Techniques d'enquête, juin 1989
Vol. 15, n° 1, pp. 31–48
Statistique Canada

Enquêtes répétées – Modélisation et estimation

D.A. BINDER et J.P. DICK[1]

RÉSUMÉ

Les auteurs examinent brièvement l'estimation de la moyenne d'une caractéristique pour une population à différentes périodes à partir d'une série d'enquêtes successives. En définissant un modèle paramétrique stochastique pour ces moyennes, il est possible d'en estimer les paramètres et d'obtenir des estimateurs des moyennes proprement dites. Les auteurs exposent le cas où les moyennes de population suivent un processus autorégressif de moyennes mobiles (ARMA) et où les erreurs de sondage peuvent aussi être exprimées par un tel processus. Enfin, les auteurs ont recours à un exemple où ils utilisent des données de l'enquête sur les voyages des Canadiens.

MOTS CLÉS: Filtre de Kalman; enquêtes avec participation répétée; modèles d'espace d'états; modèles de séries chronologiques; estimations pour petites régions.

1. INTRODUCTION

Une enquête qui revient périodiquement permet l'application de méthodes d'estimation et d'analyse qui ne conviennent pas pour une enquête unique. Par exemple, l'efficacité des méthodes d'estimation utilisées à une occasion peut dépendre des données recueillies les fois précédentes. Cela se produit lorsque des unités d'échantillonnage participent à deux enquêtes successives et que, par conséquent, on peut établir une corrélation entre les erreurs de sondage d'une période à l'autre. De même, les utilisateurs de données définissent souvent des modèles pour les séries d'estimations tirées d'enquêtes répétées. Par exemple, on suppose souvent un modèle autorégressif de moyennes mobiles (ARMA). Cependant, la plupart des méthodes qui permettent actuellement d'estimer les paramètres de ce modèle supposent que les données d'entrée ne sont pas exposées à l'erreur de sondage.

Dans cet article, nous élaborons des méthodes permettant d'estimer ces paramètres lorsque les données renferment des erreurs de sondage. La structure de covariance des erreurs que nous considérons comprend des cas où les erreurs de sondage sont corrélées dans le temps.

Lorsqu'on suppose un modèle de ce genre pour décrire le mouvement des caractéristiques de population, il est possible de déterminer l'estimateur linéaire à erreur quadratique moyenne minimum (EQMM). Cet estimateur admet la structure de modèle que l'estimateur linéaire non biaisé à variance minimum (ELNVM) n'admet pas. Nous reviendrons sur l'ELNVM dans la Section 2.

Blight et Scott (1973), Scott et Smith (1974), Scott, Smith et Jones (1977), R.G. Jones (1980) et d'autres encore ont examiné les conséquences de modèles stochastiques de ce genre pour les moyennes de population dans le temps. Les résultats de leur analyse ainsi qu'une formulation plus générale faisant intervenir des modèles d'espace d'états et des filtres de Kalman sont étudiés dans la Section 3 pour le cas où le modèle stochastique pour les caractéristiques de population est défini entièrement. Ces méthodes peuvent être élaborées dans le cadre d'un modèle de type bayesien, où la distribution a priori est définie entièrement.

[1] D.A.Binder et J.P. Dick, Division des méthodes d'enquêtes sociales, Statistique Canada, Ottawa, Ontario, K1A 0T6.

Lorsqu'on suppose un modèle ARMA en présence d'erreurs de sondage, on peut recourir aux modèles d'espace d'états pour calculer les estimations les plus vraisemblables des paramètres inconnus. Notons que cette méthode peut être assimilée à une méthode empirique de Bayes. Nous supposons que les erreurs de sondage peuvent être décrites par un processus ARMA jusqu'à un facteur multiplicatif. C'est ce que nous voyons dans la Section 4.

Dans la Section 5, nous donnons un exemple de ce modèle en nous servant des données de l'enquête sur les voyages des Canadiens. Cet exemple montre comment le fait de tenir compte des erreurs de sondage influe sur les valeurs estimées des paramètres du modèle. Par la même occasion, nous calculons des estimations lissées suivant les hypothèses du modèle. Comme, dans cet exemple, les erreurs de sondage sont indépendantes, nous n'avons pas à exposer tout l'appareil mathématique de la formulation générale. Cependant, l'exemple montre que le fait de ne pas tenir compte des erreurs de sondage peut avoir des conséquences appréciables même dans ces conditions.

La Section 6 renferme les conclusions de l'étude.

2. ESTIMATION LINÉAIRE NON BIAISÉE À VARIANCE MINIMUM DANS LES ENQUÊTES RÉPÉTÉES AVEC PARTICIPATION RÉPÉTÉE DE CERTAINES UNITÉS

Dans cette section, nous faisons un survol des articles où l'on considère la valeur d'une caractéristique de population comme la moyenne ou le total comme une constante inconnue. Dans la Section 3, nous étudions le cas où l'on suppose un modèle stochastique pour la caractéristique de population.

Dans les enquêtes à participation répétée, où des personnes sont appelées à participer plus d'une fois à la même enquête, les erreurs d'échantillonnage sont habituellement corrélées d'une fois à l'autre. Il peut aussi y avoir corrélation dans les enquêtes à plusieurs degrés où quelques-unes des unités d'échantillonnage du premier degré participent plus d'une fois, même si en définitive il ne s'agit pas des mêmes répondants.

Les estimateurs qui ne tiennent pas compte de ces corrélations et ne reposent que sur les données recueillies dans une seule période de référence sont en règle générale inefficaces comparativement à l'estimateur linéaire non biaisé à variance minimum (ELNVM). L'efficacité relative dépend du degré de corrélation entre les erreurs d'échantillonnage d'une enquête à l'autre. Lorsque le degré de corrélation est nul, comme c'est le cas dans l'exemple de la Section 5, l'ELNVM est simplement l'estimateur fondé sur les données d'une seule période de référence.

C'est Jessen (1942) qui, le premier, a utilisé les données fournies par la même personne à deux enquêtes successives. Patterson (1950) a élaboré une théorie générale pour les enquêtes répétées auxquelles certaines unités participent plus d'une fois. Il a examiné en détail le cas particulier d'un échantillon aléatoire simple tiré d'une population infinie, où la corrélation pour les personnes décroît exponentiellement d'une période à l'autre. À chaque période d'enquête, un certain nombre de personnes sont retranchées de l'échantillon de la période précédente puis d'autres y sont ajoutées. Les données recueillies concernent uniquement la période courante. Patterson a calculé l'ELNVM pour ce cas particulier.

Les hypothèses fondamentales de Patterson (1950) ont été reprises puis élargies. Eckler (1955) a désigné le plan de Patterson comme un échantillonnage avec renouvellement à un degré. Il a déterminé l'ELNVM pour le cas où des personnes fournissent des renseignements pour deux périodes successives, ce qu'il a appelé l'échantillonnage avec renouvellement à deux degrés. Il a aussi calculé l'ELNVM pour des plans d'échantillonnage avec renouvellement de degré supérieur.

Rao et Graham (1964) ont assoupli l'hypothèse de la population infinie en itégrant le facteur de correction pour population finie dans la variance de l'erreur de sondage. Singh (1968) a été le premier à considérer des plans à plusieurs degrés. Il a examiné des plans de sondage à deux degrés en supposant que la corrélation entre les réponses fournies à des périodes

différentes peut se décomposer en deux éléments: i) la corrélation entre les unités du second degré (USE) tirées des unités primaires d'échantillonnage (UPÉ) et ii) la corrélation entre les moyennes d'UPÉ de périodes successives. Si on suppose que ces deux formes de corrélation suivent un processus autorégressif du premier degré, la forme de l'ELNVM correspond à la forme générale définie par Patterson (1950).

Tikkiwal (1979) et d'autres ont examiné ce qui arriverait si on assouplissait l'hypothèse ci-dessus. Tikkiwal a conclu que si on supposait l'existence d'une structure de corrélation entièrement générale, la forme simple de l'ELNVM disparaîtrait et il faudrait recourir, en pratique, à l'approximation. Rao et Graham (1964) et Gurney et Daly (1965) ont proposé l'utilisation d'estimateurs composites, qui sont une approximation des estimateurs optimaux. Ces estimateurs s'utilisent facilement et ont une efficacité relative élevée. Pour une étude de l'utilisation de ces estimateurs, voir Binder et Hidiroglou (1988).

