Techniques d'enquête, décembre 1990
Vol. 16, n° 2, pp. 205–227
Statistique Canada

Estimation dans les enquêtes à passages répétés au moyen de séries chronologiques

WILLIAM R. BELL et STEVEN C. HILLMER[1]

RÉSUMÉ

Scott et Smith (1974) et Scott, Smith et Jones (1977) ont proposé d'utiliser les résultats de l'extraction de signal afin d'améliorer les estimations tirées des enquêtes à passages répétés; c'est ce qu'on appelle l'approche chronologique à l'estimation dans les enquêtes à passages répétés. Dans cet article, nous examinons les fondements théoriques de cette approche en précisant qu'elle repose sur la reconnaissance de deux sources de variation – variation chronologique et variation d'échantillonnage – et qu'elle peut aussi servir à résoudre d'autres problèmes qui mettent en évidence ces deux sources de variation. Avec cette approche, nous obtenons des résultats théoriques concernant la convergence selon le plan des estimateurs de séries chronologiques et l'absence de corrélation entre la série du signal et la série des erreurs d'échantillonnage. Nous constatons que, dans la perspective d'un plan de sondage, l'approche chronologique engendre, par rapport aux estimateurs classiques, une variance et une erreur quadratique moyenne moins élevées mais un biais plus grand. Nous voyons brièvement comment appliquer cette approche par la modélisation puis, à titre d'exemple, nous prenons la série des ventes au détail des établissements de restauration (ou restaurants) et des débits de boissons, tirée de la Retail Trade Survey du U.S. Bureau of the Census.

MOTS CLÉS: Enquêtes à passages répétés; séries chronologiques; extraction du signal; U.S. Retail Trade Survey.

1. INTRODUCTION

Scott et Smith (1974) et Scott, Smith et Jones (1977) (ci-après désigné SSJ) ont proposé de recourir à l'analyse de séries chronologiques et plus particulièrement à l'extraction de signal afin d'améliorer les estimations tirées des enquêtes à passages répétés. Lorsqu'on connaît la structure de covariances des estimations d'enquête (Y_t) et des erreurs d'échantillonnage correspondantes (e_t) pour une série de périodes, les résultats de l'extraction de signal peuvent servir à déterminer les fonctions linéaires à erreur quadratique moyenne minimum des observations Y_t, ces fonctions étant les estimateurs des caractéristiques de population à estimer pour une série stochastique θ_t. Concrètement, on se sert des résultats de l'extraction de signal en estimant un modèle chronologique pour la série d'observations Y_t et en estimant la structure de covariances de e_t à long terme à l'aide des données du plan de sondage.

Dans la section 2, nous exposons sommairement l'approche chronologique et ses caractéristiques fondamentales. La section 3 renferme des questions d'ordre théorique tandis que la section 4 présente des considérations relatives à l'application de l'approche. Enfin dans la section 5, nous donnons un exemple d'application en nous servant de deux séries chronologiques tirées de la Retail Trade Survey du U.S. Bureau of the Census.

2. L'APPROCHE CHRONOLOGIQUE: PRINCIPES FONDAMENTAUX ET ANALYSE GÉNÉRALE

Les estimateurs de séries chronologiques se distinguent principalement des estimateurs classiques par le fait qu'ils reconnaissent l'existence de *deux sources de variation* au lieu d'une.

[1] William R. Bell est principal de recherche à la Statistical Research Division, Bureau of the Census, Washington, D.C. 20233, É.-U. et Steven C. Hillmer est professeur à la School of Business, University of Kansas, Lawrence, Kansas 66045, É.-U.

Dans le cas des estimateurs classiques, on parle uniquement de variation due à l'échantillonnage - du fait que l'on n'a pas observé toutes les unités de la population. Dans le cas de l'analyse chronologique, on parle de variation due au fait que l'on ne peut définir parfaitement (le plus souvent de façon linéaire) une série chronologique à l'aide de données antérieures. Considérons la formule de décomposition:

$$Y_t = \theta_t + e_t, \tag{2.1}$$

où Y_t est une estimation d'enquête au temps t, θ_t est la valeur de la caractéristique de la population au temps t et e_t est l'erreur d'échantillonnage. La variabilité d'échantillonnage de e_t est le point de mire des méthodes d'estimation classiques, qui considèrent θ_t comme fixe. En revanche, l'approche chronologique prévoit que Y_t, θ_t, et e_t peuvent toutes les trois varier dans le temps pourvu qu'elles soient aléatoires et qu'on ne puisse en prévoir parfaitement la valeur à l'aide de données antérieures. L'analyse chronologique ordinaire porte directement sur Y_t et ne tient pas compte de l'erreur d'échantillonnage comme telle, si ce n'est comme un élément de l'agrégat Y_t. De fait, les analystes de séries chronologiques font comme si la variation d'échantillonnage n'existait pas et que les valeurs réelles étaient observées. Bref, ce qu'il faut surtout retenir au sujet de l'application des méthodes d'analyse chronologique à l'estimation fondée sur les enquêtes, c'est qu'il existe deux sources de variation stochastique qui sont définies, modélisées et estimées différemment l'une de l'autre.

2.1 Résultats de l'extraction du signal

Supposons qu'il existe des estimations d'enquête Y_t pour une série de périodes identifiées $t = 1, \ldots, T$. Soit $\underset{\sim}{Y} = (Y_1, \ldots, Y_T)'$ et définissons de la même façon $\underset{\sim}{\theta}$ et $\underset{\sim}{e}$, de sorte que $\underset{\sim}{Y} = \underset{\sim}{\theta} + \underset{\sim}{e}$. En supposant que les estimations Y_t sont non biaisées et que θ_t et e_t sont non corrélées (voir section 3.2)

$$E(\underset{\sim}{Y}) = E(\underset{\sim}{\theta}) \equiv \underset{\sim}{\mu} \equiv (\mu_1, \ldots, \mu_T)'$$

$$\Sigma_Y = \Sigma_\theta + \Sigma_e, \tag{2.2}$$

où E désigne l'espérance de la distribution d'échantillonnage et de la distribution du modèle chronologique, et Σ_Y est la matrice des covariances de $\underset{\sim}{Y}$, *etc.* Dans les expressions ci-dessus, $\underset{\sim}{\mu}$ et Σ_θ ont rapport à la structure chronologique de θ_t, qui n'est pas exposée à la variation d'échantillonnage. Lorsqu'on n'a pas besoin d'appliquer des différences aux séries Y_t, θ_t, ou e_t on sait très bien qu'il est possible d'exprimer l'estimateur linéaire à erreur quadratique moyenne minimum de $\underset{\sim}{\theta}$ de la façon suivante (en se servant de l'équation (2.2)) puisque $\mathrm{Cov}(\underset{\sim}{\theta}, \underset{\sim}{Y}) = \Sigma_\theta$:

$$\hat{\underset{\sim}{\theta}} = \underset{\sim}{\mu} + \Sigma_\theta \Sigma_Y^{-1} (\underset{\sim}{Y} - \underset{\sim}{\mu}) \tag{2.3}$$

$$= \underset{\sim}{\mu} + \left(I - \Sigma_e \Sigma_Y^{-1}\right)(\underset{\sim}{Y} - \underset{\sim}{\mu}) \tag{2.4}$$

$$= \underset{\sim}{\mu} + \left(I + \Sigma_e \Sigma_\theta^{-1}\right)^{-1} (\underset{\sim}{Y} - \underset{\sim}{\mu}). \tag{2.5}$$

On sait aussi que la variance de l'erreur de cet estimateur est

$$\mathrm{Var}(\hat{\underset{\sim}{\theta}} - \underset{\sim}{\theta}) = \Sigma_\theta - \Sigma_\theta \Sigma_Y^{-1} \Sigma_\theta = \Sigma_e - \Sigma_e \Sigma_Y^{-1} \Sigma_e. \tag{2.6}$$

Si on pose par hypothèse la normalité de $(\underset{\sim}{\theta},\underset{\sim}{Y})$ les équations (2.3) à (2.5) permettent d'obtenir $E(\underset{\sim}{\theta} \mid \underset{\sim}{Y})$, l'espérance conditionnelle de $\underset{\sim}{\theta}$ étant donné $\underset{\sim}{Y}$, et l'équation (2.6), la variance conditionnelle, $\text{Var}(\underset{\sim}{\theta} \mid \underset{\sim}{Y})$.

