(a) $T_j = \left[\begin{smallmatrix}0 & 1 \\ 0 & 1\end{smallmatrix}\right]$ implique que $\beta_{tkj} = \beta_{kj} + \eta_{tkj}$, de sorte que β_{kj} représente en l'occurrence une moyenne commune. On reconnaît là le fameux modèle de régression à coefficients aléatoires (Swamy 1971), que l'on retrouve souvent dans les applications économétriques. De toute évidence, si l'on pose comme hypothèse $\text{var}(\eta_{tkj}) = 0$, le modèle se ramène à un modèle à coefficients fixes.

(b) $T_j = \left[\begin{smallmatrix}1 & 0 \\ 0 & 0\end{smallmatrix}\right]$ implique que $\beta_{tkj} = \beta_{t-1,kj} + \eta_{tkj}$; on reconnaît là le modèle de trajet aléatoire. (Voir, par exemple, Cooley et Prescott (1976) ainsi que LaMotte et McWhorter (1977) pour des exemples d'application de ce modèle dans des études économétriques.) Dans ce cas précis, le coefficient β_{kj} est évidemment superflu et devrait donc être omis de sorte que $T_j \equiv 1$.

(c) $T_j = \begin{bmatrix} \rho & 1-\rho \\ 0 & 1 \end{bmatrix}$ suppose la relation autorégressive du premier ordre $(\beta_{tkj} - \beta_{kj}) = \rho(\beta_{t-1,kj} - \beta_{kj}) + \eta_{tkj}$, envisagée par Rosenberg (1973a).

(d) $T_j = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$ implique que $\beta_{tkj} = \beta_{t-1,kj} + \beta_{kj} + \eta_{tkj}$, ce qui représente une approximation locale d'une tendance linéaire (Kitagawa et Gersch 1984). Le coefficient β_{kj} désigne en l'occurrence une pente fixe.

Il convient de souligner que l'on peut se servir de différentes matrices T_j pour différents coefficients β_{tkj}. De fait, en posant $\underset{\sim}{\alpha}'_{tk} = (\beta_{tk0}, \beta_{k0}, \beta_{tk1}, \beta_{k1}, \ldots, \beta_{tkp}, \beta_{kp})$; $\tilde{T} = \text{diag}[T_0, T_1, \ldots, T_p]$, une matrice diagonale par blocs, où T_j désigne le bloc j; $\tilde{G} = I_{p+1} \otimes \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, où I_{p+1} est la matrice unité d'ordre $p + 1$ et $\otimes$ désigne le produit de Kronecker; et $\underset{\sim}{\eta}'_{tk} = (\eta_{tk0}, \eta_{tk1}, \ldots, \eta_{tkp})$, nous pouvons formuler comme suit le modèle composé valant pour les coefficients $\underset{\sim}{\beta}_{tk}$:

$$\underset{\sim}{\alpha}_{tk} = \tilde{T}\underset{\sim}{\alpha}_{t-1,k} + \tilde{G}\underset{\sim}{\eta}_{tk}; \quad E(\underset{\sim}{\eta}_{tk}) = \underset{\sim}{0}, E(\underset{\sim}{\eta}_{tk}\underset{\sim}{\eta}'_{t-d,k}) = A_d\Delta \tag{2.5}$$

où A_d égale 1 si $d = 0$ et est nul si $d \neq 0$, et où $\Delta = [\delta_{ij}]$ est définie par les variances et covariances δ_{ij} (équation 2.4).

Le modèle défini en (2.5) décrit la variation longitudinale des coefficients de régression pour une région donnée. Ordinairement, les modèles conçus pour tenir compte de la corrélation transversale des moyennes de petites régions renferment des effets aléatoires de petit domaine $\{u_k\}$ qui ne varient pas dans le temps. On peut introduire de tels effets dans le modèle général défini en (2.1) et en (2.3) en posant, par exemple, $\underset{\sim}{Y}_{tk} = \underset{\sim}{1}_{n_{tk}}u_{tk} + X_{tk}\underset{\sim}{\beta}_{tk} + \underset{\sim}{\epsilon}_{tk} = X^*_{tk}\underset{\sim}{\beta}^*_{tk} + \underset{\sim}{\epsilon}_{tk}$ et $u_{tk} = u_{t-1,k} + \eta_{tk}$, où $u_{ok} = 0$, $\text{var}(\eta_{1k}) = \sigma^2_\eta$ et $\text{var}(\eta_{tk}) = 0$ pour $t > 1$ (on peut comparer cette dernière équation avec l'exemple (b) ci-dessus). De plus, en supposant la relation autorégressive de l'exemple (c) pour la variable en cause et en tenant pour fixes les autres coefficients de régression (exemple (a) avec variance des résidus nulle), on obtient un modèle semblable à celui analysé par Choudhry et Rao (1989) sauf que dans ce dernier cas, les résidus d'observations de l'équation (2.1) peuvent être autocorrélés. Notons que l'équation (2.1) renferme maintenant deux ordonnées à l'origine aléatoires mais le modèle demeure identifiable. Choudhry et Rao supposent que les estimateurs d'enquête sont les seules données disponibles; par conséquent, il faudra recourir aux micro-données pour estimer les autocorrélations. Par ailleurs, il est toujours possible de définir un modèle qui tienne compte des autocorrélations. Pour leur part, Choudhry et Rao supposent un modèle AR(1).

Une façon plus générale de tenir compte de la corrélation transversale des moyennes de petites régions est de permettre l'existence d'une corrélation entre les résidus η_{tkj} et η_{tmj} des modèles qui décrivent la variation longitudinale des coefficients de régression β_{tkj} et β_{tmj} pour les régions k et m (équation 2.4). En règle générale, on peut raisonnablement supposer que la corrélation diminue à mesure que s'accroît la distance entre les régions. Cette relation peut être exprimée au moyen de l'équation $E(\eta_{tkj}, \eta_{tmj}) = \delta_{jj}\rho_j f_j(k,m)$, $k \neq m$, où $f_j(k,m)$ est une fonction monotone décroissante des distances $D(k,m)$. La décroissance géométrique de la corrélation est exprimée par l'équation $f_j(k,m) = \rho_j^{|k-m|-1}$. Enfin, pour exprimer une corrélation fixe, nous avons la relation $f_j(k,m) \equiv 1$. C'est sur ce dernier cas que nous concentrons notre attention maintenant. L'hypothèse d'une corrélation transversale fixe pour tous les coefficients de régression peut être traduite par l'équation

$$E(\underset{\sim}{\eta}_{tk}\underset{\sim}{\eta}'_{tm}) = D(\Delta)\emptyset, \quad k \neq m \tag{2.6}$$

où $D(\Delta)$ est une matrice diagonale dont la diagonale principale est formée des variances δ_{jj} et $\emptyset$ est une autre matrice diagonale formée des coefficients de corrélation ρ_j.

Avant de terminer cette section, nous allons exprimer le modèle défini en (2.1), (2.5) et (2.6) sous forme de modèle d'espace d'états. Ce type de présentation offre des avantages considérables sur le plan du calcul.

Soit $\underset{\sim}{Y}_t' = (\underset{\sim}{Y}_{t1}', \ldots, \underset{\sim}{Y}_{tK}')$ le vecteur de dimension $n_t = \sum_k n_{tk}$ des observations pour toutes les régions au temps t et $\underset{\sim}{\epsilon}_t' = (\underset{\sim}{\epsilon}_{t1}', \ldots, \underset{\sim}{\epsilon}_{tK}')$, les résidus de la régression correspondants. Définissons $Z_{tk} = [\underset{\sim}{1}_{ntk}, \underset{\sim}{0}_{ntk}, \underset{\sim}{x}_{tk1}, \underset{\sim}{0}_{ntk}, \ldots, \underset{\sim}{x}_{tkp}, \underset{\sim}{0}_{ntk}]$, où $\underset{\sim}{0}_{ntk}$ est un vecteur de dimension n_{tk} formé de zéros et $\underset{\sim}{x}_{tkj}$ est le vecteur des valeurs de la variable auxiliaire j, $j = 1, \ldots, p$. Soit Z_t la matrice diagonale par blocs formée des matrices Z_{tk}. Cette matrice est d'ordre $n_t \times [K \times 2 \times (p+1)]$. Définissons aussi $\underset{\sim}{\alpha}_t' = (\underset{\sim}{\alpha}_{t1}', \ldots, \underset{\sim}{\alpha}_{tK}')$, $\underset{\sim}{\eta}_t' = (\underset{\sim}{\eta}_{t1}', \ldots, \underset{\sim}{\eta}_{tK}')$, $\Sigma_t = \text{Diag}\,[\sigma_1^2 \underset{\sim}{1}_{nt1}', \ldots, \sigma_K^2 \underset{\sim}{1}_{ntK}']$, $T = I_K \otimes \tilde{T}$, et $G = I_K \otimes \tilde{G}$.

En nous servant de ces définitions, nous pouvons réexprimer le modèle défini en (2.1), (2.5) et (2.6) sous la forme concise suivante:

$$\underset{\sim}{Y}_t = Z_t \underset{\sim}{\alpha}_t + \underset{\sim}{\epsilon}_t; \; E(\underset{\sim}{\epsilon}_t) = \underset{\sim}{0}, E(\underset{\sim}{\epsilon}_t \underset{\sim}{\epsilon}_t') = \Sigma_t \tag{2.7}$$

$$\underset{\sim}{\alpha}_t = T\underset{\sim}{\alpha}_{t-1} + G\underset{\sim}{\eta}_t; \; E(\underset{\sim}{\eta}_t) = \underset{\sim}{0}, E(\underset{\sim}{\eta}_t \underset{\sim}{\eta}_t') = \Lambda, \tag{2.8}$$

où $\Lambda = [\Lambda_{k\ell}]$, $k, \ell = 1, \ldots, K$, $\Lambda_{k\ell} = \Delta$ lorsque $k = \ell$ et $\Lambda_{k\ell} = D(\Delta)\emptyset$ lorsque $k \neq \ell$. Les matrices $\Lambda_{k\ell}$ sont de dimension $(p+1) \times (p+1)$.

Le modèle défini en (2.7) et (2.8) est une représentation classique d'un modèle d'espace d'états; voir, par exemple, Anderson et Moore (1979) et Harvey (1984). Selon la formulation des modèles d'espace d'états, (2.7) est l'équation d'observation et (2.8), l'équation d'état, où $\underset{\sim}{\alpha}_t$ désigne le vecteur d'états. L'avantage évident de cette formulation est que l'on peut estimer correctement les vecteurs $\underset{\sim}{\alpha}_t$ et de là, les moyennes de population Θ_{tk}, ainsi que la variance des erreurs d'estimation au moyen du filtre de Kalman. Nous nous arrêtons à l'utilisation de ce filtre dans les sections 3 et 4.