Gurney et Daly (1965) ont de plus étendu les résultats de Patterson (1950) à un modèle linéaire. Ils ont défini la notion d'«estimation élémentaire». Il s'agit d'une estimation fondée sur des données se rattachant à une période précise et recueillies auprès d'un groupe de personnes qui sont retranchées de l'échantillon ou y sont intégrées ensemble. On peut exprimer l'espérance mathématique de ces estimations élémentaires comme une combinaison linéaire des paramètres de population, $\{\theta_t\}$. Lorsqu'on connaît la structure de corrélation, on peut calculer l'ELNVM à l'aide de la théorie générale des modèles linéaires.

Afin de mathématiser cette analyse, posons y_{tj} comme la j-ème estimation élémentaire rattachée à la période $E(y_{tj}) = \theta_t$. Si $\boldsymbol{Y}$ et $\boldsymbol{\Theta}$ sont des vecteurs ayant pour éléments y_{tj} et θ_t respectivement, nous pouvons écrire :

$$\boldsymbol{Y} = \boldsymbol{X}'\boldsymbol{\Theta} + \boldsymbol{e}, \tag{2.1}$$

où $\boldsymbol{X}$ est une matrice fixe $(n \times T)$ formée de zéros et de uns, $E(\boldsymbol{e}) = 0$ et $E(\boldsymbol{ee}') = \boldsymbol{U}$, qui est la matrice connue des variances-covariances des estimations élémentaires. Donc, l'ELNVM est défini par l'équation suivante:

$$\tilde{\boldsymbol{\Theta}} = (\boldsymbol{X}'\boldsymbol{U}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{U}^{-1}\boldsymbol{Y}, \tag{2.2a}$$

où

$$\mathrm{Var}(\tilde{\boldsymbol{\Theta}}) = (\boldsymbol{X}'\boldsymbol{U}^{-1}\boldsymbol{X})^{-1} . \tag{2.2b}$$

Ces résultats supposent qu'à chaque nouvelle enquête il faut mettre à jour toutes les estimations précédentes. Toutefois, comme les estimations tirées des enquêtes plus reculées ont souvent un effet beaucoup moindre que les estimations des enquêtes récentes, les estimateurs composites, comme ceux proposés par Gurney et Daly (1965), sont plus faciles à utiliser et ont une efficacité relative élevée. Binder et Hidiroglou (1988) ont étudié la pertinence de ces méthodes ainsi que leur application dans un certain nombre d'enquêtes. De façon générale, ils ont constaté que l'on peut obtenir de bons résultats avec des estimateurs composites, pourvu que les biais de renouvellement ne soient pas trop élevés.

3. DISSOCIATION DU SIGNAL ET DU BRUIT

Les économistes et les sociologues ont souvent tendance à considérer les paramètres $\{\theta_t\}$ comme des variables aléatoires dans leurs modèles stochastiques (Smith 1978). Cependant, si on ne tient pas compte des erreurs d'échantillonnage liées aux données d'entrée, les valeurs estimées des paramètres du modèle stochastique se trouvent biaisées.

Dans la présente section, nous allons montrer comment les hypothèses du modèle stochastique peuvent aussi permettre d'obtenir des estimateurs fondés sur un modèle et convergents selon un plan. Dans la Section 4, il est question de l'estimation des paramètres du modèle par la méthode du maximum de vraisemblance. Comme une définition erronée du modèle peut entraîner des biais appréciables, on devrait recourir à des méthodes de vérification d'hypothèses pour s'assurer que le modèle est conforme aux données. Le modèle devrait en outre refléter une connaissance approfondie du phénomène en question.

Voyons tout d'abord le cas où les erreurs de sondage sont indépendantes. (Cette hypothèse est vraisemblable dans le cas des enquêtes sans participation répétée avec de faibles taux de sondage.) En l'occurrence, l'ELNVM de θ_t est $\tilde{\theta}_t = y_t$. Or, en appliquant un modèle stochastique à la suite des paramètres , $\{\theta_t\}$, il est possible de réduire l'erreur quadratique moyenne de l'estimateur.

Scott et Smith (1974) ont proposé le modèle suivant pour les enquêtes sans participation répétée. Leur modèle des estimations d'enquête au temps t s'écrit comme suit:

$$y_t = \theta_t + e_t \tag{3.1}$$

où les résidus e_t's sont indépendants et distribués suivant une loi normale de moyenne nulle et de variancere independent $N(0,S_t^2)$. Ils ont supposé que la suite des paramètres, $\{\theta_t\}$peut être modélisée sous la forme suivante (à la condition que $\mathbf{\Theta}'_{t-1} = (\theta_1, \ldots \theta_{t-1})$,

$$\theta_t = \underline{\alpha}_t' \mathbf{\Theta}_{t-1} + \epsilon_t, \tag{3.2}$$

où les ϵ_t's sont indépendants et distribués suivant une loi normale de moyenne nulle et de variance $N(0,S_t^2)$ et sont indépendants de $\{e_t\}$, et où $\underline{\alpha}_t$ est un vecteur de constantes de dimension $(t-1)$.

En règle générale, au temps $t-1$, nous avons $Y'_{t-1} = (y_1, \ldots, y_{t-1})$, à la condition $\mathbf{\Theta}_{t-1} \sim N(\tilde{\mathbf{\Theta}}_{t-1}, \tilde{V}_{t-1})$. Par des arguments conditionnels, nous pouvons montrer que

$$\mathrm{E}(\theta_t|y_t) = \tilde{\theta}_t = \pi_t(\underline{\alpha}_t'\tilde{\mathbf{\Theta}}_{t-1}) + (1-\pi_t)y_t \tag{3.3a}$$

et que

$$\mathrm{Var}(\theta_t|y_t) = (1-\pi_t)S_t^2, \tag{3.3b}$$

où

$$\pi_t = \frac{\mathrm{Var}(y_t|\theta_t)}{\mathrm{Var}(y_t)} = \frac{S_t^2}{\underline{\alpha}_t'\tilde{V}_{t-1}\,\underline{\alpha}_t + \sigma_t^2 + S_t^2}. \tag{3.3c}$$

Notons que l'estimateur défini en (3.3a) est la moyenne pondérée de deux éléments: le premier représente la meilleure prévision linéaire de θ_t étant donné la valeur précédente de $\tilde{\mathbf{\Theta}}_{t-1}$; et le second représente la meilleure estimation de θ_t établie à l'aide des données de l'enquête. L'importance de chaque terme est déterminée par π_t, qui est le rapport entre la variance de

l'échantillon et la variance totale. Lorsque l'erreur de sondage diminue, l'importance de $\tilde{\boldsymbol{\Theta}}_{t-1}$ diminue aussi et l'estimation de θ_t dans (3.3a) se compose principalement de y_t, qui est l'estimation établie à l'aide des données d'enquête. Par conséquent, l'estimateur de θ_t est convergent selon le plan lorsque y_t l'est.

En revanche, lorsque l'erreur de sondage augmente, l'estimation de θ_t est déterminée principalement par la prévision linéaire de $\boldsymbol{\Theta}_{t-1}$. L'efficacité relative de l'estimateur, $\tilde{\theta}_t$, dans (3.3a) est définie par $1/(1 - \pi_t)$, où π_t est défini en (3.3c). On obtient les gains d'efficacité les plus élevés lorsque l'erreur de sondage est grande par rapport à σ_t^2, qui est la variance des perturbations du processus.