S'il faut appliquer des différences à Y_t, les résultats précédents ne tiennent plus. Supposons qu'il n'est pas nécessaire d'appliquer des différences à e_t mais qu'il faut appliquer une différence d'ordre un à θ_t et à Y_t (au moyen du facteur $1 - B$ où $BY_t = Y_{t-1}$). Soit la nouvelle série de données $W_t = (1 - B)Y_t = (1 - B)\theta_t + (1 - B)e_t$ pour $t = 2, \ldots, T$. Définissons $\Delta = [\Delta_{ij}]$ comme la matrice de différences $(T - 1) \times T$ où $\Delta_{ii} = -1, \Delta_{i,i+1} = 1$, et tous les autres éléments sont nuls, et écrivons $\Delta\underset{\sim}{Y} \equiv \underset{\sim}{W} = \Delta\underset{\sim}{S} + \Delta\underset{\sim}{e}$. Nous utilisons ensuite

$$\hat{\underset{\sim}{\theta}} = \underset{\sim}{Y} - \hat{\underset{\sim}{e}} = \underset{\sim}{Y} - \Sigma_e \Delta' \Sigma_W^{-1} \Delta(\underset{\sim}{Y} - \underset{\sim}{\mu}), \tag{2.7}$$

$$\text{Var}(\hat{\underset{\sim}{\theta}} - \underset{\sim}{\theta}) = \Sigma_e - \Sigma_e \Delta' \Sigma_W^{-1} \Delta \Sigma_e. \tag{2.8}$$

Les équations (2.7) et (2.8) sont aussi utiles lorsqu'il faut appliquer un opérateur plus général (par ex.: différences saisonnières) à θ_t et à Y_t (la matrice des différences Δ étant définie en conséquence), pourvu qu'il ne soit pas nécessaire d'appliquer des différences à e_t. Ces équations sont analogues à (2.4) et à (2.6) sauf que $\Delta' \Sigma_W^{-1} \Delta$ est substitué à Σ_Y^{-1}. Nous trouvons ces équations dans Bell et Hillmer (1990), qui les analysent sur le plan de l'optimalité. Jones (1980) avait exposé essentiellement les mêmes résultats mais sans les expliquer vraiment.

Scott et Smith (1974) et SSJ ont utilisé la méthode classique d'extraction de signal (c.-à-d. les équations (2.3) à (2.6)) mais en se fondant sur des fonctions génératrices de covariances plutôt que des matrices de covariances. Bell (1984) utilise la même méthode pour des modèles prévoyant l'application de différences. Une autre approche (Binder et Dick 1989; Bell et Hillmer 1989) consiste à exprimer des modèles chronologiques de θ_t et de e_t sous forme de modèles d'espace d'états puis à utiliser le filtre de Kalman, façon efficace, semble-t-il, d'obtenir les matrices définies ci-dessus. En outre, Tam (1987) se sert du filtre de Kalman dans une approche fondée explicitement sur un modèle pour l'analyse des données d'enquêtes à passages répétés. Dans le reste de cet article, il sera le plus souvent question des équations (2.3) à (2.6); cependant, nos commentaires valent tout aussi bien pour les équations (2.7) et (2.8).

Il arrive souvent que, pour des séries chronologiques Y_t et θ_t toujours positives, l'on veuille calculer le logarithme de Y_t pour faire de θ_t et de e_t des séries stationnaires. Dans un tel cas, on reformule (2.1) comme suit:

$$Y_t = \theta_t(1 + \tilde{u}_t) = \theta_t u_t, \tag{2.9}$$

où $\tilde{u}_t = e_t/\theta_t$ et $u_t = 1 + \tilde{u}_t$. En utilisant les logarithmes, on obtient

$$\log(Y_t) = \log(\theta_t) + \log(1 + \tilde{u}_t) = \log(\theta_t) + \log(u_t). \tag{2.10}$$

En posant que $\underset{\sim}{\mu}$ et Σ_θ représentent désormais $\log(\underset{\sim}{\theta}) \equiv (\log(\theta_1), \ldots, \log(\theta_T))'$, et que $\Sigma_Y = \Sigma_\theta + \Sigma_u$ représente $\log(\underset{\sim}{Y})$, la formule de l'estimateur, équivalente à (2.4), est

$$\log(\hat{\underset{\sim}{\theta}}) = \underset{\sim}{\mu} + \left[I - \Sigma_u \Sigma_Y^{-1}\right](\log(\underset{\sim}{Y}) - \underset{\sim}{\mu}). \tag{2.11}$$

Les formules équivalentes aux équations (2.6) à (2.8) sont évidentes. Pour estimer $\hat{\theta}_t$, on utilise $\exp[\log(\hat{\theta}_t)]$; par ailleurs, on pourrait utiliser $\exp[\log(\hat{\theta}_t) + \text{Var}(\log(\hat{\theta}_t) - \log(\theta_t))/2]$ pour obtenir une estimation "moins biaisée" de θ_t, qui aurait une erreur quadratique moyenne minimum (voir Granger et Newbold 1976).

Notons que pour résoudre les équations (2.3) à (2.6), il faut connaître $\underset{\sim}{\mu}$ ainsi que deux des trois matrices de covariances Σ_Y, Σ_θ et Σ_e (la troisième pouvant être déterminée à l'aide de l'équation (2.2)). En pratique, on ne connaîtra pas précisément ces matrices et il faudra donc les estimer. Par conséquent, il n'est pas possible de connaître avec exactitude l'estimateur linéaire à erreur quadratique moyenne minimum $\hat{\underset{\sim}{\theta}}$ et les équations (2.6) ou (2.8) produisent une valeur sous-estimée de l'erreur quadratique moyenne (EQM) puisqu'elles ne tiennent pas compte des erreurs de modélisation. (Voir Binder et Dick (1989) et Eltinge et Fuller (1989).) L'hypothèse fondamentale qui sert de base à l'application des équations ci-dessus est que l'on peut estimer convenablement $\underset{\sim}{\mu}$ et Σ_Y à l'aide des données de la série Y_t et d'un modèle chronologique, et tout aussi convenablement Σ_e à l'aide de micro-données d'enquête et des données du plan de sondage (et probablement aussi à l'aide d'un modèle). Nous examinons plus en détail cette question dans la section 4 et illustrons l'approche chronologique dans la section 5.

2.2 Considérations générales sur l'approche chronologique

Smith (1978), Jones (1980) et Binder et Dick (1986) analysent l'approche dite de l'estimation linéaire non biaisée à variance minimum (ELNVM). Bien que l'ELNVM et l'approche chronologique permettent toutes deux d'estimer θ_t à l'aide de données se rapportant à d'autres périodes que t, elles se distinguent l'une de l'autre en ceci que, dans la première, on tient les valeurs θ_t pour fixes et ne considère qu'une seule forme de variation, soit celle due à l'échantillonnage. L'ELNVM est utilisée spécialement dans les cas où il existe plus d'un estimateur direct de θ_t pour chaque période t et où les e_t sont corrélées d'une période à l'autre à cause de la participation répétée de certaines unités d'échantillonnage (c'est le cas notamment de nombreuses enquêtes par panel avec renouvellement). Ainsi, les résultats des moindres carrés généralisés et la corrélation des e_t font que Y_j, pour $j \neq t$, sert à estimer θ_t. Nous pouvons voir une différence par rapport à l'équation (2.1), où il existe un seul estimateur direct (Y_t) de θ_t, en ce sens que dans ce dernier cas, on obtient l'ELNVM en posant $\text{Var}(\theta_t) \to \infty$. Alors, $\Sigma_\theta^{-1} \to 0$ et l'équation (2.5) devient $\hat{\underset{\sim}{\theta}} = \underset{\sim}{Y}$, de sorte que s'il n'existe pas plusieurs estimateurs de θ_t, l'ELNVM n'utilise que Y_t pour estimer θ_t. Ces remarques concernent ordinairement l'estimation composite (Rao et Graham 1964; Wolter 1979), qui est souvent utilisée comme équivalent de l'ELNVM.