2.2 Estimateurs explicites des moyennes de petites régions

Afin de montrer comment, selon le modèle ci-dessus, les données de périodes antérieures ou de régions avoisinantes servent à "renforcer" les estimateurs pour petites régions, nous examinons le cas où le même vecteur $\underset{\sim}{x}_{tk}$ de valeurs de variables auxiliaires s'applique à toutes les unités d'une région donnée à une période donnée. Dans ce cas, nous pouvons exprimer l'équation d'observation en fonction des moyennes empiriques, c.-à-d.

$$\bar{Y}_{tk} = \underset{\sim}{x}_{tk}' \underset{\sim}{\beta}_{tk} + \bar{\epsilon}_{tk}; \; E(\bar{\epsilon}_{tk}) = 0, \; E(\bar{\epsilon}_{tk}^2) = \sigma_k^2 / n_{tk}, \; k = 1, \ldots, K. \tag{2.9}$$

Supposons que les coefficients de régression suivent un trajet aléatoire (exemple (b) de l'équation 2.3) tel que

$$\beta_{tkj} = \beta_{t-1,kj} + \eta_{tkj}; \; E(\eta_{tkj}) = 0, \; E(\eta_{tkj}\eta_{tk\ell}) = \delta_{j\ell}, j, \ell = 1, \ldots, p \tag{2.10}$$

et pour $k \neq m$,

$$E(\eta_{tkj}\eta_{tmj}) = \delta_{jj}\rho_j; \quad E(\eta_{tkj}\eta_{tm\ell}) = 0, j \neq \ell. \tag{2.11}$$

Le modèle de trajet aléatoire suppose que les coefficients s'éloignent lentement de leur valeur initiale et ne tendent pas naturellement à atteindre une valeur moyenne. Évidemment, pour des résidus η_{tkj} tels que $E(\eta_{tkj}^2) = 0$, les coefficients de régression correspondants ne varient pas dans le temps. Notons aussi que puisque $\underset{\sim}{\beta}_{tk} = \underset{\sim}{\beta}_{t-1,k} + \underset{\sim}{\eta}_{tk}$, le prédicteur de $\underset{\sim}{\beta}_{tk}$ au temps $(t-1)$ est identique au prédicteur $\hat{\underset{\sim}{\beta}}_{t-1,k}$ de $\underset{\sim}{\beta}_{t-1,k}$.

En nous servant des équations du filtre de Kalman définies dans la section 3, nous montrons en annexe que l'estimateur $\hat{\Theta}_{tk}$ de la moyenne de petite région Θ_{tk} (équation 2.2) peut avoir dans ce cas la forme suivante:

$$\hat{\Theta}_{tk} = \underset{\sim}{x}'_{tk}\hat{\underset{\sim}{\beta}}_{t-1,k} + \left(1 - \frac{\sigma_k^2}{n_{tk}v_k^2}\right)(\bar{Y}_{tk} - \underset{\sim}{x}'_{tk}\hat{\underset{\sim}{\beta}}_{t-1,k}) + \frac{\sigma_k^2}{n_{tk}v_k^2}\sum_{\substack{m=1 \\ m\neq k}}^{K}\gamma_{km}(\bar{Y}_{tm} - \underset{\sim}{x}'_{tm}\hat{\underset{\sim}{\beta}}_{t-1,m}) \tag{2.12}$$

où $\{\gamma_{km}\}$ représente les coefficients de régression partielle dans la régression de $e_{tk} = (\bar{Y}_{tk} - \underset{\sim}{x}'_{tk}\hat{\underset{\sim}{\beta}}_{t-1,k})$ par rapport aux erreurs de prédiction $\{e_{tm} = (\bar{Y}_{tm} - \underset{\sim}{x}'_{tm}\hat{\underset{\sim}{\beta}}_{t-1,m})\}$ observées dans les autres régions, et v_k^2 est la variance des résidus (inexpliquée) de la régression.

L'estimateur $\hat{\Theta}_{tk}$ renferme trois composantes: un estimateur "synthétique", $\underset{\sim}{x}'_{tk}\hat{\underset{\sim}{\beta}}_{t-1,k}$, où $\hat{\underset{\sim}{\beta}}_{t-1,k}$ est le prédicteur optimal de $\underset{\sim}{\beta}_{tk}$ fondé sur toutes les observations enregistrées jusqu'à la période $t-1$ inclusivement; un "facteur de correction", $(\bar{Y}_{tk} - \underset{\sim}{x}'_{tk}\hat{\underset{\sim}{\beta}}_{t-1,k})$, fondé sur l'erreur de prédiction pour la région k, et un "facteur d'ajustement" fondé sur les erreurs de prédiction observées pour les autres régions. Les deux premières composantes correspondent à celles de l'estimateur classique pour petite région décrit dans l'introduction. Notons que plus la taille de l'échantillon, n_{tk}, est petite, plus le poids attribué à la moyenne empirique courante $\bar{Y}_{tk}$ pour l'estimation de Θ_{tk} est faible et plus celui attribué au prédicteur chronologique $\underset{\sim}{x}'_{tk}\hat{\underset{\sim}{\beta}}_{t-1,k}$ est élevé. Le troisième élément du membre de droite de l'équation (2.12) représente l'information "empruntée" aux régions avoisinantes. Le poids imputé à cet élément dépend du degré de corrélation ρ_j entre les termes d'erreur correspondants $\{\eta_{tkj}\}$ des modèles pour les coefficients de régression (équation 2.11). Évidemment, lorsque les variables explicatives dans les diverses régions sont indépendantes de telle sorte que $\rho_j = 0$ pour tous j et, partant, $\gamma_{km} = 0$ pour tous m, le troisième élément disparaît et le prédicteur $\hat{\Theta}_{tk}$ se réduit à une moyenne pondérée de la moyenne courante $\bar{Y}_{tk}$ et du prédicteur chronologique $\underset{\sim}{x}'_{tk}\hat{\underset{\sim}{\beta}}_{t-1,k}$.

3. ESTIMATION ET INITIALISATION DE MODÈLE AU MOYEN DU FILTRE DE KALMAN

3.1 Estimation des coefficients de régression au moyen du filtre de Kalman

Dans cette section, nous présentons les équations du filtre de Kalman servant à la mise à jour et au lissage des vecteurs d'états $\underset{\sim}{\alpha}_t$ définis par les équations (2.7) et (2.8) (dans le cas qui nous occupe, il s'agit des coefficients de régression de région). Nous supposons que les matrices de variances-covariances Σ_t et Λ sont connues. L'estimation de ces matrices fait l'objet de la section 3.2. La théorie du filtre de Kalman est exposée dans de nombreux ouvrages (voir, par exemple, Anderson et Moore 1979 et Meinhold et Singpurwalla 1983); nous limiterons donc notre analyse aux aspects qui ont le plus rapport à l'estimation pour petits domaines.

Soit $\hat{\underset{\sim}{\alpha}}_{t-1}$ le meilleur prédicteur linéaire sans biais (MPLS) de $\underset{\sim}{\alpha}_{t-1}$, fondé sur l'ensemble des observations enregistrées jusqu'à la période $(t-1)$ inclusivement. Comme $\hat{\underset{\sim}{\alpha}}_{t-1}$ est MPLS de $\underset{\sim}{\alpha}_{t-1}$, $\hat{\underset{\sim}{\alpha}}_{t|t-1} = T\hat{\underset{\sim}{\alpha}}_{t-1}$ est le MPLS de $\underset{\sim}{\alpha}_t$ au temps $(t-1)$. De plus, si $P_{t-1} = E(\hat{\underset{\sim}{\alpha}}_{t-1} - \underset{\sim}{\alpha}_{t-1})(\hat{\underset{\sim}{\alpha}}_{t-1} - \underset{\sim}{\alpha}_{t-1})'$ est la matrice de variances-covariances des erreurs de prédiction au temps $(t-1)$, $P_{t|t-1} = TP_{t-1}T' + G\Lambda G'$ est la matrice de variances-covariances des erreurs de prédiction $(\hat{\underset{\sim}{\alpha}}_{t|t-1} - \underset{\sim}{\alpha}_t)$ (découle directement de (2.8)).

Lorsqu'un nouveau vecteur d'observations $[\underset{\sim}{Y}_t, Z_t]$ est connu, le prédicteur de $\underset{\sim}{\alpha}_t$ et la matrice de variances-covariances P_{t-1} sont mis à jour suivant les formules

$$\begin{aligned} \hat{\underset{\sim}{\alpha}}_t &= \hat{\underset{\sim}{\alpha}}_{t|t-1} + P_{t|t-1}Z'_tF_t^{-1}(\underset{\sim}{Y}_t - \hat{\underset{\sim}{Y}}_{t|t-1}) \\ P_t &= (I - P_{t|t-1}Z'_tF_t^{-1}Z_t)P_{t|t-1} \end{aligned} \tag{3.1}$$

où $\hat{Y}_{t|t-1} = Z_t \hat{\alpha}_{t|t-1}$ est le MPLS de Y_t au temps $(t-1)$ de sorte que $e_t = (Y_t - \hat{Y}_{t|t-1})$ est le vecteur des résidus une étape à l'avance ayant comme matrice de variances-covariances $F_t = (Z_t P_{t|t-1} Z_t' + \Sigma_t)$.

Les nouvelles données enregistrées au temps t peuvent servir aussi à la mise à jour (ou au lissage) d'anciens estimateurs des vecteurs d'états et, par voie de conséquence, à la mise à jour d'anciens estimateurs des moyennes de petites régions. En désignant par t^* le dernier mois où ont été enregistrées des observations, on exécute le lissage au moyen des équations

$$\begin{aligned} \hat{\alpha}_{t|t^*} &= \hat{\alpha}_t + P_t T' P_{t+1|t}^{-1} (\hat{\alpha}_{t+1|t^*} - T\hat{\alpha}_t) \\ P_{t|t^*} &= P_t + P_t T' P_{t+1|t}^{-1} (P_{t+1|t^*} - P_{t+1|t}) P_{t+1|t}^{-1} T P_t; \; t = 2, 3, \ldots, t^* \end{aligned} \tag{3.2}$$

où $P_{t|t^*}$ est la matrice de variances-covariances des erreurs de prédiction $(\hat{\alpha}_{t|t^*} - \alpha_t)$. Notons que $\hat{\alpha}_{t^*|t^*} = \hat{\alpha}_{t^*}$ et $P_{t^*|t^*} = P_{t^*}$, ceci définissant les valeurs initiales pour les équations de lissage.

Les estimateurs lissés de α_t permettent d'obtenir facilement des estimateurs de moyennes de petites régions ou de moyennes globales à l'aide de la relation $\hat{\Theta}_{tk} = \bar{X}_{tk} \hat{\beta}_{tk} = \bar{Z}_{tk}' \hat{\alpha}_{tk} = \bar{Z}_{tk}' A_{tk} \hat{\alpha}_t$ où $\bar{Z}_{tk}' = (1, 0, \bar{X}_{tk1}, 0, \ldots, \bar{X}_{tkp}, 0)$ et A_{tk} est la matrice indicatrice appropriée. Par conséquent, si $\Theta_t^w = \sum_{k=1}^K w_k \Theta_{tk}$, alors $\hat{\Theta}_t^w = \sum_{k=1}^K w_k \bar{Z}_{tk}' A_{tk} \hat{\alpha}_t = a_{tw}' \hat{\alpha}_t$, (par exemple). Pour des matrices de variances-covariances Σ_t et Λ données, l'EQM des erreurs d'estimation est calculée au moyen des équations

$$E(\hat{\Theta}_{tk} - \Theta_{tk})^2 = \bar{Z}_{tk}' A_{tk} P_t A_{tk}' \bar{Z}_{tk} \quad \text{et} \quad E(\hat{\Theta}_{tk}^w - \Theta_{tk}^w) = a_{tw}' P_t a_{tw}. \tag{3.3}$$

Il convient de souligner que l'EQM définie ci-dessus a rapport à la distribution conjointe des observations $\{Y_{tk}\}$ et des vecteurs de coefficients $\{\beta_{tk}\}$ de sorte qu'elle constitue une moyenne pour toutes les valeurs possibles de la moyenne de petite région.