Scott et Smith (1974) et R.G. Jones (1980) se sont aussi intéressés aux enquêtes avec participation répétée. Jones a formulé son analyse dans les termes suivants. Soit $\boldsymbol{\Theta}_t$ une variable aléatoire normale multidimensionnelle avec une moyenne nulle et une matrice des variances V_t^*. Par ailleurs, on peut généraliser les observations au temps t en un vecteur d'estimations élémentaires, y_t. La distribution conditionnelle de $Y_t = (y_1', \ldots, y_t')'$ étant donné Θ_t est supposée être de la forme:

$$Y_t = X_t' \boldsymbol{\Theta}_t + e_t, \tag{3.4}$$

où X_t est une matrice fixe formée de zéros et de uns qui fait le lien entre les paramètres et les observations et e_t est l'erreur de sondage, qui est supposée être une variable aléatoire normale multidimensionnelle avec une moyenne nulle et une matrice des covariances U_t.

À l'aide d'arguments conditionnels, on peut montrer que la meilleure estimation de θ_t étant donné Y_t est:

$$\mathrm{E}(\boldsymbol{\Theta}_t | Y_t) = \tilde{\boldsymbol{\Theta}}_t = (X_t' U_t^{-1} X_t + \tilde{V}_t^{*-1})^{-1} X_t' U_t^{-1} Y_t \tag{3.5a}$$

avec comme variance

$$\mathrm{Var}(\boldsymbol{\Theta}_t | Y_t) = (X_t' U_t^{-1} X_t + V_t^{*-1})^{-1}. \tag{3.5b}$$

Ce résultat est très général. Si jamais le modèle stochastique pour $\boldsymbol{\Theta}_t$ était très faible, l'inverse de V_t^* serait approximativement 0, ce qui donnerait l'ELNVM défini en (2.2a). R.G. Jones (1980) a obtenu les équations (3.5) à l'aide des moindres carrés stochastiques, de sorte que l'estimateur $\tilde{\boldsymbol{\Theta}}_t$ est l'estimateur linéaire à erreur quadratique moyenne minimum (EQMM) même en l'absence des hypothèses de normalité.

Si on appliquait directement (3.5), il faudrait inverser des matrices qui sont de même dimension que le vecteur de toutes les estimations élémentaires pour toutes les périodes. Calculer l'inverse de telles matrices pourrait donner des résultats incertains. Cependant, on peut souvent redéfinir l'expression (3.5) à l'aide des modèles d'espace d'états, qui permettent de décrire de nombreux modèles de séries chronologiques. Voir Harvey (1984) pour une analyse de ces modèles. Comme nous allons le voir ci-dessous, l'utilisation des modèles d'espace d'états nous évite d'inverser de grandes matrices. Pour tirer parti de la réduction de dimensions que permet l'utilisation des modèles d'espace d'états, il est nécessaire de définir une structure pour $\{\theta_t\}$ et $\{e_t\}$. On pourrait par exemple choisir un processus autorégressif de moyennes mobiles (ARMA) qui n'est pas nécessairement homogène dans le temps; ce genre de structure est souvent utilisé dans les applications de séries chronologiques.

Pour des applications comme le calcul d'estimations pour petites régions, où les tailles d'échantillons sont relativement faibles, il peut être utile de modéliser les variances de l'erreur de sondage, U_t à l'aide de processus ARMA. On ne procède pas de cette façon habituellement dans le cas des enquêtes répétées. L'utilisation de modèles ARMA faciliterait aussi l'application directe de l'équation (3.5) lorsque les dimensions de, V_t^* et U_t sont grandes et que les inverses sont numériquement instables.

Dans le modèle d'espace d'états, deux processus se déroulent simultanément. Le premier, qui est le système d'observation, décrit en détail comment les observations dépendent de l'état des paramètres du processus dans la période observée. Le second, qui est le système de transition, décrit en détail l'évolution des paramètres.

On peut formuler les modèles d'espace d'états de la façon suivante. L'équation d'observation s'écrit:

$$y_t = \boldsymbol{H}_t z_t + \underline{\omega}_t, \tag{3.6a}$$

et l'équation de transition s'écrit:

$$z_t = \boldsymbol{F}_{t-1} z_t + \boldsymbol{G}_t \underline{\epsilon}_t, \tag{3.6b}$$

où z_t est un vecteur d'états $(r \times 1)$, $\boldsymbol{H}_t$ est une matrice fixe $(n_t \times r)$, $\boldsymbol{F}_t$ est une matrice de transition fixe $(r \times r)$, $\boldsymbol{G}_t$ est une matrice fixe $(r \times m)$ et $\underline{\omega}_t$ et $\underline{\epsilon}_t$ sont des bruits aléatoires indépendants de moyenne nulle et de covariances $\mathrm{E}(\underline{\omega}_t \underline{\omega}_t') = \boldsymbol{U}_t$ et $\mathrm{E}(\underline{\epsilon}_t \underline{\epsilon}_t') = \boldsymbol{V}_t$.

À titre d'exemple, nous allons transformer le modèle analysé par Blight et Scott (1973) en un modèle d'espace d'états. Blight et Scott ont considéré les données du plan d'échantillonnage avec renouvellement à un degré de Patterson (1950). Ils ont défini $\bar{y}_t''$ comme la moyenne des nouvelles unités au temps t, et $\bar{y}'$ et $\bar{x}_t'$ comme les moyennes, au temps t et $t - 1$, respectivement, des unités ayant déjà participé. Ils ont supposé que $\bar{y}_t''$ et $\bar{y}_t' - \rho\bar{x}_t'$ sont des observations indépendantes au temps t, où ρ est la corrélation entre les réponses fournies par la même personne d'une fois à l'autre. Ils ont aussi supposé que la moyenne $\{\theta_t\}$ suit un processus autorégressif du premier degré.

Soit le vecteur d'états $z_t' = (\theta_t, \theta_{t-1})$. Nous pouvons écrire l'équation d'observation comme suit:

$$\begin{bmatrix} \hat{y}_t'' \\ \bar{y}_t' - \rho\bar{x}_t' \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & -\rho \end{bmatrix} \begin{bmatrix} \theta_t \\ \theta_{t-1} \end{bmatrix} + \begin{bmatrix} \omega_{1t} \\ \omega_{2t} \end{bmatrix},$$

où $(\omega_{1t}, \omega_{2t})'$ a une matrice des covariances diagonale.

L'équation de transition s'écrit:

$$\begin{bmatrix} \theta_t \\ \theta_{t-1} \end{bmatrix} = \begin{bmatrix} \alpha & 0 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \theta_{t-1} \\ \theta_{t-2} \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} \epsilon_t,$$

où ϵ_t est distribué suivant une loi $N(0,\sigma^2)$. Par conséquent, nous pouvons exprimer le modèle de Blight-Scott sous forme de modèle d'espace d'états.

Harvey et Phillips (1979) ont décrit une méthode permettant d'exprimer le modèle ARMA (p,q), défini par:

$$y_t - \alpha_1 y_{t-1} - \ldots - \alpha_p y_{t-p} = \epsilon_t - \beta_1 \epsilon_{t-1} - \ldots - \beta_q \epsilon_{t-q}, \tag{3.7}$$

où les ϵ_t's sont indépendants et distribués suivant une loi $N(0,\sigma^2)$, sous forme de modèle d'espace d'états. La dimension de z_t est $r = \mathrm{MAX}(p,q+1)$. Lorsque cela est nécessaire, on ajoute des zéros à, $\underline{\alpha} = (\alpha_1, \ldots, \alpha_p)$ ou à $\underline{\beta} = (\beta_1, \ldots, \beta_q)$ pour obtenir des vecteurs de dimension r. La matrice U_t est définie comme une matrice nulle. Le modèle ARMA(p,q) équivaut à (3.6) lorsque $\boldsymbol{H}_t' = (1, 0, \ldots, 0)$, $\boldsymbol{G}_t' = (1, -\beta_1, \ldots, -\beta_{r-1})$ et

$$F_t = \left[\begin{array}{c|c} \alpha_1 & \\ \vdots & I_{r-1} \\ \alpha_{r-1} & \\ \hline \alpha_r & O' \end{array}\right],$$

où I_{r-1} est la matrice unité de dimensions $(r-1) \times (r-1)$ O' est un vecteur ligne formé de zéros.