Pour ce qui a trait à l'approche chronologique, il est permis de se demander pourquoi considérer θ_t comme une série stochastique? Cette question a été étudiée par SSJ et analysée assez longuement par Smith (1978). Deux observations ressortent de ces analyses: 1) dans la modélisation, les utilisateurs de données d'enquêtes à passages répétés considèrent Y_t comme une série stochastique et ils envisageraient θ_t de la même façon s'ils en connaissaient les éléments (comme c'est le cas pour les enquêtes où le niveau d'erreur est très faible); 2) dans les textes classiques (par ex.: Patterson, 1950) sur l'estimation dans les enquêtes à passages répétés (ELNVM), on suppose l'existence d'une structure chronologique pour chacune des unités de la population tout en continuant d'affirmer que θ_t, qui est une fonction de ces unités (par ex.: le total), est une série de valeurs fixes n'ayant aucun rapport entre elles. À vrai dire, si nous supposons que θ_t est une suite de valeurs fixes qui n'ont aucun lien entre elles, alors les données relatives à chacune des périodes n'ont aucun rapport avec le comportement futur de la série réelle θ_t. Si tel était le cas, on pourrait s'interroger sur l'utilité même d'une enquête car les données ne seraient pas aussitôt publiées qu'elles seraient périmées. En réalité, il s'agit de savoir si, oui ou non, nous pouvons estimer la structure chronologique de θ_t et de e_t suffisamment bien pour que cela nous soit profitable dans le processus d'estimation, si les avantages en question sont notables et si l'opération comporte des inconvénients et lesquels?

Outre qu'elle fournit l'occasion d'améliorer l'estimation dans les enquêtes à passages répétés, l'approche chronologique peut être utile dans les cas où on ne reconnaît généralement qu'une des deux sources de variation possibles. Elle pourrait bien aussi servir d'approche globale pour tous ces cas. Mentionnons notamment l'estimation provisoire dans les enquêtes à passages répétés (Rao, Srinath et Quenneville 1989); la désaisonnalisation (Wolter et Monsour 1981;

Hausman et Watson 1985; Pfeffermann 1991); l'estimation de la tendance d'une série chronologique (et le problème connexe de la constatation d'une variation statistiquement significative à long terme) (Smith 1978); l'étalonnage, c'est-à-dire le rapprochement des résultats d'une enquête à passages répétés et de ceux d'une autre enquête ou d'un recensement qui sert à estimer les mêmes caractéristiques de la population (Hillmer et Trabelsi 1987; Trabelsi et Hillmer 1990); et l'inférence concernant les propriétés chronologiques de la série θ_t réelle par rapport aux modèles économiques (Bell et Wilcox 1990).

Enfin, notons que les erreurs non dues à l'échantillonnage ne sont pas prévues dans la formule de décomposition (2.1) ou (2.10) et qu'elles ne sont pas considérées de façon explicite dans l'approche chronologique. On ne peut encore dire si ce genre d'erreurs posent habituellement un problème plus sérieux ou moins sérieux pour l'approche chronologique que pour l'approche classique, mais il serait intéressant de connaître les effets que peuvent avoir les erreurs non dues à l'échantillonnage (réelles ou prévues) sur les estimateurs de séries chronologiques lorsque ceux-ci sont utilisés dans des conditions précises.

3. CONSIDÉRATIONS THÉORIQUES

Dans cette section, nous présentons des résultats théoriques concernant l'approche chronologique et exposons certaines propriétés des estimateurs pertinents.

3.1 Convergence des estimateurs de séries chronologiques

Suivant Fuller et Isaki (1981), nous définissons Y_t^ℓ (d'après l'échantillon ℓ au temps t) comme une série d'estimateurs de la caractéristique θ_t^ℓ de la population ℓ au temps t, où les populations et les échantillons pour $\ell = 1, 2, \ldots$ sont emboîtés. (Voir l'article précité pour plus de détails.) Définissons $\underset{\sim}{Y}^\ell, \underset{\sim}{\theta}^\ell, \underset{\sim}{e}^\ell, \underset{\sim}{\mu}^\ell, \Sigma_y^\ell, \Sigma_\theta^\ell, \Sigma_e^\ell, \underset{\sim}{\hat{\theta}}^\ell$, et $\hat{\theta}_t^\ell$ de la manière habituelle. Nous allons voir ce que deviennent les estimateurs de séries chronologiques $\underset{\sim}{\hat{\theta}}^\ell$ lorsque les estimateurs $\underset{\sim}{Y}^\ell$ sont convergents, c.-à-d. $Y_t^\ell \to \theta_t^\ell$ d'une manière quelconque lorsque $\ell \to \infty$ pour $t = 1, \ldots, T$, la longueur de la série (T) étant constante. Pour le moment, nous supposons que $\underset{\sim}{\mu}^\ell$, Σ_θ^ℓ, et Σ_e^ℓ sont connues pour chaque ℓ, ce qui signifie généralement que les modèles chronologiques (y compris les valeurs des paramètres) pour les composantes sont connus. Puisque $\underset{\sim}{\mu}^\ell$ et Σ_θ^ℓ sont en réalité des paramètres de superpopulation pour la série chronologique que nous voulons estimer (θ_t^ℓ), nous supposerons que chacun d'eux est unique pour toutes les populations ℓ, c'est-à-dire $\underset{\sim}{\mu}^\ell \equiv \underset{\sim}{\mu}$ et $\Sigma_\theta^\ell \equiv \Sigma_\theta$ (une matrice définie positive) pour tous ℓ. Nous posons aussi cette hypothèse en partie pour des raisons de commodité, puisque nous obtiendrions les mêmes résultats en supposant que $\underset{\sim}{\mu}^\ell \to \underset{\sim}{\mu}$ et $\Sigma_\theta^\ell \to \Sigma_\theta$ lorsque $\ell \to \infty$.

D'après l'équation (2.5), il semble que $\underset{\sim}{Y}^\ell \to \underset{\sim}{\theta}^\ell$ implique que $\underset{\sim}{\hat{\theta}}^\ell \to \underset{\sim}{\theta}^\ell$ pourvu que $\Sigma_e^\ell \to 0$. Cette condition nous porte à croire que Y_t^ℓ doit converger en moyenne quadratique vers θ_t^ℓ. Nous considérons donc les estimateurs Y_t^ℓ de θ_t^ℓ tels que $E[(Y_t^\ell - \theta_t^\ell)^2] = E[(e_t^\ell)^2] \to 0$ lorsque $\ell \to \infty$. Puisque $E[(e_t^\ell)^2] = \mathrm{Var}(e_t^\ell) + [E(e_t^\ell)]^2$ cela implique que $\mathrm{Var}(e_t^\ell) \to 0$ et $E(e_t^\ell) \to 0$ également. Le fait de supposer que Y_t^ℓ converge en moyenne quadratique vers θ_t^ℓ, pour $t = 1, \ldots, T$, implique donc que $\Sigma_e^\ell \to 0$. Nous pouvons maintenant établir ce qui suit:

Résultat 3.1: Considérons l'estimateur $\underset{\sim}{\hat{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_T)'$ défini en (2.4). Si Y_t^ℓ converge en moyenne quadratique vers θ_t^ℓ lorsque $\ell \to \infty$, pour $t = 1, \ldots, T$, alors $\hat{\theta}_t^\ell$ converge en moyenne quadratique vers θ_t^ℓ lorsque $\ell \to \infty$, pour $t = 1, \ldots, T$.