3.2 Estimation des matrices de variances-covariances et initialisation du filtre

Pour appliquer le filtre de Kalman, il est nécessaire d'estimer les éléments inconnus des matrices Σ_t et Λ et d'initialiser le filtre, c'est-à-dire d'estimer le vecteur α_o et la matrice de variances-covariances correspondante (P_o) des erreurs d'estimation. Dans cette section, nous décrivons des méthodes d'estimation simples qui peuvent servir à ces calculs.

Dans l'hypothèse que les résidus ϵ_t et η_t des équations (2.7) et (2.8) sont distribués suivant une loi normale, nous pouvons exprimer la fonction de vraisemblance logarithmique des vecteurs $Y_{m+1}, \ldots, Y_{t^*}$, étant donné les m premiers vecteurs $Y_1, \ldots, Y_m$, de la façon suivante:

$$L(\lambda) = \text{constante} - \frac{1}{2} \sum_{t=m+1}^{t^*} (\log | F_t | + e_t' F_t^{-1} e_t) \tag{3.4}$$

où λ renferme les variances et les covariances inconnues du modèle, exprimées sous forme vectorielle. La grandeur scalaire m représente le nombre de périodes nécessaires pour construire des valeurs initiales pour le filtre de Kalman. (En ce qui concerne le modèle de trajet aléatoire exposé dans la section 2.2, $m = 1$, à la condition que l'on dispose de suffisamment de données pour chaque région de manière à pouvoir déterminer les estimateurs par les m.c.o. (moindres carrés ordinaires) des vecteurs de coefficients). L'expression (3.4) découle de la décomposition de l'erreur de prédiction; voir Schweppe (1965) et Harvey (1981) pour plus de détails. Pour des matrices Σ_t et Λ données, on détermine les résidus e_t et les matrices de variances-covariances F_t à l'aide des équations du filtre de Kalman (3.1).

Il faut initialiser le filtre de Kalman pour calculer la fonction de vraisemblance; la meilleure façon d'exécuter cette opération est de recourir à la méthode proposée par Harvey et Phillips (1979). Selon cette méthode, les éléments non stationnaires du vecteur d'états sont initialisés avec une variance de l'erreur très élevée (ce qui équivaut à poser par hypothèse une distribution a priori non informative), de telle sorte que l'on puisse considérer comme nulles les estimations d'états correspondantes. (Pour ce qui a trait au modèle de trajet aléatoire, le fait d'initialiser avec une distribution a priori non informative permet d'établir les estimateurs par les m.c.o. au bout d'une période; voir Meinhold et Singpurwalla 1983, pour une formulation bayesienne du filtre de Kalman.) Quant aux éléments stationnaires du vecteur d'états, ils sont initialisés à l'aide des moyennes et des variances inconditionnelles correspondantes, qui peuvent compter parmi les paramètres inconnus qui définissent les arguments de la fonction de vraisemblance.

On peut maximiser la fonction de vraisemblance (3.4) par la méthode de la fonction de caractérisation avec pas variable. Plus particulièrement, désignons par $\underset{\sim}{\lambda}_{(o)}$ les valeurs estimées initiales des éléments inconnus dans $\underset{\sim}{\lambda}$. La méthode de la fonction de caractérisation consiste alors à résoudre itérativement le système d'équations

$$\underset{\sim}{\lambda}_{(i)} = \underset{\sim}{\lambda}_{(i-1)} + r_i\{I[\underset{\sim}{\lambda}_{(i-1)}]\}^{-1} g[\underset{\sim}{\lambda}_{(i-1)}] \tag{3.5}$$

où $\underset{\sim}{\lambda}_{(i-1)}$ est l'estimateur de $\underset{\sim}{\lambda}$ obtenu à la $(i-1)$ième itération, $I[\underset{\sim}{\lambda}_{(i-1)}]$ est la matrice d'information calculée en fonction de $\underset{\sim}{\lambda}_{i-1}$ et $g[\underset{\sim}{\lambda}_{(i-1)}]$ est le gradient de la fonction de vraisemblance logarithmique calculé à $\underset{\sim}{\lambda}_{i-1}$. Le coefficient r_i désigne un pas variable qui fait que la condition $L[\underset{\sim}{\lambda}_{(i)}] \geq L[\underset{\sim}{\lambda}_{(i-1)}]$ est respectée à chaque itération. La valeur de ce coefficient peut être déterminée au moyen d'une recherche par quadrillage dans la région $[0,1]$. Les formules pour l'élément k du vecteur gradient et l'élément $k\ell$ de la matrice d'information sont définies dans Watson et Engle (1983).

Une fois que les variances et les covariances du modèle ont été estimées, on peut substituer ces estimations aux paramètres réels des équations du filtre de Kalman (équ. 3.1 et 3.2) de manière à obtenir les estimateurs des coefficients de régression et des matrices de variances-covariances et, de là, les estimateurs pour petites régions et les variances correspondantes (voir équation 3.3). Soulignons toutefois que les matrices de variances-covariances estimées ne tiennent pas compte de la variation qui découle de la nécessité d'estimer les éléments inconnus compris dans $\underset{\sim}{\lambda}$. Pour tenir compte de cette variation additionnelle dans l'élaboration de modèles d'espace d'états, Ansley et Kohn (1986) proposent l'utilisation de facteurs de correction d'ordre $1/t^*$ en ayant recours à des formules d'approximation de Taylor du premier ordre. Pour sa part, Hamilton (1986) propose une méthode du type Monte Carlo qui consiste à échantillonner des unités provenant d'une distribution normale multidimensionnelle dont la moyenne est définie par l'estimateur du maximum de vraisemblance du vecteur $\underset{\sim}{\lambda}$ et la matrice de variances-covariances, par l'inverse de la matrice d'information, et à estimer les vecteurs d'états pour chaque réalisation aléatoire des valeurs de paramètres. Cette méthode offre plus de souplesse au point de vue des hypothèses et nous renseigne davantage sur la sensibilité des estimateurs du filtre de Kalman par rapport à l'erreur dans les estimateurs de la variance et de la covariance. En revanche, elle est plus exigeante sur le plan du calcul.

4. MODIFICATIONS VISANT À PRÉVENIR LES DÉFAILLANCES DU MODÈLE

4.1 Description du problème et modifications proposées

L'utilisation d'un modèle paraît inévitable dans l'estimation pour petites régions compte tenu de la faible taille des échantillons. Ceci étant dit, il faut se demander comment prévenir les défaillances du modèle. Tester le modèle à chaque fois que de nouvelles données sont connues est peu pratique; il faut plutôt envisager un "dispositif intégré" qui conserve la robustesse des estimateurs lorsque le modèle n'est pas valide.

On peut, par exemple, modifier les estimateurs par régression calculés aux diverses périodes de manière qu'ils satisfassent certaines contraintes linéaires que l'on aura établies en posant la moyenne globale des données brutes égale à sa valeur ajustée probable selon le modèle. Plus précisément, nous proposons d'ajouter à l'équation de modèle (2.1) des contraintes linéaires du type

$$\sum_k W_{tk}^{(\ell)} \sum_i Y_{tki} = \sum_k W_{tk}^{(i)} \sum_i \underset{\sim}{x}'_{tki} \underset{\sim}{\beta}_{tk} \quad \ell = 1, 2, \ldots, L(t), \; t = 1, \ldots, t^* \qquad (4.1)$$

où les coefficients $W_{tk}^{(\ell)}$ sont des poids normalisés fixes tels que $\sum_k n_{tk} W_{tk}^{(\ell)} = 1$. Un exemple de contrainte linéaire est l'équation

$$\sum_{k=1}^{K} N_{tk} \hat{M}_{tk} \Big/ \sum_{k=1}^{K} N_{tk} = \sum_{k=1}^{K} N_{tk} (\underset{\sim}{\bar{x}}'_{tk} \underset{\sim}{\beta}_{tk}) \Big/ \sum_{k=1}^{K} N_{tk} \qquad (4.2)$$

où $\hat{M}_{tk}$ est l'estimateur d'enquête (direct) pour la région k. Lorsque $\underset{\sim}{\bar{x}}_{tk} \simeq \underset{\sim}{\bar{X}}_{tk}$, l'équation (4.2) fait que le prédicteur de la moyenne globale de population (fondé sur un modèle) concorde avec l'estimateur d'enquête correspondant. On peut justifier cette contrainte en faisant valoir que, même s'ils ne sont pas assez fiables pour l'estimation des moyennes de petites régions, vu la faible taille des échantillons, les estimateurs d'enquête peuvent être jugés très satisfaisants lorsqu'ils servent collectivement à estimer la moyenne globale. Notons, de plus, que les organismes statistiques doivent de toute façon définir des contraintes pour les besoins de leurs analyses. Battese, Harter et Fuller (1988) et Pfeffermann et Barnard (1991) utilisent une contrainte semblable pour analyser des enquêtes transversales. Il arrive souvent que l'on puisse former des groupes avec les petites régions, à la condition qu'il y ait suffisamment de données dans chacun des groupes pour pouvoir estimer les moyennes de groupes à l'aide des estimateurs d'enquête. On peut alors introduire plusieurs contraintes comme celle définie en (4.2), la sommation s'étendant cette fois aux régions du même groupe. À cet égard, il convient de souligner que, compte tenu de la corrélation entre les coefficients de régression dans les diverses régions, l'application d'une contrainte à un sous-ensemble des régions influera sur les estimations de toutes les régions. Un exemple illustrera cette propriété.

Il importe de souligner que la série de contraintes définie en (4.1) ne représente pas de l'information externe sur les valeurs probables des coefficients de régression. Elle constitue plutôt un "système de contrôle" qui vise à ce que les estimateurs du modèle réagissent plus rapidement à une variation éventuelle des coefficients de régression. C'est pourquoi les variances des estimateurs par régression modifiés sont légèrement plus élevées que celles des estimateurs optimaux. Évidemment, lorsqu'il ne se produit aucun changement et que les variances des moyennes globales sont suffisamment faibles, on s'attend que les contraintes soient à peu près satisfaites sans même qu'elles soient introduites explicitement. Comme nous l'avons mentionné plus haut, il est possible d'introduire plusieurs contraintes différentes à chaque période mais il faut absolument que les variances des moyennes globales correspondantes soient suffisamment faibles pour que les modifications soient vraiment nécessaires et qu'elles ne se confondent pas avec les fluctuations aléatoires des données brutes.

4.2 Inférence avec contraintes linéaires

Dans la section 4.1, nous avons proposé de modifier l'équation de modèle (2.1) en y ajoutant la série de contraintes (4.1) de manière que les estimateurs par régression résistent bien à une variation subite des valeurs des coefficients.