Selon cette méthode, le vecteur d'états $z_t = (z_{1t}, \ldots, z_{rt})'$ est défini comme suit:

$$z_{it} = \alpha_i y_{t-1} + \alpha_{i+1} y_{t-2} + \ldots + \alpha_r y_{t-(r-i+1)}$$

$$- \beta_{i-1}\epsilon_t - \beta_i \epsilon_{t-1} - \ldots - \beta_{r-1}\epsilon_{t-(r-i)},$$

pour $i = 2, 3, \ldots, r$ et $z_{1t} = y_t$ tel que défini en (3.7).

Une condition nécessaire pour la stationnarité est que $\mathrm{Var}(z_t) = \mathrm{Var}(z_{t-1})$ pour tous les t. D'après l'expression (3.6b), nous voyons que cette condition implique:

$$\mathrm{Var}(z) = \boldsymbol{F}'\mathrm{Var}(z)\boldsymbol{F} + \boldsymbol{G}\boldsymbol{V}\boldsymbol{G}',$$

où $V_t \equiv V$ est constante pour tous les t. Pearlman (1980) a indiqué que cela pouvait servir à définir les conditions initiales pour z_1.

On peut souvent inclure le processus des erreurs de sondage dans le modèle d'espace d'états lorsqu'il est possible de supposer une structure pour ces erreurs de sondage. Nous l'avons déjà démontré en ce qui concerne le modèle de Blight et Scott (1973). Scott et Smith (1974) et Miazaki (1985) ont examiné une série de modèles qui étaient des cas particuliers d'un processus (θ_t) ARMA (p,q), pour $\{e_t\}$ et d'un processus ARMA (p^*,q^*) et ont étudié les observations scalaires qui satisfont $y_t = \theta_t + e_t$. Les modèles d'espace d'états pour ces processus peuvent être formulés suivant la méthode de Harvey et Phillips décrite ci-dessus; en l'occurrence, le vecteur d'états z_t est le résultat de l'enchaînement des vecteurs d'états de chaque processus ARMA.

Supposons, par exemple, que $\{\theta_t\}$ suit un processus ARMA(3,0) avec des paramètres $(\alpha_1, \alpha_2, \alpha_3)$ et une variance de modèle σ^2 et que $\{e_t\}$ suit un processus ARMA (0,1) avec un paramètre β^* et une variance de modèle s^2. La seconde partie de cette hypothèse est plausible

dans le cas d'une enquête où le plan d'échantillonnage correspond à l'échantillonnage avec renouvellement à deux degrés d'Eckler, où l'estimation d'enquête pour θ_t est représentée par $\bar{y}_t$, qui est la moyenne calculée pour l'ensemble des personnes qui participent à l'enquête pour la t-ième période.

Nous pouvons exprimer cela sous forme de modèle d'espace d'états en posant

$$F_t = \left[\begin{array}{ccc|cc} \alpha_1 & 1 & 0 & 0 & 0 \\ \alpha_2 & 0 & 1 & 0 & 0 \\ \alpha_3 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right], \quad G_t = \left[\begin{array}{c|c} 1 & 0 \\ 0 & 0 \\ 0 & 0 \\ \hline 0 & 0 \\ 0 & -\beta^* \end{array}\right], \quad V_t = \left[\begin{array}{c|c} \sigma^2 & 0 \\ \hline 0 & s^2 \end{array}\right], \tag{3.8}$$

$U_t = 0$ et $H_t' = (1\,0\,0|1\,0)$.Les trois premiers éléments du vecteur d'états se rapportent au processus $\{\theta_t\}$ tandis que les deux derniers se rapportent au processus $\{e_t\}$.

Notons que les modèles d'espace d'états tiennent compte de l'erreur de mesure, qui est représentée par $\underline{\omega}_t$ dans l'équation (3.6a). Or, à moins qu'il s'agisse d'une enquête sans participation répétée avec erreurs d'échantillonnage indépendantes, on ne peut se servir des termes d'erreur de mesure pour modéliser l'erreur de sondage. C'est pourquoi nous avons intégré l'erreur de mesure (de sondage) dans le vecteur d'états.

À partir du modèle général d'espace d'états, il est possible de déterminer les équations du filtre de Kalman. Si, comme dans Meinhold et Singpurwalla (1983), nous définissons la distribution conditionnelle de z_{t-1} étant donné Y_{t-1} comme $N(\tilde{z}_{t-1|t-1}, \tilde{P}_{t-1|t-1})$, il est possible de construire des relations récursives pour $\tilde{z}_{t|t}$ et $\tilde{P}_{t|t}$ Harvey (1984) montre que ces relations sont l'équivalent du filtre de Kalman.

De façon générale, le filtre de Kalman se compose de deux parties. La première est une prévision une étape à l'avance du vecteur d'états et de sa covariance; la seconde est une mise à jour de la matrice des moyennes et des covariances du vecteur d'espace d'états une fois que les nouvelles observations sont connues.

Utilisant la même notation qu'en (3.6), nous posons $Y_1 = y_1$ et $Y_{t+1}' = (Y_t', y_{t+1}')'$, alors la moyenne et la variance de la prévision une étape à l'avance sont

$$\mathrm{E}(z_1) = \tilde{z}_{1|0} \tag{3.9a}$$

$$\mathrm{Var}(z_1) = \tilde{P}_{1|0} \tag{3.9b}$$

$$\mathrm{E}(z_t|Y_{t-1}) = \tilde{z}_{t|t-1} = F_t\,\tilde{z}_{t-1|t-1} \tag{3.9c}$$

$$\mathrm{Var}(z_t|Y_{t-1}) = \tilde{P}_{t|t-1} = F_t\tilde{P}_{t-1|t-1}F_t' + G_tV_tG_t'. \tag{3.9d}$$

Les nouvelles valeurs de la moyenne et de la variance pour le vecteur d'états au temps t une fois que les observations au temps t sont connues sont déterminées par les équations suivantes:

$$\begin{aligned} E(z_t | Y_t) &= \tilde{z}_{t|t} \\ &= \tilde{z}_{t|t-1} + \tilde{P}_{t|t-1} H_t (H_t' \tilde{P}_{t|t-1} H_t + U_t)^{-1} (y_t - H_t' z_{t|t-1}) \end{aligned} \tag{3.10a}$$

$$\mathrm{Var}(z_t | Y_t) = \tilde{P}_{t|t} = \tilde{P}_{t|t-1} - \tilde{P}_{t|t-1} H_t (H_t' \tilde{P}_{t|t-1} H_t + U_t)^{-1} H_t' \tilde{P}_{t|t-1} \tag{3.10b}$$

Les équations (3.9) et (3.10) sont les équations bien connues du filtre de Kalman. La formulation utilisée ici est essentiellement bayesienne mais on peut obtenir des résultats équivalents au moyen de projections orthogonales; voir Young (1984).

Nous pouvons constater à quel point les équations du filtre de Kalman simplifient les calculs dans les enquêtes par sondage en comparant les équations (3.9) et (3.10) avec celle obtenue par R.G. Jones (1980) (équation 3.5). Notons que pour en arriver à cette équation, Jones a dû inverser une matrice dont les dimensions étaient déterminées par le vecteur des estimations d'enquête.

Le filtre de Kalman peut aussi servir à calculer des estimations lissées que l'on définit par $E(z_t | Y_t)$ pour $T > t$. Pour plus de détails sur cette extrapolation rétrospective, voir Harvey (1984).

Remarques

1. Bien que le filtre de Kalman suppose un modèle de population infinie, il arrive souvent que, grâce au théorème de la limite centrée, les erreurs de sondage suivent une distribution approximativement normale lorsque l'enquête repose sur un grand échantillon. De la même façon, comme les estimateurs lissés de $\{\theta_t\}$ sont identiques à ceux calculés par R.G. Jones (1980) en (3.5a), il s'agit là des estimateurs linéaires à EQM minimum même en l'absence des hypothèses de normalité.

2. Le modèle d'espace d'états peut tenir compte des observations manquantes à une période donnée. Si on ne connaît pas y_t au temps t, les équations de mise à jour analogues à (3.9) deviennent $\tilde{z}_{t|t} = \tilde{z}_{t|t-1}$ et $\tilde{P}_{t|t} = \tilde{P}_{t|t-1}$ comme dans R.H. Jones (1980). Cependant, les estimations lissées correspondant aux observations manquantes dépendront largement du modèle choisi étant donné l'absence d'estimations d'enquête. Il y a donc ici une forte probabilité de mal définir le modèle.