Démonstration: D'après l'équation $\underset{\sim}{Y}^\ell = \underset{\sim}{\theta}^\ell + \underset{\sim}{e}^\ell$ où $\Sigma_e^\ell \to 0$, nous avons $\Sigma_Y^\ell \to \Sigma_\theta$ (même si $\underset{\sim}{\theta}^\ell$ et $\underset{\sim}{e}^\ell$ sont corrélées). D'après (2.4), nous avons

$$\underset{\sim}{\hat{\theta}}^\ell - \underset{\sim}{\theta}^\ell = (\underset{\sim}{Y}^\ell - \underset{\sim}{\theta}^\ell) - \Sigma_e^\ell (\Sigma_Y^\ell)^{-1} (\underset{\sim}{Y}^\ell - \underset{\sim}{\mu}). \tag{3.1}$$

Le premier terme du membre de droite converge en moyenne quadratique vers $\underset{\sim}{0}$; le second terme du membre de droite a pour moyenne $\underset{\sim}{0}$ et pour variance $\Sigma_e^\ell (\Sigma_Y^\ell)^{-1} \Sigma_e^\ell$, qui converge vers 0 lorsque $\ell \to \infty$. Comme les deux termes convergent en moyenne quadratique vers $\underset{\sim}{0}$, il en est de même pour $\hat{\underset{\sim}{\theta}}^\ell - \underset{\sim}{\theta}^\ell$.

La convergence en probabilité est une notion plus connue en théorie des sondages. Si Y_t^ℓ converge en probabilité vers θ_t^ℓ lorsque $\ell \to \infty$, pour $t = 1, \ldots, T$, cela ne veut pas nécessairement dire que $\Sigma_e^\ell \to 0$, qui est représentatif de la convergence en moyenne quadratique, un mode de convergence plus strict. Si nous supposons qu'il existe des variables aléatoires ζ_t ayant une variance finie de telle sorte que $| e_t^\ell | \leq \zeta_t$ (avec quasi-certitude) uniformément en ℓ, alors la convergence en probabilité de Y_t^ℓ vers θ_t^ℓ impliquera une convergence en moyenne quadratique de Y_t^ℓ vers θ_t^ℓ (Chung 1968, p. 64). Par conséquent, nous établissons ce qui suit à l'aide du Résultat 3.1:

Résultat 3.2: Si Y_t^ℓ converge en probabilité vers θ_t^ℓ lorsque $\ell \to \infty$, pour $t = 1, \ldots, T$ et qu'il existe des variables aléatoires ζ_t ayant une variance finie de telle sorte que $| Y_t^\ell - \theta_t^\ell | \leq \zeta_t$ uniformément en ℓ, alors $\hat{\theta}_t^\ell$ converge en probabilité vers θ_t^ℓ lorsque $\ell \to \infty$, pour $t = 1, \ldots, T$.

Les résultats ci-dessus montrent que si l'erreur rattachée aux estimations originales Y_t de θ_t est faible (Σ_e est petite), l'écart $\hat{\theta}_t - \theta_t$ sera aussi faible. Cela vient de ce que, dans l'équation (3.1), $\hat{\underset{\sim}{\theta}} - \underset{\sim}{Y}$ s'amenuise lorsque Σ_e s'amenuise; ainsi, lorsque les estimations originales Y_t sont entachées d'une faible erreur, l'approche chronologique n'a pas vraiment d'effet sur ces estimations. Binder et Dick (1986) ont noté le phénomène et précisé que, dans un tel cas, le modèle chronologique utilisé importe peu. Autrement dit, (3.1) convergera vers $\underset{\sim}{0}$ si $\Sigma_e^\ell \to 0$, et la condition de convergence n'a rien à voir avec $\underset{\sim}{\mu}$ ou Σ_θ. En vertu des résultats ci-dessus, nous pouvons donc remplacer $\underset{\sim}{\mu}$, Σ_θ et Σ_e^ℓ par les estimations $\hat{\underset{\sim}{\mu}}^\ell$, $\hat{\Sigma}_\theta^\ell$ et $\hat{\Sigma}_e^\ell$ (qui viennent le plus souvent de modèles estimés - voir sections 4 et 5), à la condition que $\hat{\underset{\sim}{\mu}}^\ell$ et $\hat{\Sigma}_\theta^\ell$ convergent vers une valeur quelconque lorsque $\ell \to \infty$ (la valeur a peu d'importance, pourvu que la limite de $\hat{\Sigma}_\theta^\ell$ soit définie positive) et que $\hat{\Sigma}_e^\ell \to 0$, ce qui devrait être le cas si $\Sigma_e^\ell \to 0$. En outre, il est clair que les résultats ci-dessus valent pour les séries non stationnaires, où $\hat{\underset{\sim}{\theta}}$ est défini par (2.7) au lieu de (2.4). Autant l'approche chronologique donne peu de résultats lorsque les estimations originales Y_t sont entachées d'une faible erreur, autant elle est profitable dans le cas contraire - c'est-à-dire lorsque $\mathrm{Var}(e_t)$ est élevée.

Nous pouvons aussi étudier la convergence d'estimateurs dans les cas où nous utilisons des logarithmes et estimons $\log(\theta_t)$ au moyen de l'équation (2.11). Posons alors $\Sigma_u^\ell = \mathrm{Var}(\log(\underset{\sim}{u}^\ell))$ où $\log(\underset{\sim}{u}^\ell) \equiv (\log(u_1^\ell), \ldots, \log(u_T^\ell))'$. Si nous utilisons les logarithmes, il est raisonnable de supposer que Y_t^ℓ et θ_t^ℓ ne peuvent avoir de valeur nulle, c'est-à-dire $| Y_t^\ell | \geq \kappa$ et $| \theta_t^\ell | \geq \kappa$ (avec quasi-certitude) pour tous t et tous ℓ pour une constante $\kappa > 0$.

Résultat 3.3: Si Y_t^ℓ converge en moyenne quadratique vers θ_t^ℓ lorsque $\ell \to \infty$, pour $t = 1, \ldots, T$, alors $\log(Y_t^\ell)$ converge en moyenne quadratique vers $\log(\theta_t^\ell)$ et $\log(\hat{\theta}_t^\ell)$ converge également en moyenne quadratique vers $\log(\theta_t^\ell)$ lorsque $\ell \to \infty$, pour $t = 1, \ldots, T$.

Démonstration: L'équivalent logarithmique de (3.1) est

$$\log(\hat{\underset{\sim}{\theta}}^\ell) - \log(\underset{\sim}{\theta}^\ell) = (\log(\underset{\sim}{Y}^\ell) - \log(\underset{\sim}{\theta}^\ell) - \Sigma_u^\ell (\Sigma_Y^\ell)^{-1} (\log(\underset{\sim}{Y}^\ell) - \underset{\sim}{\mu}).$$

Si nous pouvons montrer que $\Sigma_u^\ell \to 0$, la démonstration sera faite car cette convergence implique que $\log(\underset{\sim}{Y}^\ell)$ converge en moyenne quadratique vers $\log(\underset{\sim}{\theta}^\ell)$, et le second terme du membre de droite de l'équation suit la même tendance que le terme correspondant de l'équation (3.1). Notons que

$$E[(\tilde{u}_t^\ell)^2] = E[(e_t^\ell)^2 / (\theta_t^\ell)^2] \leq (E(e_t^\ell)^2)/\kappa^2 \to 0 \quad \text{lorsque} \quad \ell \to \infty,$$

par conséquent, $E[(\tilde{u}_t^\ell)^2] = E[(u_t^\ell - 1)^2] \to 0$. Donc, $\mathrm{Var}(u_t^\ell) \to 0$ et $E(u_t^\ell) \to 1$. D'après l'inégalité de Jensen (Chung 1968, p. 45), puisque $\exp(\cdot)$ est une fonction convexe,

$$1 \leq \exp\left(E\left[\log\left(u_t^\ell\right)^2\right]\right) \leq E\left(\exp\left[\log\left(u_t^\ell\right)^2\right]\right) = E\left[\left(u_t^\ell\right)^2\right].$$

Or, $E[(u_t^\ell)^2] = \text{Var}(u_t^\ell) + [E(u_t^\ell)]^2 \to 1$, de sorte que $\exp(E[\log(u_t^\ell)^2]) \to 1$ et, partant, $E[\log(u_t^\ell)^2] \to 0$. Nous obtenons ainsi $\text{Var}(\log(u_t^\ell)) \to 0$, le résultat recherché.