La meilleure façon de procéder sur le plan du calcul est d'ajouter au vecteur $\underset{\sim}{Y}_t$ de l'équation (2.7) la grandeur scalaire $\sum_k W_{tk}^{(\ell)} \sum_i Y_{tki}$, d'ajouter aux matrices Z_t les vecteurs lignes

correspondants ($W_{t1}^{(t)} 1'_{nt1} Z_{t1}$, ..., $W_{tK}^{(t)} 1'_{ntK} Z_{tK}$) et de poser les variances des résidus égales à zéro. Le nouveau système d'équations, y compris (2.8), constitue un pseudo-modèle d'espace d'états qui peut être estimé à l'aide des équations du filtre de Kalman (3.1). Notons que la pseudo-matrice de variances-covariances $\Sigma_t^{(P)}$ du vecteur élargi des résidus n'est plus définie positive (les $L(t)$ dernières lignes et $L(t)$ dernières colonnes de $\Sigma_t^{(P)}$ sont formées de zéros) mais cela ne complique pas le calcul.

L'application du filtre de Kalman au pseudo-modèle présente un inconvénient; en effet, les matrices de variances-covariances des estimateurs par régression ne tiennent pas compte de la variabilité réelle des moyennes globales dans le membre de gauche de l'équation (4.1). Pour remédier à cela, nous proposons de modifier la formule de mise à jour de la matrice de variances-covariances P_t (équation 3.1) de manière que l'on tienne compte des variances et des covariances des moyennes globales.

Désignons par $\underset{\sim}{Y}_t^{(A)}$ et $Z_t^{(A)}$ le vecteur Y et la matrice Z élargis au temps t et par $\Sigma_t^{(A)}$, la vraie matrice de variances-covariances des résidus $[\underset{\sim}{Y}_t^{(A)} - Z_t^{(A)} \underset{\sim}{\alpha}_t]$. La matrice $\Sigma_t^{(A)}$ est d'ordre $[n_t + L(t)]$, les éléments de Σ_t occupant les n_t premières lignes et n_t premières colonnes, avec les variances et les covariances des moyennes $\Sigma_k W_{tk}^{(t)} \Sigma_i Y_{tki}$, et les éléments du vecteur $\underset{\sim}{Y}_t$ occupant le reste des lignes et des colonnes. En désignant par $\hat{\underset{\sim}{\alpha}}_{t-1}^{(A)}$ le prédicteur robuste de $\underset{\sim}{\alpha}_{t-1}$ déterminé à $(t-1)$ à l'aide du pseudo-modèle et par $P_{t-1}^{(A)}$, la vraie matrice de variances-covariances des erreurs $(\hat{\underset{\sim}{\alpha}}_{t-1}^{(A)} - \underset{\sim}{\alpha}_{t-1})$, nous pouvons déterminer l'estimateur d'état modifié au temps t par la formule

$$\hat{\underset{\sim}{\alpha}}_t^{(A)} = T\hat{\underset{\sim}{\alpha}}_{t-1}^{(A)} + P_{t|t-1}^{(A)} Z_t^{(A)'} \left(F_t^{(P)}\right)^{-1} \left[\underset{\sim}{Y}_t^{(A)} - Z_t^{(A)} T\hat{\underset{\sim}{\alpha}}_{t-1}^{(A)}\right] \tag{4.3}$$

où $P_{t|t-1}^{(A)} = (TP_{t-1}^{(A)} T' + G\Lambda G')$ et $F_t^{(P)} = Z_t^{(A)} P_{t|t-1}^{(A)} Z_t^{(A)'} + \Sigma_t^{(P)}$ (comparer avec les équations (3.1)). Nous montrons en annexe que la matrice réelle des variances-covariances $(P_t^{(A)})$ des erreurs $(\hat{\underset{\sim}{\alpha}}_t^{(A)} - \underset{\sim}{\alpha}_t)$ satisfait l'équation récursive

$$P_t^{(A)} = \left[I - K_t^{(P)} Z_t^{(A)}\right] P_{t|t-1}^{(A)} + K_t^{(P)} \left[\Sigma_t^{(A)} - \Sigma_t^{(P)}\right] K_t^{(P)}, \tag{4.4}$$

où $K_t^{(P)} = P_{t|t-1}^{(A)} Z_t^{(A)'} (F_t^{(P)})^{-1}$ est le pseudo-gain de Kalman. Le premier terme du membre de droite de l'équation (4.4) représente la formule de mise à jour usuelle du filtre de Kalman (comparer avec (3.1)). Le second terme est un facteur de correction qui, tient compte des variances et des covariances réelles des moyennes $\Sigma_k W_{tk}^{(t)} \Sigma_i Y_{tki}$, qu'ignore le premier terme.

Le filtre de Kalman modifié défini en (4.3) et (4.4) produit le prédicteur robuste $\hat{\underset{\sim}{\alpha}}_t^{(A)}$ au lieu du prédicteur optimal (fondé sur un modèle) $\hat{\underset{\sim}{\alpha}}_t$ mais utilise la bonne matrice de variances-covariances suivant le modèle. Il peut donc servir régulièrement à estimer les vecteurs de coefficients et, par voie de conséquence, les moyennes de petites régions; de plus, lorsque le modèle est valide, il donnera des résultats comparables à ceux obtenus avec le filtre optimal. Dans les cas où le modèle ne serait pas valide, l'équation (4.4) pourrait être incorrecte (suivant le genre de défaillance qui affecte le modèle) mais les prédicteurs $\hat{\underset{\sim}{\alpha}}_t^{(A)}$ satisferont quand même les contraintes linéaires (4.1). De même, il est possible de modifier les équations de lissage (3.2) pour qu'elles satisfassent les contraintes linéaires.

5. RÉSULTATS EMPIRIQUES

5.1 Description des données et du modèle ajusté

Afin d'illustrer les grandes caractéristiques de la classe de modèles définis dans la section 2, nous avons ajusté ces modèles aux prix de vente des maisons à Jérusalem. Le Bureau central de la statistique d'Israël fait un relevé de ces prix à chaque mois et utilise régulièrement ces

données pour le calcul des indices mensuels des prix du logement (IPL) corrigés en fonction des changements qualitatifs. On calcule un IPL pour chaque ville ou chaque ensemble de villes et pour chaque catégorie de logement, définie par le nombre de pièces (de 1 à 5). Le nombre mensuel de transactions est très peu élevé pour de nombreuses combinaisons "ville/catégorie de logement" et il arrive qu'il n'y ait aucune transaction dans le cas des logements d'une pièce. Le tableau ci-dessous donne le nombre mensuel moyen de transactions enregistrées entre juillet 1987 et novembre 1989, ainsi que l'écart-type (É.T.) correspondant.

Catégorie	1	2	3	4	5
Moyenne	2.7	29.0	101.9	39.7	5.6
É.T.	2.6	12.9	50.4	18.8	3.5

La nécessité de tenir compte des changements qualitatifs tient à ce que les transactions ne sont soumises à aucun contrôle, ce qui engendre de fortes variations qualitatives d'un mois à l'autre, particulièrement dans les cases à faible fréquence. Pour chaque transaction, on enregistre les variables de mesure de la qualité (VMQ) suivantes: $\tilde{X}^{(1)}$ - superficie du logement; $\tilde{X}^{(2)}$ - âge du logement; $X^{(3)}$, $X^{(4)}$ - variables auxiliaires servant à désigner les districts de la ville.

Dans un article récent, Pfeffermann, Burck et Ben-Tuvia (1989) analysent en détail les problèmes liés au calcul de l'IPL et la méthode de calcul utilisée en Israël. Ils proposent le modèle ci-dessous comme solution de rechange pour le modèle en usage actuellement. L'indice triple "*tki*" sert à désigner la transaction *i* dans la catégorie *k* au mois *t*, Y_{tki} étant le logarithme du prix de vente correspondant et $X_{tki}^{(j)} = \log(\tilde{X}_{tki}^{(j)})$, $j = 1, 2$.

$$Y_{tki} = \beta_{tk0} + \beta_{tk1}X_{tki}^{(1)} + \beta_{tk2}X_{tki}^{(2)} + \beta_{tk3}X_{tki}^{(3)} + \beta_{tk4}X_{tki}^{(4)} + \epsilon_{tki} \tag{5.1}$$

$$\begin{aligned} \beta_{tk0} &= \beta_{t-1,k0} + \beta_{k0} + \eta_{tk0} \\ \beta_{tkj} &= \beta_{t-1,kj} + \eta_{tkj},\ j = 1, \ldots, 4, \end{aligned} \tag{5.2}$$

où les termes d'erreur ϵ_{tki} et η_{tkj} satisfont aux hypothèses (2.1), (2.4) et (2.5). Notons que le modèle posé pour l'ordonnée à l'origine correspond à la formule d'approximation locale d'une tendance linéaire, définie dans l'exemple (d) de la section (2.1). Le modèle posé pour les autres coefficients correspond au modèle de trajet aléatoire défini dans l'exemple (b).

À partir du modèle de régression défini en (5.1), on peut construire un IPL corrigé en fonction des changements qualitatifs. En prenant pour valeur des VMQ la moyenne de population correspondante, qui est constante en pratique (la valeur de ces variables étant révisée à tous les cinq ans environ), on peut calculer des prix de vente moyens à l'aide de (5.1) et ces moyennes sont comparables d'un mois à l'autre puisqu'elles se rapportent à des logements de même nature.

Pfeffermann, Burck et Ben-Tuvia examinent les raisons qui motivent le choix du modèle (5.2) pour les coefficients de régression. Ils présentent aussi des résultats empiriques qui confirment la validité du modèle. Seulement, ces résultats ont été obtenus en ajustant le modèle aux données de chaque case prise individuellement, c'est-à-dire sans tenir compte de la corrélation transversale entre les coefficients de régression. Dans la présente étude, nous approfondissons cet aspect du modèle et, chose non moins importante, nous donnons un aperçu de l'efficacité des modifications proposées dans la section 4 pour prévenir les défaillances du modèle.

5.2 Estimation du modèle

Le modèle défini en (5.1) et (5.2) peut être exprimé sous forme de modèle d'espace d'états comme dans les équations (2.7) et (2.8). En fait, les vecteurs α_t et les matrices Z_t, T et G impliquent en l'occurrence des structures simples puisque pour $j = 1, \ldots, 4$, $\beta_{kj} \equiv 0$

(voir exemple (b) dans la section 2.1). Ainsi, $\underset{\sim}{\alpha}'_{tk} = (\beta_{tk0}, \beta_{k0}, \beta_{tk1}, \ldots, \beta_{tk4})$, $Z_{tk} = [1_{ntk}, 0_{ntk}, X_{tk}^{(1)}, \ldots, X_{tk}^{(4)}]$, $\tilde{T} = [\underset{\sim}{e}_1, \underset{\sim}{e}_1 + \underset{\sim}{e}_2, \underset{\sim}{e}_3, \ldots, \underset{\sim}{e}_6]$, une matrice 6×6 où $\underset{\sim}{e}_j$ renferme le nombre un à la position j et des zéros aux autres positions, et $\tilde{G} = [\underset{\sim}{e}_1, \underset{\sim}{e}_3, \ldots, \underset{\sim}{e}_6]$, une matrice 6×5. La matrice Δ est définie comme en (2.5). Le vecteur $\underset{\sim}{\alpha}_t$ et les matrices Z_t, T, G et Λ sont déterminés à l'aide des vecteurs $\{\underset{\sim}{\alpha}_{tk}\}$ et des matrices $\{Z_{tk}\}$, $\tilde{T}$, $\tilde{G}$ et Δ de la même manière qu'en (2.7) et (2.8).