3. La fonction de vraisemblance, dont nous nous servons dans la Section 4 pour calculer les estimations les plus vraisemblables des paramètres inconnus, peut être définie en situation de non-réponse à l'aide de la méthode proposée par R.H. Jones (1980). Toutefois, les données manquantes tendront à accroître les erreurs types des valeurs estimées des paramètres. Dans l'exemple de la Section 5, il y a des observations manquantes à une période donnée.

4. ESTIMATION DES PARAMÈTRES DANS UN MODÈLE D'ESPACE D'ÉTATS

Lorsque les données proviennent du modèle ARMA (équation 3.7) et que les paramètres $\underline{\alpha}$, $\underline{\beta}$, et σ^2 sont inconnus, on peut calculer l'estimation la plus vraisemblable des paramètres inconnus à l'aide de la fonction de vraisemblance issue du modèle d'espace d'états. C'est la méthode qu'ont proposée Harvey et Phillips (1979), R.H. Jones (1980) et d'autres.

On peut aussi utiliser les modèles d'espace d'états habituels lorsque les erreurs de mesure touchant les données d'entrée sont indépendantes. C'est ce que nous faisons dans l'exemple de la Section 5, où nous montrons ce qui arrive aux estimations de paramètres lorsque les erreurs de sondage sont prises en considération.

L'estimation de ces paramètres par la méthode du maximum de vraisemblance lorsqu'il y a corrélation entre les erreurs de sondage est un sujet qui n'a jamais été approfondi. Dans le cas d'un modèle avec des observations stationnaires unidimensionnelles $\{y_t\}$ Scott, Smith et Jones (1977) ont proposé d'utiliser la fonction d'autocovariance estimée des observations, $\{y_t\}$ pour estimer les paramètres du processus ARMA. En l'occurrence, le modèle de données s'écrit $y_t = \theta_t + e_t$. On peut estimer les variances et les covariances des erreurs de sondage, $\{e_t\}$ à l'aide de méthodes fondées sur le plan; voir par exemple Wolter (1985).

Comme il n'existe pas vraiment d'ouvrage où l'on parle de l'estimation efficace des autocovariances des erreurs de sondage dans l'hypothèse d'une série stationnaire, on recourra dans la pratique à des méthodes improvisées. Il serait donc utile de pousser la recherche sur la modélisation des erreurs de sondage. Dans l'exemple de la Section 5, cette lacune ne pose pas de problème car nous avons pu supposer que les erreurs de sondage étaient indépendantes.

En supposant que nous connaissons l'autocovariance de $\{e_t\}$ nous pouvons estimer l'autocovariance de $\{\theta_t\}$ par la formule $\mathrm{Cov}\,(\theta_t, \theta_{t-s}) = \mathrm{Cov}(y_t, y_{t-s}) - \mathrm{Cov}(e_t, e_{t-s})$. Cependant, cette méthode n'est pas parfaitement efficace (Smith, 1978). En outre, elle ne tiendrait pas compte des erreurs de sondage non stationnaires.

Miazaki (1985) a examiné le cas où $\{\theta_t\}$ suit un processus ARMA $(p,0)$. Elle a aussi supposé que $\{e_t\}$ suit un processus ARMA $(0,q)$ qui peut être estimé directement à partir de l'enquête. Ensuite, elle a exprimé les observations $\{y_t\}$ suivant un processus ARMA $(p,p+q)$ qu'elle a estimé par des méthodes du maximum de vraisemblance assujetties à des contraintes.

Dans les modèles d'espace d'états, on peut parfois représenter la non-stationnarité des erreurs de sondage en substituant des matrices non homogènes à $\boldsymbol{V}_t$, thqui est la matrice des variances des «perturbations» aléatoires découlant de l'équation de transition (3.6b). Dans l'équation (3.7) par exemple, on substituerait s^2 à s_t^2 pour tenir compte des erreurs de sondage non homogènes. C'est ce que nous faisons dans l'exemple de la Section 5.

De façon générale, Harvey et Phillips (1979) écrivent comme suit la fonction exacte de vraisemblance pour les modèles d'espace d'états définis par les équations (3.5). Soit

$$\hat{y}_{t|t-1} = \mathrm{E}(y_t | \boldsymbol{Y}_{t-1}) = \boldsymbol{H}_t' \tilde{z}_{t|t-1}$$

et

$$\boldsymbol{R}_t = \mathrm{Var}(y_t | \boldsymbol{Y}_{t-1}) = \boldsymbol{H}_t' \tilde{\boldsymbol{P}}_{t|t-1} \boldsymbol{H}_t + \boldsymbol{U}_t,$$

la fonction de vraisemblance logarithmique pour $\boldsymbol{Y}_T' = (y_1', \ldots, y_T')$ est

$$\log f(\boldsymbol{Y}_t) = (1/2) \textstyle\sum_{t=1}^{T} \log|\boldsymbol{R}_t| - (1/2) \sum_{t=1}^{T} (y_t - \hat{y}_{t|t-1})' \boldsymbol{R}_t^{-1} (y_t - \hat{y}_{t|t-1}). \quad (4.1)$$

Les paramètres inconnus de l'équation (4.1) sont compris dans $\hat{y}_{t|t-1}$ et $\boldsymbol{R}_t$. Suivant l'algorithme utilisé pour maximiser (4.1) par rapport aux paramètres inconnus, on pourrait devoir calculer les dérivées première et seconde de (4.1) par rapport à ces paramètres. Pour cela, on doit normalement calculer les dérivées de $\tilde{z}_{t|t-1}$ et de $\hat{\boldsymbol{P}}_{t|t-1}$. Celles-ci peuvent être calculées à l'aide des formules récursives définies en (3.8) et en (3.9). Par exemple, (3.9c) donne $\partial\hat{z}_{t|t-1} = (\partial \boldsymbol{F}_t)\hat{z}_{t-1|t-1} + \boldsymbol{F}_t(\partial\hat{z}_{t-1|t-1})$. On peut déterminer de la même façon les autres expressions tirées de (3.8) et (3.9).

On peut inclure des paramètres de régression dans l'équation (4.1) en remplaçant y_t par l'écart entre y_t et la droite de régression. Tam (1987) a élargi ce concept en considérant un modèle où le processus stochastique correspondant est déterminé par un modèle d'espace d'états des coefficients de régression en évolution.

Pour maximiser la fonction de vraisemblance (4.1) par rapport aux paramètres inconnus, il faut une méthode itérative. Nous sautons ici les détails de la méthode utilisée dans l'exemple de la Section 5 puisqu'on est encore à élaborer des méthodes efficaces.

Une fois qu'on a estimé les paramètres, on peut calculer des valeurs lissées pour le vecteur d'états, $\tilde{z}_{t|T} = \mathrm{E}(z_t | Y_T)$ pour $T > t$, à l'aide des formules d'extrapolation rétrospective définies par le filtre de Kalman; voir Harvey (1984). Par exemple, si $y_t = \theta + e_t$ comme en (3.1), l'extrapolation rétrospective nous permettra d'écrire $y_t = \tilde{\theta}_{t|T} + \tilde{e}_{t|T}$, de sorte que $\tilde{\theta}_{t|T}$ devient l'estimation lissée de la moyenne au temps t une fois Y_t connu.

Pour calculer l'erreur type de l'estimation lissée, il faut se rappeler, à plus forte raison lorsque la série de données est courte, que les paramètres inconnus ont été estimés à partir des données d'enquête; voir Jones (1979). Hamilton (1986) propose de calculer l'erreur type de l'estimation lissée par des simulations de Monte Carlo. Il produit un ensemble de variables aléatoires normales multidimensionnelles dont la moyenne est déterminée par l'estimation la plus vraisemblable des paramètres et la variance déterminée par l'inverse de la matrice d'information de Fisher estimée. Il estime ensuite $\mathrm{E}(\tilde{P}_{t|T})$ et $\mathrm{Var}(\tilde{z}_{t|T})$, où l'espérance mathématique et la variance sont calculées pour l'ensemble des valeurs de paramètres générées. La somme de ces deux éléments est la matrice des covariances estimées du vecteur d'états estimé. Cette méthode suppose un grand échantillon, de sorte qu'on peut dire que la distribution d'échantillonnage des estimations des paramètres est approximativement normale.