Nous pourrions aussi observer une convergence en probabilité en fixant une borne pour $\log(u_t^\ell)$. Étant donné que $\log(\hat{\theta}_t)$ est un estimateur de $\log(\theta_t)$, nous formulons ci-dessous un corollaire du Résultat 3.3 concernant l'utilisation de $\exp[\log(\hat{\theta}_t)]$ comme estimateur de θ_t.

Corollaire 3.4: Si Y_t^ℓ converge en moyenne quadratique vers θ_t^ℓ lorsque $\ell \to \infty$, pour $t = 1, \ldots, T$, alors (voir (2.11)) $\exp[\log(\hat{\theta}_t^\ell)]$ converge en probabilité vers θ_t^ℓ lorsque $\ell \to \infty$, pour $t = 1, \ldots, T$.

Démonstration: Puisque la convergence en moyenne quadratique de $\log(\hat{\theta}_t^\ell)$ vers $\log(\theta_t^\ell)$ implique qu'il y a aussi convergence en probabilité, la démonstration du corollaire est faite étant donné que $\exp(\cdot)$ est une fonction continue (Chung 1968, p. 66).

Nous arrivons évidemment aux mêmes conclusions pour ce qui a trait à l'utilisation de $\exp[\log(\hat{\theta}_t^\ell) + \text{Var}(\log(\hat{\theta}_t^\ell) - \log(\theta_t^\ell))/2]$ comme estimateur de θ_t, car $\text{Var}(\log(\hat{\theta}_t^\ell) - \log(\theta_t^\ell)) \to 0$ lorsque $\ell \to \infty$.

3.2 Absence de corrélation entre θ et e

Dans sa forme classique, l'extraction de signal de séries chronologiques, représentée par les équations (2.3) à (2.8), suppose invariablement que θ_t et e_t sont non corrélées à toutes les avances et à tous les retards (suivant l'hypothèse de la normalité, nous dirions que ces séries sont indépendantes l'une de l'autre). Dans les articles antérieurs sur l'approche chronologique appliquée à l'estimation dans les enquêtes à passages répétés, on se limite à supposer la même chose. Or, il n'est pas sûr que cette hypothèse soit fondée car θ_t et e_t dépendent des mêmes unités de population. Heureusement, nous pouvons établir qu'elle est juste suivant des conditions assez générales. (Tam (1987) explique comment une telle hypothèse peut ne plus être valide dans une approche fondée explicitement sur un modèle.)

Posons y_{it} comme la valeur de la caractéristique étudiée pour l'unité i de la population au temps t, et $\Omega_t = \{y_{it}: i = 1, \ldots, N_t\}$ comme la série de valeurs pour les N_t unités. Considérons les périodes $t = 1, \ldots, T$ et posons $\underset{\sim}{\Omega} = (\Omega_1, \ldots, \Omega_T)'$. Les y_{it} sont des variables aléatoires, tout comme $\theta_t = \theta_t(\Omega_t)$, qui est une fonction des y_{it}. L'échantillon au temps t, s_t (ceci désignant l'indice et non la valeur des unités choisies), a une probabilité de sélection $p(s_t \mid \underset{\sim}{\Omega})$. L'estimateur Y_t de θ_t est une fonction des valeurs y_{it} pour les unités échantillonnées, donc fonction de Ω_t et de s_t, c.-à-d. $Y_t = Y_t(\Omega_t, s_t)$. Nous pourrions faire en sorte que Y_t dépende d'échantillons se rapportant à d'autres périodes que t, mais nous nous en abstiendrons pour des raisons de simplicité.

Nous considérons des estimateurs Y_t de θ_t qui sont *non biaisés selon le plan* et que nous définirons $E(Y_t \mid \underset{\sim}{\Omega}) \equiv \sum_{s_t} Y_t p(s_t \mid \underset{\sim}{\Omega}) = \theta_t$. Nous pourrions exprimer aussi bien ces estimateurs par la formule $E(Y_t \mid \Omega_t) \equiv \sum_{s_t} Y_t p(s_t \mid \Omega_t) = \theta_t$; il faudrait alors supposer que le processus d'échantillonnage est tel que $p(s_t \mid \underset{\sim}{\Omega}) = p(s_t \mid \Omega_t)$, de sorte que $E(Y_t \mid \underset{\sim}{\Omega}) = E(Y_t \mid \Omega_t)$. Si le plan d'échantillonnage est non informatif, s_t et $\underset{\sim}{\Omega}$ sont indépendants, ce qui implique que $p(s_t \mid \underset{\sim}{\Omega}) = p(s_t \mid \Omega_t) = p(s_t)$, et les deux formules ci-dessus se ramènent à $\sum_{s_t} Y_t p(s_t) = \theta_t$. Cette dernière formule correspond à la définition habituelle, qui suppose en règle générale que les y_{it}, et par voie de conséquence Ω_t et θ_t, sont fixes. (L'hypothèse $p(s_t \mid \underset{\sim}{\Omega}) = p(s_t \mid \Omega_t)$ fait que le processus d'échantillonnage au temps t ($p(s_t \mid \underset{\sim}{\Omega})$) dépend des valeurs de la population au temps t (Ω_t) et suppose que les valeurs de la population pour d'autres périodes que t (Ω_j pour $j \neq t$) n'apportent aucune nouvelle information à propos de s_t. Ce genre de situation peut se produire lorsque l'échantillonnage se fait avec une probabilité proportionnelle à la taille d'une variable auxiliaire au temps t et que cette variable auxiliaire est en corrélation avec les valeurs y_{it} uniquement à cette période.) Les hypothèses avancées ici pourraient, à la rigueur, être généralisées.

Résultat 3.5: Si Y_t est non biaisé selon le plan pour tous t, alors θ_t et e_t sont des séries chronologiques non corrélées.

Démonstration: Considérons $\text{Cov}(\theta_t, e_j)$ pour deux périodes quelconques t et j. Comme Y_j est non biaisé selon le plan, $E(e_j \mid \underset{\sim}{\Omega}) = E(Y_j - \theta_j \mid \underset{\sim}{\Omega}) = 0$ et par conséquent, $E[E(e_j \mid \underset{\sim}{\Omega})] = E(e_j) = 0$. De plus, $E(\theta_t \cdot e_j \mid \underset{\sim}{\Omega}) = \theta_t \cdot E(e_j \mid \underset{\sim}{\Omega}) = 0$, ce qui implique que $E(\theta_t \cdot e_j) = 0$. Donc, $\text{Cov}(\theta_t, e_j) = E(\theta_t \cdot e_j) - E(\theta_t)E(e_j) = 0$.

Remarque: Si $E(e_j \mid \underset{\sim}{\Omega})$ ne dépend pas d'$\underset{\sim}{\Omega}$, alors on dira que e_j est indépendante d'$\underset{\sim}{\Omega}$ "en espérance", ce qui est plus fort que l'absence de corrélation entre e_j et $\underset{\sim}{\Omega}$ mais moins fort que l'indépendance stochastique (sauf s'il y a normalité). En réalité, il suffit que $E(e_t \mid \underset{\sim}{\Omega}) = E(Y_t \mid \underset{\sim}{\Omega}) - \theta_t$ ne dépende pas d'$\underset{\sim}{\Omega}$ pour que θ_t et e_t soient des séries non corrélées. On verra cela dans les cas où Y_t, comme estimateur de θ_t, est entaché d'un biais additif constant (ne dépendant pas d'Ω_t) ou (si l'on se fonde sur le résultat 3.6 ci-dessous) d'un biais en pourcentage (multiplicatif) constant.