Après avoir exprimé le modèle sous forme de modèle d'espace d'états, nous avons cherché à estimer les variances et les covariances inconnues à l'aide de la méthode de la fonction de caractérisation décrite dans la section 3.2. Or, nous nous sommes rendu compte que la puissance de l'ordinateur qui a servi à cette étude, un IBM 1481, était de beaucoup inférieure à ce qu'il aurait fallu avoir pour obtenir la convergence dans des conditions raisonnables. Il convient de souligner que le modèle d'espace d'états combiné renferme vingt-cinq paramètres inconnus ($\dim(\underset{\sim}{\lambda}) = 25$) tandis que les vecteurs d'états et les matrices de variances-covariances de ces vecteurs ont une dimension de 30 ($\dim(\underset{\sim}{\alpha}_t) = 30$). Le nombre mensuel d'observations varie de 55 à 353. Le programme qui a été élaboré pour cette étude utilise des différences finies, de sorte qu'à chaque itération de la méthode de la fonction de caractérisation, il doit se faire une exploration intégrale des données, chaque exploration comportant $[\dim(\underset{\sim}{\lambda}) + 1]$ opérations de calcul pour le vecteur d'états $\hat{\underset{\sim}{\alpha}}_t$ et la matrice de variances-covariances P_t (équation 3.1) pour chaque période. Ces calculs sont nécessaires pour évaluer les fonctions de vraisemblance logarithmiques et les différences finies correspondantes. C'est donc dire que les frais de calcul augmentent avec le nombre d'observations, la dimension du vecteur d'états et le nombre de paramètres inconnus.

Pour contourner ce problème, nous avons estimé séparément la variance σ_k^2 (équation 2.1) et la matrice Δ (équation 2.5) pour chacune des catégories de logement au moyen des séries d'observations respectives, puis avons estimé les coefficients de corrélation ρ_j (équation 2.6) par une simple recherche par quadrillage. Nous avons constaté que le fait de poser ρ_j égal à ½, pour chaque j, donnait des résultats satisfaisants tant au point de vue de la tendance des résidus une étape à l'avance (erreurs de prédiction une étape à l'avance) qu'au point de vue du lissage des coefficients de régression relatifs aux logements d'une pièce et de cinq pièces, pour lesquels il y a très peu d'observations à chaque mois. Précisons qu'en estimant séparément les variances et les covariances qui définissent la corrélation longitudinale des coefficients de régression propres à chaque catégorie, on déroge quelque peu aux hypothèses du modèle, mais il faut dire qu'il y a de toute évidence une perte d'efficacité si les variances et les covariances sont identiques d'une catégorie à l'autre.

5.3 Résultats

Pfeffermann, Burck et Ben-Tuvia (1989) montrent jusqu'à quel point les modèles chronologiques peuvent être ajustés convenablement aux données relatives aux diverses catégories de logement. Dans la présente étude, nous cherchons plutôt à comparer les résultats obtenus lorsqu'on tient compte de la corrélation transversale avec ceux obtenus lorsqu'on n'en tient pas compte, et à montrer dans quelle mesure les modifications proposées en (4.1) permettent de prévenir les défaillances du modèle.

Afin de rendre la comparaison la plus claire possible, nous avons augmenté délibérément les valeurs de Y de 5% à chacune des ces périodes: octobre 1987, novembre 1988, janvier 1989 et mai 1989. Ainsi, toutes les valeurs de Y enregistrées de octobre 1987 à octobre 1988 pour les cinq catégories de logement ont été "gonflées" de 5%, les valeurs de Y enregistrées en novembre et en décembre 1988 ont été haussées de 10.25% (c'est-à-dire 5% ajouté au 5% précédent) et ainsi de suite. Ce genre de défaillances peut survenir (quoique dans des conditions évidemment beaucoup moins sévères) par suite d'une dévaluation de la monnaie par exemple, et il importe d'en tenir compte dans la modélisation des prix de vente. Voir Pfeffermann, Burck et Ben-Tuvia (1989) pour une analyse approfondie. Des défaillances semblables peuvent survenir par exemple pour les séries de taux de chômage dans les périodes de récession profonde.

Tableau 1

Fluctuation moyenne des résidus et des résidus une étape à l'avance, compte tenu ou non de la corrélation transversale et du caractère de robustesse; selon la catégorie de logement

Cat. de logement (n° de pièces)	Fluctuation des résidus une étape à l'avance				Fluctuation des résidus			
	$\rho \equiv ½$		$\rho \equiv 0$		$\rho \equiv ½$		$\rho \equiv 0$	
	Rob.	Non rob.	Rob.	Non rob.	Rob.	Non rob.	Rob.	Non rob.
1	.141	.134	.176	.218	.021	.027	.056	.092
2	.070	.090	.084	.123	.021	.039	.023	.070
3	.065	.090	.070	.197	.017	.042	.019	.143
4	.067	.123	.072	.198	.019	.066	.021	.141
5	.067	.114	.077	.193	.023	.033	.065	.106

Le tableau 1 donne la fluctuation moyenne (FM) des résidus $\hat{\epsilon}_{tki} = (Y_{tki} - \hat{\beta}_{tk0} - \sum_{j=1}^{4} X_{tki}^{(j)} \hat{\beta}_{tkj})$ et des résidus une étape à l'avance $e_{tki} = [Y_{tki} - (\hat{\beta}_{t-1,k0} + \beta_{k0}) - \sum_{j=1}^{4} X_{tki}^{(j)} \hat{\beta}_{t-1,kj}]$ du modèle (voir équations 5.1 et 5.2) pour chacune des catégories de logement. Les formules utilisées pour calculer cette fluctuation sont $\mathrm{FM}_k(\epsilon) = 1/N \sum_{t=1}^{N} (1/n_t \sum_{i=1}^{n_t} \hat{\epsilon}_{tki}^2)$; $\mathrm{FM}_k(e) = 1/N \sum_{t=1}^{N} (1/n_t \sum_{i=1}^{n_t} e_{tki}^2)$, où $t = 1, \ldots, N$ désigne les mois compris entre juillet 1987 et novembre 1989. Nous définissons quatre estimateurs des coefficients de régression, selon que le modèle tient compte ($\rho_j \equiv ½$) ou non ($\rho_j \equiv 0$) de la corrélation transversale et selon qu'on a modifié ou non les estimateurs de manière à prévenir les défaillances du modèle (dans le tableau, ce dernier critère est représenté par les rubriques "robuste" (rob.) et "non robuste" (non rob.)). Pour modifier les estimateurs, nous avons ajouté à l'équation d'observation de chaque mois trois contraintes linéaires comme celle définie en (4.2). Ces contraintes ont pour effet de faire concorder la moyenne globale des valeurs ajustées, dans chacun des trois districts, avec la moyenne correspondante des valeurs observées. Lorsque nous avons introduit les contraintes, nous avons ajusté le modèle à l'aide du filtre de Kalman modifié défini par les équations (4.3) et (4.4).

Afin d'illustrer l'efficacité des quatre séries d'estimateurs par régression aux divers mois et plus particulièrement autour des mois où nous avons "gonflé" les observations, nous avons représenté graphiquement la fluctuation mensuelle des résidus une étape à l'avance et des résidus ordinaires pour les logements de 3 et de 5 pièces. Les graphiques pertinents sont reproduits dans les figures 1 à 4. Notons que les valeurs indiquées dans le tableau 1 pour les logements de 3 et de 5 pièces sont les moyennes respectives des valeurs reproduites dans les quatre graphiques.

Le tableau et les graphiques nous permettent de tirer les conclusions suivantes:

Lorsqu'on tient compte de la corrélation transversale et qu'on introduit les contraintes linéaires visant à rendre robuste l'estimateur, on obtient de meilleurs résultats que dans tous les autres cas. Cela est particulièrement notable dans le cas des logements d'une pièce et de cinq pièces, pour lesquels il y a très peu d'observations à chaque mois. En ce qui a trait aux trois autres catégories de logement, on n'observe que de faibles différences entre le scénario ($\rho \equiv ½$, est. robuste) et le scénario ($\rho \equiv 0$, est. robuste), ce qui était prévisible étant donné que la quantité d'information empruntée aux cases avoisinantes (dans un contexte plus général: petites régions) va en diminuant lorsque le nombre d'observations mensuelles augmente. Toutefois, la situation est tout autre lorsqu'on fait abstraction des contraintes linéaires. Dans un tel cas, on obtient de bien meilleurs résultats lorsqu'on tient compte de la corrélation transversale que lorsqu'on n'en tient pas compte et ce, pour toutes les catégories de logement. Ainsi, grâce à l'emprunt d'information, les estimateurs des coefficients de régression s'adaptent beaucoup plus rapidement à une variation subite des données, comme le montrent clairement les figures 1 à 4. Les quatre sommets de chaque graphique correspondent aux mois où il y a eu un "gonflement" des données et comme nous pouvons le voir, les courbes se rapportant au scénario

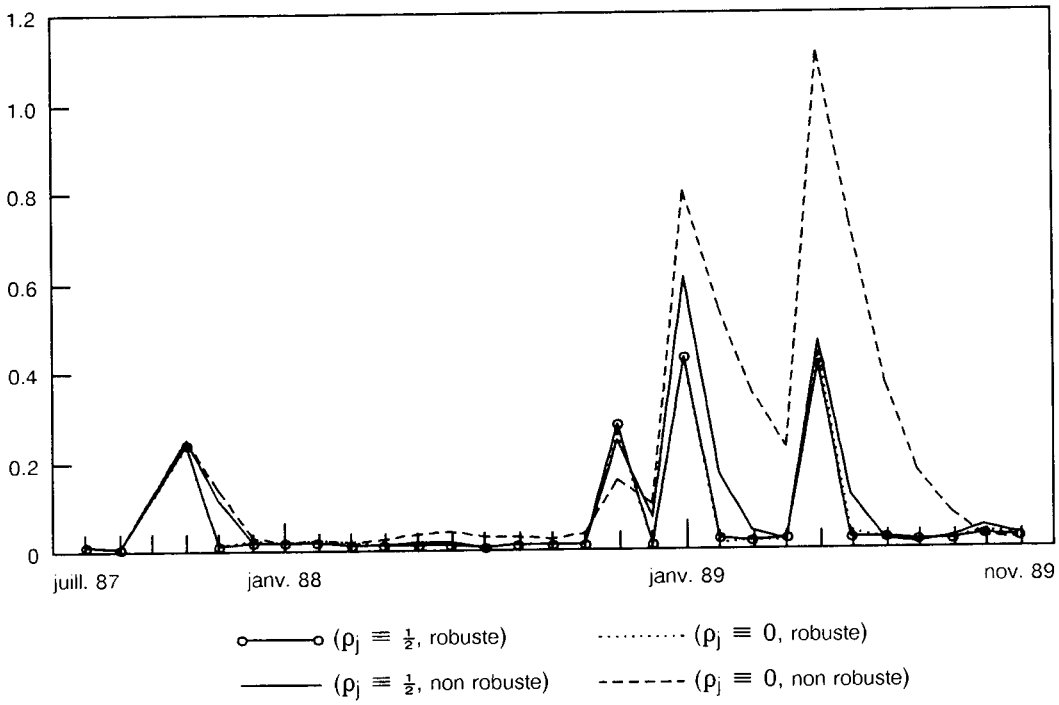

Figure 1 Fluctuation mensuelle des résidus une étape à l'avance; logements de 3 pièces