Dans l'exemple de la Section 5, nous calculons une valeur approchée pour l'écart type des erreurs d'échantillonnage des estimations lissées en ne tenant pas compte de la variation causée par l'estimation de certains paramètres du modèle. Nous comparons ensuite ces valeurs aux erreurs types de la distribution d'échantillonnage établie à l'aide de données simulées.

5. ANALYSE DE DONNÉES

Dans cette section, nous allons voir l'incidence des erreurs de sondage sur les valeurs estimées des paramètres d'un modèle autorégressif du premier degré comprenant des termes de régression. Dans l'exemple que nous présentons, nous supposons que les erreurs de sondage sont indépendantes d'une enquête à l'autre. Le cadre d'analyse que nous venons d'exposer permet de traiter les cas plus complexes où il y a corrélation des erreurs de sondage et où la caractéristique de population est expliquée par des modèles ARMA de degré supérieur. Nous avons choisi cet exemple pour montrer que le fait de tenir compte des erreurs de sondage peut avoir des conséquences appréciables même pour un modèle relativement simple.

Pour notre exemple, nous nous sommes servi des données fournies par des résidents de la Saskatchewan à l'enquête sur les voyages des Canadiens (EVC). L'EVC est réalisée par Statistique Canada dans le but de recueillir des statistiques descriptives sur les habitudes de voyage des résidents canadiens et les caractéristiques des voyageurs canadiens. Elle est un supplément de l'enquête sur la population active (EPA). L'EPA est une enquête mensuelle à échantillon constant où il y a six groupes de renouvellement. Or, l'EVC revient au plus quatre fois par année et utilise au moins un groupe de renouvellement mais peut en utiliser jusqu'à trois. Les groupes utilisés pour les trimestres où il y a une EVC sont choisis de manière qu'il n'y ait pas de chevauchement d'une fois à l'autre.

Les erreurs de sondage sont supposées indépendantes. Cette hypothèse n'est pas parfaitement exacte. En effet, l'EPA est une enquête à plusieurs degrés et les unités primaires d'échantillonnage (UPÉ) ne se succèdent pas aussi rapidement que les groupes de renouvellement dans l'échantillon. Les mêmes UPÉ servent à plusieurs reprises. Ainsi, bien que l'échantillon de

l'EVC soit choisi de manière qu'il n'y ait pas de chevauchement des panels d'une fois à l'autre, l'hypothèse de l'indépendance n'est approximativement juste que lorsqu'il y a une faible corrélation des erreurs d'échantillonnage d'un trimeste à l'autre à l'intérieur de la même UPÉ. Cette hypothèse n'a pas été vérifiée.

Les coefficients de variation (en pourcentage) ont été calculés au moyen de la fonction suivante:

$$\mathrm{CV} = \alpha y^{-\beta}/\sqrt{\text{nombre de groupes de renouvellement}},$$

où y est l'estimation d'enquête en milliers. C'est la fonction que nous SCR recommandons d'utiliser pour les données de l'EVC concernant les résidents de la Saskatchewan; voir Statistique Canada (1985). Dans cette étude, nous nous servons d'un modèle de régression log-linéaire appliqué à des données de 1979 pour estimer la valeur des paramètres α et β les valeurs estimées sont respectivement 91.7528 et 0.353253. Pour les besoins de notre exemple, ces coefficients de variation ont été arrondis au dixième pourcent près.

Nous avons supposé le modèle:

$$y_t = \theta_t + e_t, \tag{5.1}$$

où les e_t's sont des erreurs de sondage indépendantes, $e_t \sim N(0,s_t^2)$ et

$$\theta_t = \gamma_0 + \gamma_1 t + \gamma_2 Q_{1t} + \gamma_3 Q_{2t} + \gamma_4 Q_{3t} + \epsilon_t, \tag{5.2}$$

où $\{\epsilon_t\}$ suit un processus ARMA (1,0) avec des paramètres (α,σ^2). Les termes de régression de l'équation (5.2) sont, dans l'ordre, l'ordonnée à l'origine, un terme indiquant le numéro du trimestre, où t varie linéairement en fonction du temps, son intervalle de variation étant $[-15.5, 15.5]$, et enfin des termes saisonniers pour les trois premiers trimestres de chaque année, où

$$\begin{aligned} Q_{it} &= 1 \text{ si la } t\text{-ième observation est dans le } i\text{-ième trimestre;} \\ &= -1 \text{ si la } t\text{-ième observation est dans le quatrième trimestre;} \\ &= 0 \text{ dans les autres cas;} \end{aligned}$$

pour $i = 1, 2, 3$.

On pourrait trouver de meilleurs modèles pour ces données mais compte tenu de la faible dimension de la série, les tests d'hypothèses pour d'autres modèles ne seraient pas très puissants.

Pour calculer l'estimation la plus vraisemblable des paramètres inconnus de ce modèle, il faut tenir compte des hypothèses qui ont été posées à propos des erreurs de sondage. La plupart des utilisateurs de données officielles ne se préoccupent pas de l'erreur de sondage et supposent implicitement que les données d'entrée ne sont pas entachées d'erreur. Cela n'a pas de conséquence grave lorsque la variance de l'erreur de sondage est faible par rapport à la variance de l'erreur de modèle.

Les tableaux 1 et 2 donnent les estimations d'enquête et les coefficients de variation correspondants tandis que les tableaux 3 et 4 contiennent les résultats de l'estimation par la méthode du maximum de vraisemblance. Dans ce dernier cas, deux estimations sont indiquées pour chaque modèle: une pour le cas où on tient compte de l'erreur d'échantillonnage et une pour le cas où on n'en tient pas compte. On suppose que le modèle (5.2) s'applique dans les deux cas.

Tableau 1

Voyages-personnes d'une nuit ou plus sur le territoire de la Saskatchewan
Résidents de la Saskatchewan[1]

Année	Trimestre	Nbre de groupes de renouvellement	Estimation brute (milliers)	Estimation lissée (milliers)	C.V. des estimations (brutes) (%)	C.V. des estimations (lissées) (%)	Précision simulée (%)	Biais simulé (%)
1979	Hiver	1	598	611	9.6	5.9	6.9	0.1
	Printemps	1	808	813	8.6	4.8	4.9	0.4
	Été	3	1033	1103	4.6	3.0	3.1	0.0
	Automne	3	678	683	5.3	4.3	4.5	1.2
1980	Hiver	1	578	608	9.7	5.5	5.8	0.1
	Printemps	3	837	837	4.9	3.7	3.6	0.0
	Été	1	1451	1169	7.0	3.3	3.5	0.3
	Automne	1	744	724	8.9	5.1	5.9	0.8
1981	Hiver	3	631	632	5.4	4.3	5.0	-0.1
	Été	3	1262	1172	4.2	2.9	3.3	0.1
1982	Hiver	1	565	613	9.8	5.5	6.4	-0.4
	Printemps	1	901	838	8.3	4.5	5.1	0.8
	Été	3	1167	1147	4.4	2.9	3.1	0.1
	Automne	1	721	706	9.0	5.1	5.6	0.2
1984	Hiver	1	585	598	9.6	5.8	6.7	-1.2
	Printemps	1	788	804	8.7	4.6	5.2	-0.4
	Été	3	1068	1107	4.5	2.9	3.6	-0.5
	Automne	1	711	686	9.0	5.3	6.7	0.7
1986	Hiver	1	793	630	8.7	6.2	7.1	-1.3
	Printemps	3	798	808	5.0	3.9	3.9	-0.4
	Été	3	1053	1096	4.5	3.0	3.3	-0.3
	Automne	3	650	663	5.4	4.4	4.2	0.2

[1] Il n'y a pas eu d'enquête sur les voyages des Canadiens au printemps et à l'automne de 1981 ni en 1983 et 1985.