Considérons maintenant la formule de décomposition logarithmique (2.10) pour le cas où les Y_t sont non biaisés selon le plan. Nous supposons que $\tilde{u}_j$ est représenté par $O_p(r_\ell)$, où $r_\ell \to 0$ lorsque $\ell \to \infty$, comme dans le modèle de superpopulation de la section précédente; toutefois, pour des raisons de commodité, nous omettons ici l'indice supérieur ℓ des variables aléatoires. (Voir Wolter (1985, p. 222) pour une définition du degré dans le symbole de probabilité $O_p(r_\ell)$. Par exemple, pour l'estimation d'une moyenne de population, nous aurions souvent $\text{Var}(\tilde{u}_j) \leq K/n_{j\ell}$, où K est une constante et $n_{j\ell}$, la taille de l'échantillon au temps j pour la population ℓ. Alors, d'après Wolter (1985, théorème 6.2.1), $\tilde{u}_j = O_p(n_{j\ell}^{-.5})$.) En appliquant la méthode de linéarisation de Taylor à $\log(u_j) = \log(1 + \tilde{u}_j)$, nous obtenons (d'après Wolter (1985, théorème 6.2.2)

$$\log(u_j) = \tilde{u}_j + O_p(r_\ell^2). \tag{3.2}$$

Cette équation nous amène au résultat suivant:

Résultat 3.6: Si Y_t est non biaisé selon le plan pour tous t et $\tilde{u}_j$ est $O_p(r_\ell)$, alors les termes $O_p(r_\ell^3)$, $\log(\theta_t)$ et $\log(u_t)$ représentent des séries chronologiques non corrélées.

Démonstration: D'après le théorème 6.2.5 de Wolter (1985), $\text{Cov}(\log(\theta_t), \log(u_j)) = \text{Cov}(\log(\theta_t), \tilde{u}_j) + O_p(r_\ell^3)$. Notons que $E(\tilde{u}_j \mid \underset{\sim}{\Omega}) = E(e_j \mid \underset{\sim}{\Omega})/\theta_j = 0$ suppose que $E(\tilde{u}_j) = 0$, et $E(\log(\theta_t)\tilde{u}_j \mid \underset{\sim}{\Omega}) = \log(\theta_t)E(\tilde{u}_j \mid \underset{\sim}{\Omega}) = 0$ suppose que $E(\log(\theta_t)\tilde{u}_j) = 0$; par conséquent, $\text{Cov}(\log(\theta_t), \tilde{u}_j) = 0$, ce qu'il fallait démontrer.

3.3 Propriétés de plan des estimations d'extraction de signal

Inconditionnellement, l'estimateur $\hat{\underset{\sim}{\theta}}$ définie en (2.3) est non biaisé ($E(\hat{\underset{\sim}{\theta}}) = E(\underset{\sim}{\theta}) = \underset{\sim}{\mu}$) et a une EQM minimum telle que définie en (2.6). Lorsqu'on envisage cet estimateur par rapport à un plan, il est facile de constater que ces propriétés ne sont plus aussi vraies. Supposons que nous avons des estimateurs $\underset{\sim}{Y}$ non biaisés selon le plan, c.-à-d. $E(\underset{\sim}{Y} \mid \underset{\sim}{\Omega}) = \underset{\sim}{\theta}$. D'après (2.2) et (2.4) nous avons $\hat{\underset{\sim}{\theta}} - \underset{\sim}{\theta} = (I - \Sigma_e \Sigma_Y^{-1})\,\underset{\sim}{e} - \Sigma_e \Sigma_Y^{-1}(\underset{\sim}{\theta} - \underset{\sim}{\mu})$. Par quelques opérations algébriques, nous pouvons exprimer le biais, la variance et l'EQM de plan de $\hat{\underset{\sim}{\theta}}$:

$$\begin{aligned}
E(\hat{\underset{\sim}{\theta}} \mid \underset{\sim}{\Omega}) - \underset{\sim}{\theta} &= -\Sigma_e \Sigma_Y^{-1}(\underset{\sim}{\theta} - \underset{\sim}{\mu}), \\
\text{Var}(\hat{\underset{\sim}{\theta}} - \underset{\sim}{\theta} \mid \underset{\sim}{\Omega}) &= \Sigma_e - \Sigma_e \Sigma_Y^{-1} \Sigma_e - \Sigma_e \Sigma_Y^{-1} \Sigma_\theta \Sigma_Y^{-1} \Sigma_e, \\
E[(\hat{\underset{\sim}{\theta}} - \underset{\sim}{\theta})(\hat{\underset{\sim}{\theta}} - \underset{\sim}{\theta})' \mid \underset{\sim}{\Omega}] &= \Sigma_e - \Sigma_e \Sigma_Y^{-1} \Sigma_e \\
&\quad - \Sigma_e \Sigma_Y^{-1} \left[\Sigma_\theta - (\underset{\sim}{\theta} - \underset{\sim}{\mu})(\underset{\sim}{\theta} - \underset{\sim}{\mu})'\right] \Sigma_Y^{-1} \Sigma_e.
\end{aligned} \tag{3.3}$$

Nous voyons que l'utilisation de $\hat{\underset{\sim}{\theta}}$ dans une perspective de plan engendre une variance moins élevée mais un biais plus grand puisque $\Sigma_e - \text{Var}(\hat{\underset{\sim}{\theta}} - \underset{\sim}{\theta} \mid \underset{\sim}{\Omega})$ est une matrice semi-positive. Quant à savoir si cette application rend l'EQM conditionnelle (3.3) inférieure à Σ_e, l'EQM de $\underset{\sim}{Y}$, cela dépend des deux derniers termes de (3.3) et, par conséquent, de $\underset{\sim}{\theta}$. Il peut y avoir des valeurs de $\underset{\sim}{\theta}$ pour lesquelles l'EQM conditionnelle de $\hat{\underset{\sim}{\theta}}$ est supérieure à Σ_e mais, en règle générale, l'extraction de signal a pour effet de réduire l'EQM d'une valeur équivalant à $\Sigma_e \Sigma_Y^{-1} \Sigma_e$ puisque l'espérance inconditionnelle du terme entre crochets dans l'équation (3.3) est nulle. (Évidemment, l'équation (3.3) est inutilisable en pratique puisqu'elle dépend de $\underset{\sim}{\theta}$.) De plus, comme nous l'avons mentionné plus tôt, l'erreur de modélisation accroîtra l'EQM de $\hat{\underset{\sim}{\theta}}$; cela nous amène à nous poser une autre question fondamentale, à laquelle il est encore plus difficile de répondre (voir Eltinge et Fuller 1989): l'EQM inconditionnelle réelle de $\hat{\underset{\sim}{\theta}}$ est-elle inférieure, égale ou supérieure à Σ_e?

4. CONSIDÉRATIONS RELATIVES À L'APPLICATION

Nous avons vu dans la section 2 que pour appliquer l'approche chronologique à l'estimation dans les enquêtes, il fallait estimer la structure d'autocovariances des erreurs d'échantillonnage de même que la moyenne et la structure d'autocovariances du signal et calculer les valeurs estimées $\hat{\theta}_t$ et $\text{Var}(\hat{\theta}_t - \theta_t)$. Pour ce qui est des opérations d'estimation, on se sert habituellement de modèles chronologiques. Ces opérations sont d'ailleurs analysées dans Bell et Hillmer (1989). La présente section renferme des remarques générales. Nous supposons que Y_t est un estimateur non biaisé selon le plan de θ_t. Dans la section suivante, nous donnons un exemple d'application en nous servant de deux séries chronologiques tirées de la Retail Trade Survey du U.S. Bureau of the Census.

Les autocovariances d'erreurs d'échantillonnage, $\text{Cov}(e_t, e_{t+k})$, peuvent être estimées comme les variances d'échantillonnage, $\text{Var}(e_t)$; l'opération est simple et peut s'effectuer suivant plusieurs méthodes. (Voir Wolter (1985).) Dans la pratique, il peut s'avérer difficile de raccorder des microdonnées d'enquête sur plusieurs périodes afin d'estimer directement les covariances d'erreurs d'échantillonnage. Néanmoins, nous allons supposer ici que nous disposons d'estimations de ce genre, $\widehat{\text{Cov}}(e_t, e_{t+k})$, pour une série de périodes t et de décalages k. Malheureusement, s'il devait y avoir des erreurs d'échantillonnage importantes (circonstances où les méthodes d'analyse chronologique peuvent s'avérer particulièrement utiles), les estimations d'autocovariance afficheraient probablement elles-mêmes une forte variance. Cela donne à penser que l'on pourrait faire une sorte de calcul de moyennes afin d'améliorer ces estimations.