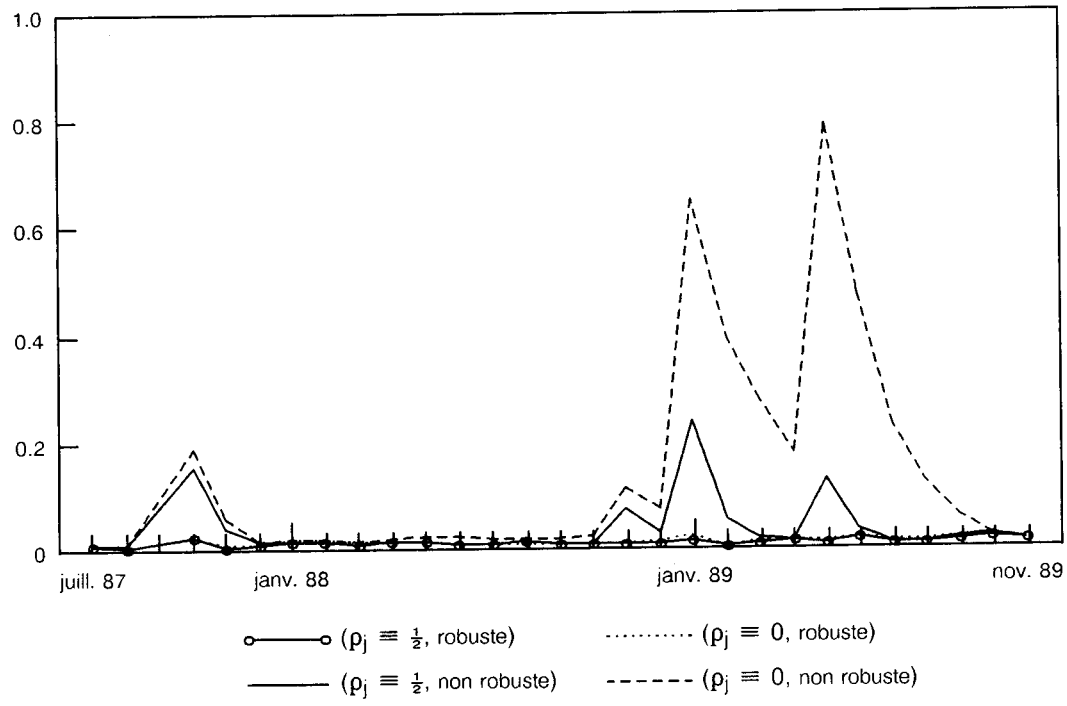

Figure 2 Fluctuation mensuelle des résidus; logements de 3 pièces

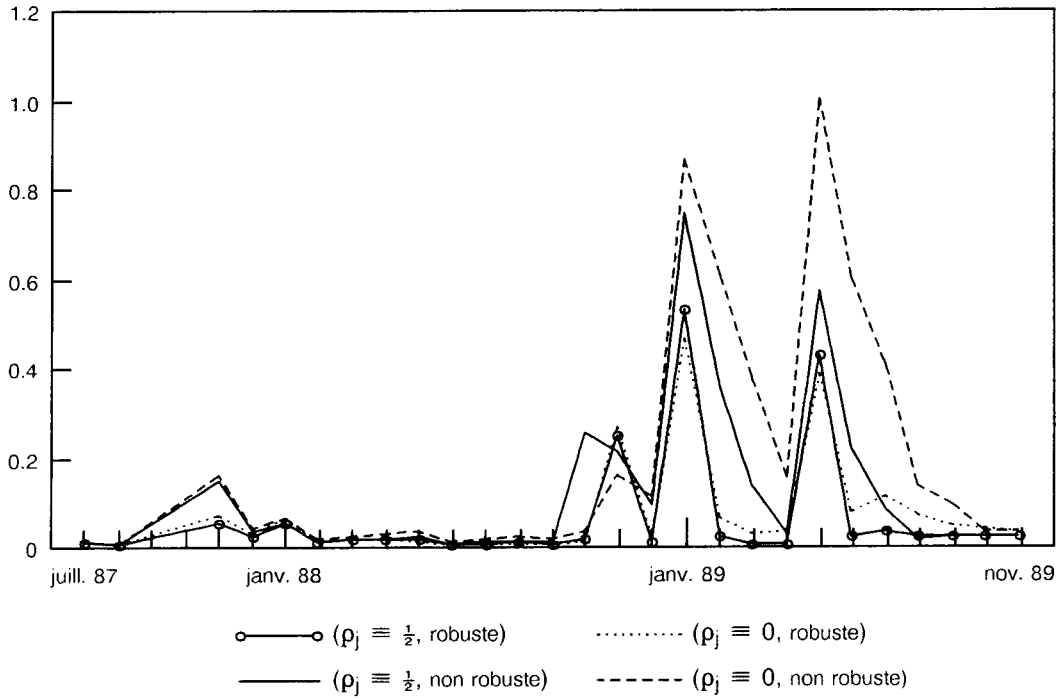

Figure 3 Fluctuation mensuelle des résidus une étape à l'avance; logements de 5 pièces

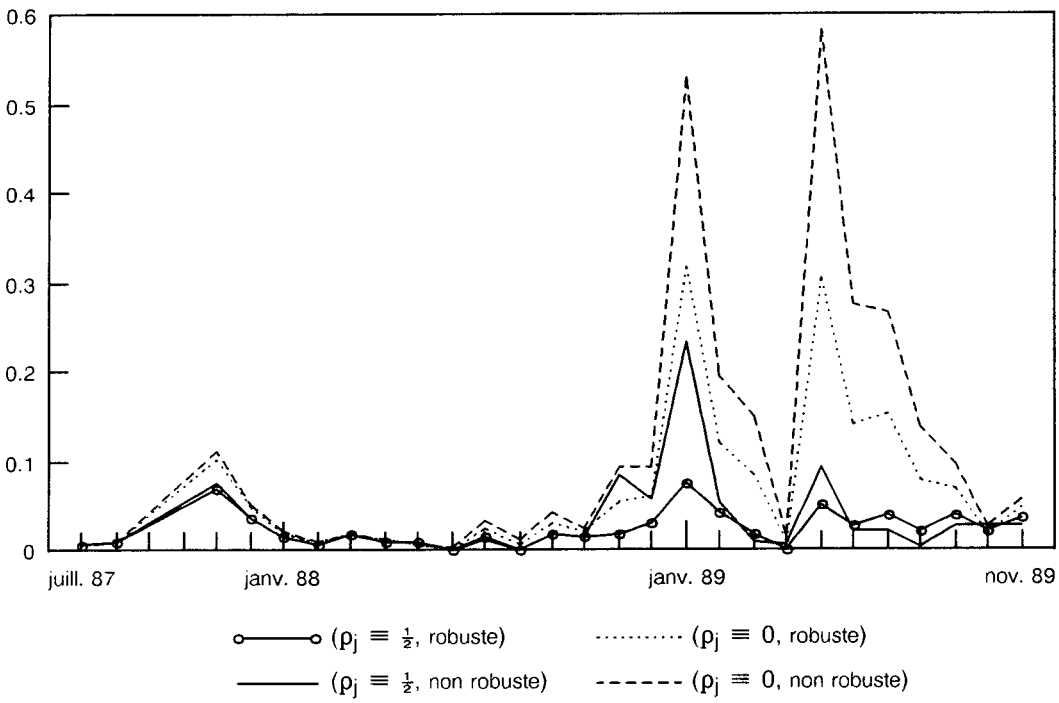

Figure 4 Fluctuation mensuelle des résidus; logements de 5 pièces

($\rho \equiv \frac{1}{2}$, est. non robuste) reprennent leur forme d'avant gonflement beaucoup plus rapidement que celles qui se rapportent au scénario ($\rho \equiv 0$, est. non robuste).

Une autre comparaison intéressante est celle que l'on peut faire entre le cas où on a introduit des contraintes linéaires et celui où on n'en a pas introduites. De toute évidence, l'introduction de contraintes a un effet très favorable lorsqu'on tient compte de l'autocorrélation et cet effet est encore plus favorable lorsque l'autocorrélation est égale à zéro. À cet égard, il est intéressant de comparer les figures qui présentent la fluctuation mensuelle des résidus une étape à l'avance avec celles qui présentent la fluctuation mensuelle des résidus ordinaires. Dans les quatre mois où il y a eu "gonflement" des données, la fluctuation des résidus une étape à l'avance est forte, phénomène parfaitement normal étant donné que ces résidus sont la différence entre les observations et les prédicteurs correspondants des mois précédents. Or, lorsqu'on fait intervenir les contraintes linéaires, la fluctuation revient à son niveau normal dans le mois qui suit le mois du "gonflement". Pour ce qui a trait aux résidus ordinaires, dans l'hypothèse de contraintes linéaires, la fluctuation n'est pas vraiment plus forte dans les mois de "gonflement" pour les logements de 3 pièces et à peine plus forte, si l'on tient compte de l'autocorrélation, pour les logements de 5 pièces. Toutefois, si l'on fait abstraction de l'autocorrélation dans ce dernier cas, la fluctuation des résidus est beaucoup plus forte dans les mois de "gonflement" que dans les autres mois, même lorsqu'on fait intervenir les contraintes. Cela s'explique facilement. En effet, nous savons que les contraintes linéaires s'appliquent aux moyennes globales des valeurs ajustées dans chaque district; or, comme le nombre d'observations pour les logements de 5 pièces ne représente qu'une petite fraction du nombre total d'observations, les contraintes influent relativement peu sur les coefficients de régression estimés pour cette catégorie de logements. En revanche, ces mêmes contraintes ont un effet notable sur les coefficients estimés pour les autres catégories de logement, de sorte que si l'on tient compte de la corrélation transversale, les estimateurs relatifs aux logements de 5 pièces subiront quand même des modifications puisqu'ils sont en corrélation avec les autres coefficients.