Les valeurs simulées figurant dans les deux dernières colonnes sont fondées sur un échantillon de taille n = 100.

Les valeurs estimées des paramètres de régression sont sensiblement les mêmes suivant l'une ou l'autre des hypothèses. En revanche, les valeurs estimées des paramètres du modèle autorégressif diffèrent considérablement. La variance de modèle, notamment, augmente de façon substantielle lorsqu'on ne tient pas compte de l'erreur d'échantillonnage car la variation due à l'erreur de sondage n'est pas expliquée par le modèle. La situation est toute autre pour les coefficients de régression parce que les estimateurs de ces coefficients demeurent non biaisés bien qu'ils soient quelque peu inefficaces.

Une fois les paramètres du modèle estimés, on peut se servir du modèle proposé pour redresser le nombre estimé de voyages-personnes d'une nuit ou plus. Les résultats analysés ci-dessous indiquent jusqu'à quel point cette opération permet de réduire le coefficient de variation des estimations lorsque les hypothèses du modèle sont exactes. Ce genre d'opération s'apparente aux méthodes d'estimation fondées sur un modèle pour les petites régions.

Tableau 2

Voyages-personnes d'une nuit ou plus sur le territoire du Manitoba - Résidents de la Saskatchewan[1]

Année	Trimestre	Nbre de groupes de renouvellement	Estimation brute (miliers)	Estimation lissée (milliers)	C.V. des estimations (%)	C.V. des estimations (%)	Précision simulée (%)	Biais simulé (%)
1979	Hiver	1	27	34	28.6	13.4	14.1	0.5
	Printemps	1	33	48	26.7	11.0	10.2	0.9
	Été	3	78	80	11.4	6.6	7.1	1.3
	Automne	3	55	48	12.9	10.1	10.8	0.6
1980	Hiver	1	24	30	29.7	13.6	14.5	0.5
	Printemps	3	63	50	12.3	9.5	9.4	0.7
	Été	1	86	80	19.0	6.6	6.3	0.8
	Automne	1	75	46	19.9	11.0	12.2	0.5
1981	Hiver	3	42	34	14.2	11.3	13.2	1.0
	Été	3	79	82	11.3	5.9	5.7	0.1
1982	Hiver	1	33	34	26.5	12.5	13.2	-2.8
	Printemps	1	46	44	23.7	10.7	10.0	1.6
	Été	3	78	82	11.4	5.7	5.4	0.1
	Automne	1	30	42	27.6	10.9	11.4	0.3
1984	Hiver	1	36	34	25.7	13.8	16.8	-1.3
	Printemps	1	48	43	23.4	11.4	11.5	0.1
	Été	3	82	82	11.1	6.1	7.3	-0.2
	Automne	1	30	40	27.7	11.5	11.4	0.6
1986	Hiver	1	33	33	26.7	16.3	19.9	-0.8
	Printemps	3	38	41	14.6	10.9	11.7	-0.1
	Été	3	90	81	10.8	7.1	8.8	-0.3
	Automne	3	42	40	14.1	11.2	10.5	1.7

[1] Il n'y a pas eu d'enquête sur les voyages des Canadiens au printemps et à l'automne de 1981 ni en 1983 et 1985. Les valeurs simulées figurant dans les deux dernières colonnes sont fontées sur un échantillon de taille = 100.

Tableau 3

Valeurs estimées des paramètres pour les voyages-personnes à l'intérieur de la Saskatchewan[1]

Paramètre	Compte non tenu de l'erreur d'échantillonnage	Compte tenu de l'erreur d'échantillonnage				
	Valeur estimée	Valeur estimée	Erreur type	Précision simulée	Biais simulé	Valeur t du biais
RÉGRESSION						
Ordonnée à l'origine (γ_0)	831.4	815.0	15.6	14.4	1.8	1.29
Terme Linéaire (γ_1)	-0.84	-0.86	1.52	1.51	-0.10	-0.65
1er trimestre (γ_2)	-209.6	-203.8	21.8	24.6	-3.5	-1.41
2e trimestre (γ_3)	-4.0	7.1	22.9	23.8	0.4	0.17
3e trimestre (γ_4)	340.1	316.0	21.2	23.4	-0.4	-0.18
ARMA						
Autorégressif (α)	0.14	0.47	0.66	0.68	-0.39	-6.77
Variance de modèle (σ^2)	7930.5	879.3	1205.6	770.0	-488.2	-8.16

[1] Les valeurs simulées et les valeurs t sont fondées sur un échantillon de taille n = 100.

Tableau 4

Valeur estimées des paramètres pour les voyages-personnes de la Saskatchewan au Manitoba[1]

Paramètre	Compte non tenu de l'erreur d'échantillonnage	Compte tenu de l'erreur d'échantillonnage				
	Valeur estimée	Valeur estimée	Erreur type	Précision simulée	Biais simulé	Valeur t du biais
RÉGRESSION						
Ordonnée à l'origine (γ_0)	51.2	50.5	1.9	2.0	0.4	1.57
Terme linéaire (γ_1)	-0.17	-0.13	0.18	0.17	-0.04	-2.01
1er trimestre (γ_2)	-20.1	-17.2	3.4	3.5	-0.6	-1.52
2e trimestre (γ_3)	-5.9	-6.1	3.6	3.7	-0.1	-0.32
3e trimestre (γ_4)	30.7	30.8	3.7	3.7	0.0	-0.07
ARMA						
Autorégressif (α)	0.14	-0.75	0.66	0.71	0.49	7.90
Variance de modèle (σ^2)	100.0	5.7	18.7	9.5	-0.3	-0.29

[1] Les valeurs simulées et les valeurs t sont fondées sur un échantillon de taille n = 100.

Les estimations lissées et les coefficients de variation correspondants figurent dans les tableaux 1 et 2. On calcule ces coefficients de variation en tenant compte de l'erreur d'échantillonnage des coefficients de régression, $\gamma_0, \ldots, \gamma_4$. On peut procéder ainsi puisque, étant donné α et σ^2, les estimations lissées sont une fonction linéaire des estimations originales, de sorte que les variances peuvent être calculées à l'aide de cette fonction linéaire et de la variance des résidus de régression du modèle hypothétique. Or, on n'a pas tenu compte jusqu'ici des erreurs d'échantillonnage pour les valeurs estimées de α et σ^2. Nous voyons ci-dessous quels effets cela peut avoir.

Les estimations lissées pour le nombre de voyages-personnes en Saskatchewan sont généralement proches des estimations d'enquête, sauf peut-être en ce qui concerne l'été de 1980 et l'hiver de 1986. Les estimations lissées pour le nombre de voyages-personnes au Manitoba sont également proches des estimations d'enquête, sauf peut-être en ce qui concerne l'automne de 1980. Les trois exceptions peuvent être des valeurs aberrantes ou peuvent être attribuables à un événement spécial qui a contribué à accroître le tourisme durant ces trois trimestres. En règle générale, il y a deux façon d'inclure ces phénomènes dans le modèle: (i) soit accroître la variance de modèle dans le modèle d'espace d'états pour ces périodes ou prévoir des variables auxiliaires appropriées pour les événements spéciaux; (ii) soit accroître la variance d'échantillonnage pour les valeurs aberrantes. Seule une connaissance plus approfondie des circonstances nous permettrait de déterminer si ce sont là les ajustements qui conviennent. L'analyse que nous faisons ici peut nous permettre de faire ressortir les cas exceptionnels possibles.

Comme notre analyse n'a pas tenu compte jusqu'ici de l'effet de l'erreur d'échantillonnage liée à l'estimation de α et de σ^2, nous avons effectué une simulation visant à évaluer cet effet. Jones (1979), Hamilton (1986) et Tam (1987) estiment qu'il ne faut pas négliger cette erreur d'échantillonnage, surtout lorsque la série chronologique comporte peu d'observations.