Premièrement, si nous supposons que e_t est stationnaire en covariance, de sorte que $\text{Cov}(e_t, e_{t+k}) \equiv \gamma_e(k)$ dépend de k mais non de t, alors chaque valeur $\widehat{\text{Cov}}(e_t, e_{t+k})$ est une estimation de $\gamma_e(k)$ et il suffit de faire la moyenne de ces valeurs, c.-à-d. calculer $\hat{\gamma}_e(k) = (T-k)^{-1} \sum_t \widehat{\text{Cov}}(e_t, e_{t+k})$ si nous connaissons $\widehat{\text{Cov}}(e_t, e_{t+k})$ pour $t = 1, \ldots, T-k$. Par ailleurs, on peut calculer la moyenne de $\widehat{\text{Corr}}(e_t, e_{t+k}) = \widehat{\text{Cov}}(e_t, e_{t+k}) / [\widehat{\text{Var}}(e_t)\, \widehat{\text{Var}}(e_{t+k})]^{.5}$ par rapport à t pour estimer $\text{Corr}(e_t, e_{t+k})$, qui dépend aussi de k mais non de t, et on peut, comme avant, faire la moyenne des estimations de la variance.

Supposons maintenant que e_t est stationnaire en covariance relative, de sorte que $\text{Cov}(e_t/\theta_t, e_{t+k}/\theta_{t+k}) = \text{Cov}(\tilde{u}_t, \tilde{u}_{t+k}) \equiv \gamma_u(k)$ dépend de k mais non de t. Si $\tilde{u}_t$ est $O_p(r_t)$ pour tous t, comme dans la section 3.2, alors d'après l'équation (3.2) et le théorème 6.2.5 de Wolter (1985), $\text{Cov}(\log(u_t), \log(u_{t+k})) = \text{Cov}(\tilde{u}_t, \tilde{u}_{t+k}) + O_p(r_t^3) \approx \gamma_u(k)$. Considérant que $\widehat{\text{Cov}}(e_t, e_{t+k}) / (Y_t Y_{t+k})$ est une estimation de $\text{Cov}(\tilde{u}_t, \tilde{u}_{t+k})$, on peut faire la moyenne de ces estimations par rapport à t pour estimer $\gamma_u(k)$. Par ailleurs, en nous servant du corollaire 5.1.5 de Fuller (1976), nous pouvons montrer que $\text{Corr}(\log(u_t), \log(u_{t+k})) = \text{Corr}(\tilde{u}_t, \tilde{u}_{t+k}) + O_p(r_t^3)$, et en considérant que $\{\widehat{\text{Cov}}(e_t, e_{t+k}) / Y_t Y_{t+k}\} / \{[\widehat{\text{Var}}(e_t)\, \widehat{\text{Var}}(e_{t+k})]^{.5} / Y_t Y_{t+k}\} = \widehat{\text{Corr}}(e_t, e_{t+k})$ donne une estimation de $\rho_u(k) \equiv \text{Corr}(\tilde{u}_t, \tilde{u}_{t+k})$, nous pouvons faire la

moyenne des autocorrélations estimées de e_t par rapport à t pour estimer $\rho_u(k)$, qui équivaut à peu près à l'autocorrélation de $\log(u_t)$. On peut, comme avant, faire la moyenne des estimations de la variance relative.

En réalité, les valeurs estimées de variances et d'autocovariances seront le plus souvent des estimations de $\text{Var}(e_t \mid \underset{\sim}{\Omega})$ et de $\text{Cov}(e_t, e_{t+k} \mid \underset{\sim}{\Omega})$. Elles pourront peut-être servir d'estimations pour $\text{Var}(e_t)$ et $\text{Cov}(e_t, e_{t+k})$ si, par exemple, elles peuvent convenir dans une perspective de modèle. Dans le cas contraire, et si Y_t est non biaisé selon le plan de sorte que $E(e_t \mid \underset{\sim}{\Omega}) = 0$, il ne sera pas moins justifié de faire la moyenne des estimations d'autocovariances par rapport à t. Premièrement, si nous supposons que e_t est stationnaire, alors $\gamma_e(k) \equiv \text{Cov}(e_t, e_{t+k}) = E[\text{Cov}(e_t, e_{t+k} \mid \underset{\sim}{\Omega})]$; nous pouvons donc calculer la moyenne des valeurs estimées de $\text{Cov}(e_t, e_{t+k} \mid \underset{\sim}{\Omega})$ pour estimer $\gamma_e(k)$. Ou encore, si e_t est stationnaire en covariance relative, $E(\tilde{u}_t \mid \underset{\sim}{\Omega}) = E(e_t \mid \underset{\sim}{\Omega})/\theta_t = 0$; par conséquent, nous avons $\gamma_u(k) \equiv \text{Cov}(\tilde{u}_t, \tilde{u}_{t+k}) = E[\text{Cov}(\tilde{u}_t, \tilde{u}_{t+k} \mid \underset{\sim}{\Omega})] = \text{Cov}(\log(u_t), \log(u_{t+k})) + O_p(r_\ell^3)$, et nous pouvons faire la moyenne des valeurs estimées de $\text{Cov}(\tilde{u}_t, \tilde{u}_{t+k} \mid \underset{\sim}{\Omega})$ pour estimer $\gamma_u(k)$. Quant au calcul de la moyenne des valeurs estimées des corrélations conditionnelles (par rapport à $\underset{\sim}{\Omega}$), il est plus difficile d'en démontrer le bien-fondé puisque $E[\text{Corr}(e_t, e_{t+k} \mid \underset{\sim}{\Omega})] \neq \text{Corr}(e_t, e_{t+k})$; cependant, l'inverse pourrait être vrai si nous posons certaines hypothèses. En général, les méthodes d'estimation des structures d'autocovariances des erreurs d'échantillonnage nécessitent une étude plus poussée.

À partir d'une structure estimée des covariances des erreurs d'échantillonnage et de renseignements pertinents sur le plan de l'enquête, nous pouvons tenter de définir un modèle chronologique (y compris les paramètres) qui reproduira fidèlement cette structure. C'est ce à quoi nous nous attachons dans la section 5.

Nous allons maintenant élaborer un modèle pour le signal, θ_t. Comme le signal est ce qui explique en très grande partie le mouvement de la plupart des séries publiées Y_t (autrement, ces séries ne seraient pas publiées), l'élaboration de modèles pour les signaux θ_t peut s'inspirer des cas antérieurs de modélisation des séries Y_t où l'on fait abstraction des erreurs d'échantillonnage. Cela donne à croire que les modèles de signaux feront une place importante aux transformations non linéaires, à l'application de différences et aux fonctions de moyenne de régression. Le logarithme est la transformation non linéaire qui est la plus couramment appliquée aux séries chronologiques et à ce propos, $\log(Y_t)$ nous permet de modéliser $\log(\theta_t)$ grâce à l'équation (2.10), avec les conséquences que l'on sait pour l'erreur d'échantillonnage. Les remarques qui suivent s'appliquent plus particulièrement à l'utilisation de (2.1) mais elles valent aussi pour l'utilisation de (2.10). Bien que l'on pourrait envisager d'autres types de transformation que la transformation logarithmique, celles-là ne permettraient pas, en règle générale, d'exprimer convenablement les séries transformées Y_t en fonction de séries transformées θ_t et de l'erreur d'échantillonnage. Par conséquent, pour la modélisation de nombreuses séries chronologiques, il suffit, semble-t-il, de choisir entre la transformation logarithmique ou pas de transformation du tout.