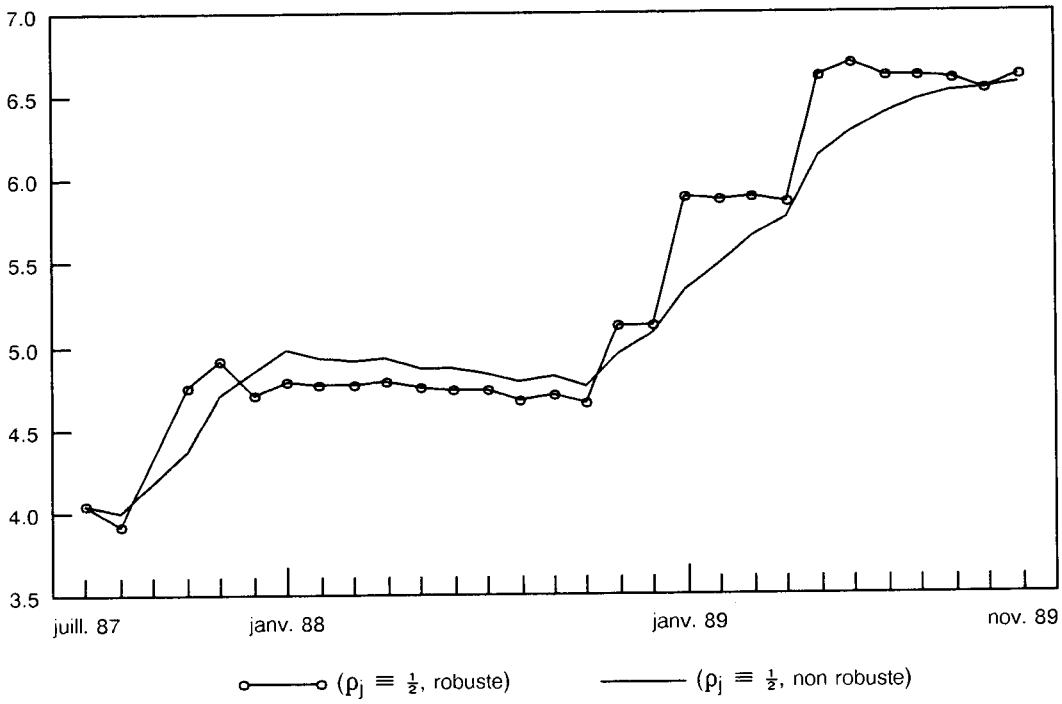

Figure 5 Valeurs estimées mensuelles de l'ordonnée à l'origine; logements de 3 pièces

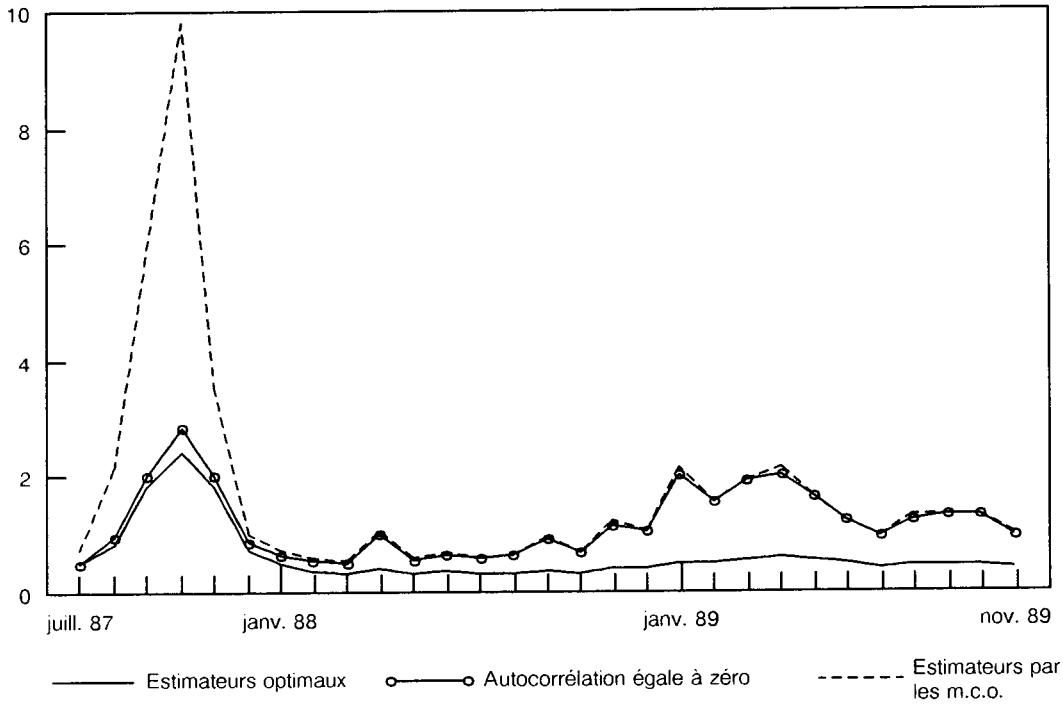

Figure 6 Variance des estimateurs de moyenne de case ($\times 10^4$); logements de 3 pièces

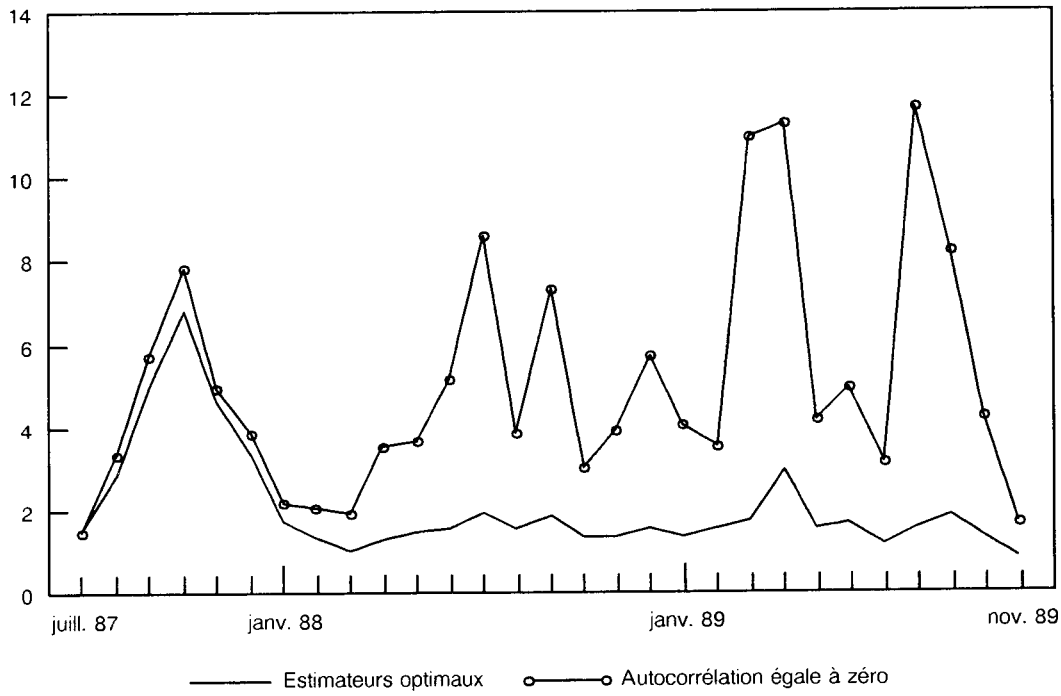

Figure 7 Variance des estimateurs de moyenne de case ($\times 10^4$); logements de 5 pièces

La figure 5 montre dans quelle mesure les contraintes linéaires permettent de prévenir les variations subites de données; pour cela, nous avons représenté graphiquement les valeurs estimées mensuelles de l'ordonnée à l'origine pour les logements de 3 pièces.

Dans l'hypothèse de contraintes linéaires, l'ordonnée à l'origine réagit aussitôt à une variation brusque des données. En l'absence de contraintes, l'adaptation s'étend sur plusieurs mois. Le même graphique pour les logements de 5 pièces n'est pas aussi simple puisque, compte tenu de la faible taille des échantillons mensuels, le "gonflement" des données s'est répercuté sur les autres coefficients de régression.

Jusqu'à maintenant, notre analyse a porté exclusivement sur la distribution empirique des résidus et des résidus une étape à l'avance du modèle. Or, une application importante de l'estimation pour petites régions est la prévision de moyennes de petites régions (équation 2.2). De toute évidence, lorsqu'un modèle produit des résidus qui ont des propriétés "classiques", on peut s'attendre qu'il produise aussi de bons estimateurs pour les moyennes de population. Néanmoins, il serait intéressant de comparer les variances théoriques des estimateurs de moyenne de petite région calculées, les unes suivant l'hypothèse de la corrélation transversale et les autres suivant l'hypothèse de l'absence de corrélation transversale, en vertu du modèle qui tient compte de cette corrélation ($\rho_j \equiv \frac{1}{2}$). Cette comparaison permet de mesurer la perte d'efficacité subie lorsqu'on fait abstraction de l'autocorrélation.

Les figures 6 et 7 présentent les variances mensuelles des estimateurs de moyenne de case pour les logements de 3 et de 5 pièces. (Les variances ont été multipliées par 10^4.) La figure pour les logements de 3 pièces présente, en plus, les variances des estimateurs par les moindres carrés ordinaires (m.c.o.) de la moyenne de population, c'est-à-dire les variances des estimateurs obtenus lorsqu'on estime les coefficients de régression à chaque mois par les m.c.o. Ce genre d'estimateurs n'est pas vraiment utile pour les logements de 5 pièces à cause de la trop petite taille des échantillons mensuels.

Ce qu'il faut retenir surtout de ces deux graphiques, c'est qu'en tenant compte de la corrélation transversale, on peut réduire de façon substantielle (selon la taille de l'échantillon) la variance des estimateurs. On le voit clairement en ce qui a trait aux logements de 5 pièces; cette constatation vaut aussi pour les logements de 3 pièces même si la taille des échantillons dans ce cas est relativement élevée. D'ailleurs, à cause de la taille ordinairement élevée des échantillons de logements de 3 pièces, les estimateurs par les m.c.o. se rapprochent sensiblement des estimateurs que l'on utilise pour estimer les moyennes de population en faisant abstraction de la corrélation transversale. Notons toutefois l'écart appréciable qui sépare la variance de l'estimateur par les m.c.o. de celle des deux autres estimateurs en octobre 1987. Ce mois-là, il n'y a eu que 10 observations pour les logements de 3 pièces; ce cas illustre très bien l'utilité de recourir à des données de périodes antérieures même lorsqu'on fait abstraction de la corrélation transversale. (Le nombre d'observations pour novembre 1987 est de 28; pour tous les autres mois, on compte au moins 46 observations.)

Les deux graphiques nous amènent à une autre conclusion importante: la variance des estimateurs optimaux selon le modèle est beaucoup plus stable que celle des estimateurs qui font abstraction de la corrélation transversale. Il convient de souligner à cet égard que les différences de variances d'un mois à l'autre ne dépendent pas uniquement de la taille des échantillons mensuels mais aussi de la valeur des variables explicatives (matrice de plan) et de la quantité d'observations de périodes antérieures. C'est néanmoins la taille des échantillons qui explique en majeure partie les différences de variance entre les estimateurs, particulièrement vers la fin de la série.

REMERCIEMENTS

Cet article a été rédigé pendant que le premier auteur était en congé sabbatique à Statistique Canada en vertu du programme de bourses de recherche de cet organisme. Les auteurs tiennent à exprimer leur reconnaissance à l'arbitre pour ses commentaires utiles.