Pour ce qui a trait à la simulation, nous avons produit des séries de données aléatoires suivant le modèle posé en (5.1) et (5.2). Nous nous sommes servi des estimations les plus vraisemblables comme valeurs des paramètres et du même schéma de non-réponse que pour la

série de données originales. Nous avons produit dans ces conditions cent séries de données pour chaque modèle. Les tableaux 1 et 2 donnent le biais (en pourcentage) des valeurs lissées ainsi que la précision (en pourcentage) relative à la différence entre les valeurs lissées et les valeurs réelles fondées sur ces simulations.

Afin de déterminer si 100 simulations étaient suffisantes pour estimer la précision, nous avons calculé une estimation du coefficient de variation de l'estimateur de la précision. Les simulations nous ont permis d'établir une estimation non biaisée de la variance de l'estimateur de l'erreur quadratique moyenne. Nous avons ensuite estimé la variance de l'estimateur de la précision par la méthode de linéarisation de Taylor. Les valeurs estimées des coefficients de variation allaient de 6 à 11 % pour les voyages faits sur le territoire de la Saskatchewan et de 5 à 9 % pour les voyages faits au Manitoba. Ainsi, ces estimations de la précision nous donnent une idée assez juste des conséquences de l'omission de l'erreur d'échantillonnage des paramètres autorégressifs.

Selon les tableaux 1 et 2, les biais obtenus par simulation sont tous faibles; de fait pour les deux séries de 22 observations chacune, on n'a relevé que quatre biais significatifs à un seuil de 5 % en utilisant un test t-ordinaire.

Nous remarquons aussi que les estimations de la précision obtenues par simulation (en pourcentage) sont en règle générale plus élevées que les valeurs figurant dans la colonne «C.V. des estimations lissées». Cela est normal puisque les simulations tiennent compte des erreurs d'échantillonnage qui découlent de l'estimation de α et de θ^2. Néanmoins, les C.V. des estimations lissées donnent une approximation acceptable des valeurs simulées, de sorte que nous pouvons dire que le fait de ne pas tenir compte de l'erreur d'échantillonnage de α et de θ^2 ne modifie pas réellement les coefficients de variation.

Les tableaux 3 et 4 contiennent des résultats de la simulation pour les paramètres estimés. En ce qui a trait aux coefficients de régression, un seul des biais simulés est significatif à un seuil de 5 %. Les erreurs types sont toutes conformes aux résultats de la simulation.

En revanche, les choses ne sont pas aussi simples en ce qui concerne les valeurs estimées de α et de σ^2. Les biais des valeurs estimées de α sont fortement significatifs et comme nous pouvons le voir dans les tableaux 3 et 4, le biais d'une des valeurs estimées de σ^2 est aussi fortement significatif. Les erreurs types simulées ne sont pas très proches de l'approximation asymptotique de l'erreur type obtenue en inversant la matrice d'information de Fisher. Il semble que l'échantillon de notre exemple ne soit pas assez grand pour permettre une approximation asymptotique très précise. C'est un problème que l'on retrouve souvent lorsqu'on analyse de courtes séries chronologiques.

6. CONCLUSION

Lorsque la variance de l'erreur de sondage est faible par rapport à la variance de l'erreur de modèle, il y a peu de différence entre l'estimation lissée et l'estimation linéaire non biaisée à variance minimum et l'erreur type correspondante ne subit pas de réelle diminution même lorsque le modèle hypothétique est exact. Cependant, en ce qui a trait à l'estimation pour petits domaines, où l'erreur d'échantillonnage est élevée, l'erreur type de l'estimation lissée peut être beaucoup moindre que celle de l'estimation d'enquête. Par exemple, le lissage des estimations a donné des résultats plus notables dans le cas des voyages faits au Manitoba puisque l'erreur d'échantillonnage était plus élevée dans ce cas-là.

Une des conséquences de la modélisation pour les enquêtes répétées est qu'une définition erronée du modèle peut se traduire par un estimateur à EQM minimum fortement biaisé. Il est donc essentiel de choisir un modèle qui soit conforme aux données et qui reflète une connaissance approfondie du phénomène en question. Comme notre exemple a trait à une petite série de données, un grand nombre de modèles statistiques seraient conformes aux données.

La simulation que nous avons faite donne à penser que même pour de petites séries de données, l'approximation asymptotique de la variance des estimations lissées est tout à fait acceptable. Cependant, lorsqu'il s'agit d'applications plus classiques de l'analyse chronologique, l'approximation asymptotique de l'erreur d'échantillonnage des valeurs estimées des paramètres peut laisser à désirer.

REMERCIEMENTS

Les auteurs tiennent à remercier un rédacteur associé de la revue et les arbitres pour leurs précieux commentaires sur des versions antérieures de cet article. Ils tiennent à exprimer plus particulièrement leur reconnaissance à l'arbitre qui, par ses commentaires profonds et éclairants, a contribué grandement à l'amélioration de cet article. Ils remercient également Pierre Hubert, chef de la Section des voyages, du tourisme et des loisirs, Division de l'éducation, de la culture et du tourisme, pour avoir mis à leur disposition les données de l'enquête sur les voyages des Canadiens. Des éléments de cet article sont tirés du mémoire de maîtrise présenté par le second auteur à l'Université de Guelph.

BIBLIOGRAPHIE

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

BLIGHT, B.J.N., et SCOTT, A.J. (1973). A stochastic model for repeated surveys. *Journal of the Royal Statistical Society,* Sér. B, 35, 61-68.

ECKLER, A.R. (1955). Rotation sampling. *Annals of Mathematical Statistics.* 26, 664-685.

GURNEY, M., et DALY, J.F. (1965). A multivariate approach to estimation in periodic sample surveys. *Proceedings of the Social Statistics Section, American Statistical Association*, 247-257.

HAMILTON, J.D. (1986). A standard error for the estimated state vector of a state-space model. *Journal of Econometrics*, 33, 387-397.

HARVEY, A.C. (1984). A unified view of statistical forecasting procedures. *Journal of Forecasting*, 3, 245-275.

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

JESSEN, R.J. (1942). Statistical investigation of a farm survey for obtaining farm facts. *Iowa Agricultural Experimental Station Research Bulletin*, 304, 54-59.

JONES, R.G. (1979). The efficiency of time series estimators for repeated surveys. *Australian Journal of Statistics*, 21, 45-56.

JONES, R.G. (1980). Best linear unbiased estimators for repeated surveys. *Journal of the Royal Statistical Society*, Sér. B, 42, 221-226.

JONES, R.H. (1980). Maximum likelihood fitting of ARMA models to time series with missing observations. *Technometrics*, 22, 389-395.

KALMAN, R.E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82, 35-45.

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

MIAZAKI, E.S. (1985). Estimation for time series subject to the error of rotation sampling. Thèse de doctorat Iowa State University, Ames, Iowa.

PATTERSON, H.D. (1950). Sampling on successive occasions with partial replacement of units. *Journal of the Royal Statistical Society*, Sér. B, 12, 241-255.

PEARLMAN, J.G. (1980). An algorithm for the exact likelihood of a high-order autoregressive-moving average process. *Biometrika*, 67, 232-233.

RAO, J.N.K., et GRAHAM, J.E. (1964). Rotation designs for sampling on repeated occasions. *Journal of the American Statistical Association*, 59, 492-509.

SCOTT, A.J., et SMITH, T.M.F. (1974). Analysis of repeated surveys using time series methods. *Journal of the American Statistical Association*, 69, 674-678.

SCOTT, A.J., SMITH, T.M.F., et JONES, R.G. (1977). The application of time series methods to the analysis of repeated surveys. *International Statistics Review*, 45, 13-28.

SINGH, D. (1968). Estimates in successive sampling using multi-stage design. *Journal of the American Statistical Association*, 63, 99-112.

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

STATISTIQUE CANADA (1985). Canadian Travel Survey: Estimation and variance estimation procedures. Rapport technique, Statistique Canada.

TAM, S.M. (1987). Analysis of repeated surveys using a dynamic linear model. *International Statistical Review*, 55, 63-73.

TIKKIWAL, B.D. (1979). Successive sampling — a review. *Bulletin of the International Statistics Institute*, 48, 367-384.

WOLTER, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.

YOUNG, P. (1984). *Recursive Estimation and Time Series Analysis: An Introduction*. New York: Springer-Verlag.