ANNEXE

a) Détermination de l'équation (2.12)

Lorsque $\underset{\sim}{x}_{tki} = \underset{\sim}{x}_{tk}$, $\hat{\Theta}_{tk} = \underset{\sim}{x}'_{tk}\hat{\underset{\sim}{\beta}}_{tk} = \underset{\sim}{z}'_{tk}\hat{\underset{\sim}{\alpha}}_{tk}$ de sorte que $\hat{\underset{\sim}{\Theta}}_t = (\hat{\Theta}_{t1}, \ldots, \hat{\Theta}_{tK})' = Z_t\hat{\underset{\sim}{\alpha}}_t$. En outre, pour ce qui a trait au modèle de trajet aléatoire, la matrice T est la matrice unité et, d'après l'équation (3.1),

$$Z_t\hat{\underset{\sim}{\alpha}}_t = Z_t\hat{\underset{\sim}{\alpha}}_{t-1} + \left(Z_tP_{t|t-1}Z'_t\right)F_t^{-1}\left(\underset{\sim}{Y}_t - Z_t\hat{\underset{\sim}{\alpha}}_{t-1}\right) =$$

$$\left(I - \Sigma_tF_t^{-1}\right)\underset{\sim}{Y}_t + \Sigma_tF_t^{-1}Z_t\hat{\underset{\sim}{\alpha}}_{t-1} \tag{A1}$$

puisque $F_t = (Z_tP_{t|t-1}Z'_t + \Sigma_t)$. Supposons, pour des raisons de commodité, que $k = 1$ et posons

$$F_t = \begin{bmatrix} f_{11}, \underset{\sim}{f}'_1 \\ \underset{\sim}{f}_1, F_{22} \end{bmatrix} \quad \text{et} \quad H_t = F_t^{-1} = \begin{bmatrix} h_{11}, \underset{\sim}{h}'_1 \\ \underset{\sim}{h}_1, H_{22} \end{bmatrix} \quad \text{où} \quad f_{11} \quad \text{et} \quad h_{11}$$

sont des grandeurs scalaires, $\underset{\sim}{f}'_1$ et $\underset{\sim}{h}'_1$ sont de dimension $[1 \times (K-1)]$ et F_{22} et H_{22} sont de dimension $[(K-1) \times (K-1)]$. En nous servant de la notation ci-dessus, nous pouvons déduire de l'expression (A1) l'équation suivante:

$$\hat{\Theta}_{t1} = \left(1 - \frac{\sigma_1^2}{n_{t1}}h_{11}\right)\bar{Y}_{t1} + \frac{\sigma_1^2}{n_{t1}}h_{11}\left(\underset{\sim}{x}'_{t1}\hat{\underset{\sim}{\beta}}_{t-1,1}\right) - \frac{\sigma_1^2}{n_{t1}}\sum_{k=2}^{K}h_{11}\frac{h_{1k}}{h_{11}}\bar{e}_{tk}. \tag{A2}$$

Désignons par $\underset{\sim}{\gamma}'_1 = (\gamma_{12}, \ldots, \gamma_{1K}) = \underset{\sim}{f}'_1F_{22}^{-1}$ les coefficients de régression partielle dans la régression de $\bar{e}_{t1}$ par rapport à $(\bar{e}_{t2}, \ldots, \bar{e}_{tK})$, et désignons par $v_1^2 = (f_{11} - \underset{\sim}{f}'_1F_{22}^{-1}\underset{\sim}{f}_1)$ la variance des résidus de la régression.

L'équation (2.12) découle directement de (A2) puisque

$$\underset{\sim}{f}'_1F_{22}^{-1} = -\frac{1}{h_{11}}\underset{\sim}{h}'_1; \quad \left(f_{11} - \underset{\sim}{f}'_1F_{22}^{-1}\underset{\sim}{f}_1\right)^{-1} = h_{11} \tag{A3}$$

d'après les propriétés de l'inverse d'une matrice partitionnée.

b) Détermination de l'équation (4.4)

D'après (4.3),

$$\hat{\underset{\sim}{\alpha}}_t^{(A)} = \left(I - K_t^{(P)}Z_t^{(A)}\right)T\hat{\underset{\sim}{\alpha}}_{t-1}^{(A)} + K_t^{(P)}\underset{\sim}{Y}_t^{(A)}. \tag{A4}$$

Par conséquent,

$$\hat{\underset{\sim}{\alpha}}_t^{(A)} - \underset{\sim}{\alpha}_t = \left(I - K_t^{(P)}Z_t^{(A)}\right)\left(T\hat{\underset{\sim}{\alpha}}_{t-1}^{(A)} - \underset{\sim}{\alpha}_t\right) + K_t^{(P)}\left(\underset{\sim}{Y}_t^{(A)} - Z_t^{(A)}\underset{\sim}{\alpha}_t\right). \tag{A5}$$

Les erreurs de prédiction $\left(T\hat{\underset{\sim}{\alpha}}_{t-1}^{(A)} - \underset{\sim}{\alpha}_t\right)$ sont indépendantes des résidus $\left(\underset{\sim}{Y}_t^{(A)} - Z_t^{(A)}\underset{\sim}{\alpha}_t\right)$; par conséquent,

$$P_t^{(A)} = E\left[\left(\hat{\underset{\sim}{\alpha}}_t^{(A)} - \underset{\sim}{\alpha}_t\right)\left(\hat{\underset{\sim}{\alpha}}_t^{(A)} - \underset{\sim}{\alpha}_t\right)'\right] = Q_tP_{t|t-1}^{(A)}Q'_t + K_t^{(P)}\Sigma_t^{(A)}K_t^{(P)\prime} \tag{A6}$$

où, pour des raisons de commodité, nous avons posé $Q_t = (I - K_t^{(P)}Z_t^{(A)})$.

D'après la définition de la matrice $F_t^{(P)}$ (voir au-dessous de l'équation 4.3), nous pouvons réécrire l'équation (A6) de la façon suivante:

$$P_t^{(A)} = Q_t P_{t|t-1}^{(A)} - P_{t|t-1}^{(A)} Z_t^{(A)'} K_t^{(P)'} + K_t^{(P)} F_t^{(P)} K_t^{(P)'} + K_t^{(P)} \left(\Sigma_t^{(A)} - \Sigma_t^{(P)} \right) K_t^{(P)'} \quad \text{(A7)}$$

par des transformations algébriques simples, l'expression ci-dessus devient l'équation (4.4).

BIBLIOGRAPHIE

ANDERSON, B.O.D., et MOORE, J.B. (1979). *Optimal Filtering*. Englewood Cliffs: Prentice-Hall.

ANSLEY, C.F., et KOHN, R. (1986). Prediction mean squared error for State Space models with estimated parameters. *Biometrika*, 73, 467-473.

BATTESE, G.E., HARTER, R.M., et FULLER, W.A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83, 28-36.

BINDER, D.A., et DICK, J.P. (1989). Enquêtes répétées - Modélisation et estimation. *Techniques d'enquête*, 15, 31-48.

BINDER, D.A., et HIDIROGLOU, M.A. (1988). Sampling in time. *Handbook of Statistics*, Vol. 6, (éds.), P.R. Krishnaiah et C.R. Rao, Amsterdam: Elsevier Science, 187-211.

CHOUDHRY, G.H., et RAO, J.N.K. (1989). Small area estimation using models that combine time series and cross-sectional data. *Proceedings of the Statistics Canada Symposium on Analysis of Data in Time* (à paraître).

COOLEY, T.F., et PRESCOTT, E.C. (1976). Estimation in the presence of stochastic parameter variation. *Economitrica*, 44, 167-184.

DIELMAN, T.E. (1983). Pooled cross-sectional and time series data: A survey of current statistical methodology. *The American Statistician*, 37, 111-122.

HAMILTON, J.D. (1986). A standard error for the estimated state vector of a State-Space model. *Journal of Econometrics*, 33, 388-397.

HARVEY, A.C. (1981). *Time Series Models*. Deddington, Oxford: Philip Allan.

HARVEY, A.C., et PHILLIPS, G.D.A. (1979). Maximum likelihood estimation of regression models with autoregressive-moving average disturbances. *Biometrika*, 66, 49-58.

KITAGAWA, G., et GERSCH, W. (1984). A smoothness priors State-Space modeling of time series with trend and seasonality. *Journal of the American Statistical Association*, 79, 378-389.

JOHNSON, L.W. (1977). Stochastic parameter regressions: An annotated bibliography. *Revue Internationale de Statistique*, 45, 257-272.

JOHNSON, L.W. (1980). Stochastic parameter regression: An additional annotated bibliography. *Revue Internationale de Statistique*, 48, 95-102.

LaMOTTE, L.R., et McWHORTER, A. (1977). Estimation, testing and forecasting with random coefficient regression models. *Proceedings of the Section on Business and Economic Statistics, American Statistical Association*, 814-817.

MADDALA, G.S. (1977). *Econometrics*. Kogakusta: McGraw-Hill.

MEINHOLD, R.J., et SINGPURWALLA, N.D. (1983). Understanding the Kalman filter. *The American Statistician*, 37, 123-127.

PFEFFERMANN, D. (1991). Estimation and seasonal adjustment of population means using data from repeated surveys. *Journal of Business and Economic Statistics*, 163-175.

PFEFFERMANN, D.,et BARNARD, C. (1991). Some new estimators for small area means with application to the assessment of farmland values. *Journal of Business and Economic Statistics* (à paraître).

PFEFFERMANN, D., BURCK, L., et BEN-TUVIA, S. (1989). A time series model for estimating housing price indexes adjusted for changes in quality. *L'analyse des données dans le temps*. Ottawa: Statistique Canada (à paraître).

PFEFFERMANN, D., et SMITH, T.M.F. (1985). Regression models for grouped populations in cross-section surveys. *Revue Internationale de Statistique*, 53, 37-59.

ROSENBERG, B. (1973a). The analysis of cross section of time series by stochastically convergent parameter regression. *Annals of Economic and Social Measurement*, 2, 399-428.

ROSENBERG, B. (1973b). A survey of stochastic parameter regression. *Annals of Economic and Social Measurement*, 2, 381-397.

SÄRNDAL, C.E., et HIDIROGLOU, M.A. (1989). Small domain estimation: A conditional analysis. *Journal of the American Statistical Association*, 84, 266-275.

SCHWEPPE, F. (1965). Evaluation of likelihood functions for gaussian signals. *IEE Transactions on Information Theory*, 11, 61-70.

SMITH, T.M.F. (1978). Principles and problems in the analysis of repeated surveys. *Survey Sampling and Measurement*, (éd.) N.K. Nawboodivi, New York: Academic Press, 201-216.

SWAMY, P.A.V.B. (1971). *Statistical Inference in Random Coefficient Regression Models*. Berlin: Springer-Verlag.

TILLER, R. (1989). A Kalman filter approach to labour force estimation using survey data. *Proceedings of the Section on Business and Economic Statistics, American Statistical Association* (à paraître).

WATSON, M.W., et ENGLE, R.F. (1983). Alternative algorithms for the estimation of dynamic factor, mimic and varying coefficient regression models. *Journal of Econometrics*, 23, 385-400.