Dans l'hypothèse où e_t a une moyenne nulle (suivant la propriété d'être non biaisé selon le plan) et qu'il n'est pas nécessaire de lui appliquer des différences, on devra appliquer les mêmes différences à θ_t et à Y_t et celles-ci auront la même fonction de moyenne. On peut souvent modéliser la fonction de moyenne à l'aide d'une fonction de régression linéaire, $\mu_t = \underset{\sim}{X}_t'\underset{\sim}{\beta}$, pour un vecteur de variables de régression $\underset{\sim}{X}_t$ et de paramètres $\underset{\sim}{\beta}$. On se sert souvent de modèles ARMMI (autorégressifs à moyennes mobiles intégrés) pour représenter les différences devant être appliquées et décrire la structure d'autocovariances des valeurs θ_t, une fois les différences appliquées. Une façon convenable d'élaborer un modèle pour θ_t consiste tout d'abord à modéliser Y_t en faisant abstraction de l'erreur d'échantillonnage, puis à définir un modèle ayant les mêmes termes de régression et le même ordre ARMMI pour θ_t. On peut ensuite estimer les paramètres du modèle de θ_t à l'aide des données de la série Y_t et du modèle élaboré antérieurement pour e_t, les paramètres de ce dernier étant fixes. Par un test de diagnostic, on peut être amené à apporter des modifications au modèle de θ_t. Le modèle ajusté

définitif peut ensuite servir à l'estimation d'extraction de signal de θ_t. Les calculs relatifs à l'ajustement de modèle et à l'extraction de signal sont complexes; Bell et Hillmer (1989) font une analyse des algorithmes du filtre de Kalman. Ces algorithmes ont trouvé une application dans un logiciel élaboré récemment avec le concours de spécialistes de l'analyse chronologique de la Statistical Research Division du U.S. Bureau of the Census. C'est ce logiciel dont nous nous servons dans la section suivante.

5. EXEMPLE: U.S. RETAIL TRADE SURVEY - VENTES DES RESTAURANTS ET DES DÉBITS DE BOISSONS

À titre d'exemple, nous allons analyser la série chronologique des ventes (en millions de dollars) des établissements de restauration (ou restaurants) et des débits de boissons, qui sont estimées dans l'enquête mensuelle sur le commerce de détail aux É.-U. L'échantillon de la Retail Trade Survey est composé d'un panel de grandes entreprises qui sont prélevées dans un répertoire avec une probabilité égale à un et qui déclarent leurs ventes mensuellement, et de trois panels avec renouvellement formés d'entreprises de moindre importance qui sont prélevées dans un répertoire suivant un échantillonnage aléatoire simple stratifié. Notons aussi l'existence d'un panel aréolaire avec renouvellement, formé d'entreprises qui ne figurent pas à un répertoire. À chaque trimestre, on introduit dans l'échantillon global un certain nombre d'entreprises nouvellement créées, puis on supprime du même échantillon les entreprises réputées disparues selon les derniers rapports d'activité. Les entreprises qui font partie des panels avec renouvellement déclarent leurs ventes du mois courant et du mois précédent, les unes (panels de répertoire) à tous les 3 mois, les autres (panel aréolaire) à tous les 6 ou 12 mois. On établit alors des estimations d'Horvitz-Thompson (HT) des ventes du mois courant et du mois précédent; les séries chronologiques qui en découlent sont désignées par Y_t' et Y_{t-1}'. Ces séries servent ensuite à construire des estimateurs composites de la manière décrite dans Wolter (1979). Enfin, les estimations composites ainsi obtenues forment la série chronologique Y_t. (Bien qu'il serait intéressant d'analyser directement Y_t' et Y_{t-1}' au lieu de faire les opérations ci-dessus, les estimations qui forment ces séries ne sont pas conservées suffisamment longtemps pour pouvoir servir à la modélisation de séries chronologiques saisonnières.) Les variances d'échantillonnage sont estimées au moyen de la méthode des groupes aléatoires (Wolter 1985, chapitre 2), dans le cas des panels de répertoire (16 groupes), ou de la méthode de regroupement des strates, dans le cas du panel aréolaire. Pour plus de détails sur la Retail Trade Survey, voir Isaki et coll. (1976); Wolter et coll. (1976); Wolter (1979); Garrett, Detlefsen et Veum (1987) ainsi que Bell et Wilcox (1990).

La Retail Trade Survey présente plusieurs contraintes pour notre analyse. Premièrement, le plan de sondage est remanié et l'échantillon reconstitué à tous les cinq ans environ, les dernières révisions ayant eu lieu en septembre 1977, en janvier 1982 et en janvier 1987. Cela a pour effet de créer un bris de continuité dans la structure de covariances de e_t à tous les cinq ans; le filtre de Kalman peut remédier à cette situation, comme l'indiquent Bell et Hillmer (1989). Nous utiliserons des données pour la période allant de septembre 1977 à décembre 1986, de sorte qu'une reconstitution de l'échantillon coincide à peu près avec le milieu de notre série. Deuxièmement, lorsqu'un nouvel échantillon d'entreprises est introduit, on se sert d'ELNVM approximatifs pour les trois premiers mois, avant d'utiliser les estimateurs composites (Wolter 1979). Cela crée un effet transitoire dans les autocorrélations d'erreurs d'échantillonnage, effet dont nous ne tiendrons pas compte. Enfin, les estimations mensuelles de l'enquête sont ajustées en fonction de totaux annuels établis à la suite d'une enquête annuelle et du recensement économique quinquennal. Pour éliminer cette contrainte, nous allons nous servir de données qui ne sont pas étalonnées. Toutefois, il ne faudra pas s'étonner que les données que nous utilisons ici ne concordent pas avec les estimations publiées.

Tableau 1

Corrélations d'erreurs d'échantillonage pour les estimations d'Horvitz-Thompson

	Décalage					
	4	8	12	16	20	24
Restaurants						
D'après le calcul d'une moyenne[1]	.72	.71	.79	.63	.65	.77
D'après (5.1)[2]	.75	.69	.81	.60	.53	.61
Débits de boissons						
D'après le calcul d'une moyenne[1]	.70	.67	.78	.60	.60	.61
D'après (5.1)[2]	.72	.66	.80	.56	.50	.59
Nombre de corrélations ayant servi au calcul de moyennes	23	19	15	11	7	3
Poids utilisés dans le calcul de $\hat{\phi}$	1	1	1	.5	0	0

[1] Il existait des estimations brutes de Corr(e_t', e_j') et de Corr(e_{t-1}', e_{j-1}') pour toutes les paires de mois de janvier 1973 à mars 1975. On a fait la moyenne des corrélations pour les décalages indiqués après avoir appliqué la transformation de Fisher puis transformé à nouveau les résultats ainsi obtenus.

[2] Corrélations calculées à l'aide du modèle (5.1) pour $m = 4$ avec comme paramètres $\hat{\phi}^4 = .604$, $\hat{\phi}_{12} = .723$ (restaurants) et $\hat{\phi}^4 = .580$, $\hat{\phi}_{12} = .714$ (débits de boissons). Ces valeurs ont été déterminées à l'aide des poids figurant dans le tableau pour minimiser la somme pondérée des carrés des écarts entre les corrélations calculées à l'aide de (5.1) et celles ayant servi au calcul de moyennes. Les décalages 20 et 24 n'ont pas servi au calcul (poids nul) à cause du faible nombre de corrélations estimées correspondant à ces décalages.

5.1 Élaboration de modèles d'erreur d'échantillonnage

Nous allons, en premier lieu, élaborer un modèle pour la structure de corrélation des erreurs d'échantillonnage. Exprimons par $Y_t' = \theta_t + e_t'$ l'estimation HT pour le mois courant (t) et par $Y_{t-1}' = \theta_{t-1} + e_{t-1}'$ l'estimation HT pour le mois précédent $(t-1)$. Nous allons utiliser le même modèle pour e_t' et e_{t-1}'. Les valeurs estimées de Corr(e_t', e_{t-1}') sont extrêmement élevées - en règle générale, 0.98 et plus. Bien que cette forte corrélation soit en partie artificielle (à cause des entreprises qui déclarent le même chiffre pour les ventes du mois courant et celles du mois précédent et à cause, peut-être, de la façon dont on impute les valeurs manquantes), il est difficile de faire la distinction entre les caractéristiques de e_t' et celles de e_{t-1}' lorsqu'on ne dispose pas d'autres renseignements.

Comme les trois panels avec renouvellement sont tirés de façon à peu près indépendante (Wolter 1979), les autocorrélations et les corrélations croisées pour (e_t', e_{t-1}') devraient être non nulles uniquement pour les décalages qui sont des multiples de 3. On peut faire la moyenne des valeurs estimées de ces corrélations avec décalage pour plusieurs périodes successives en supposant que la série est stationnaire en corrélation. Bien qu'en règle générale, on ne produise pas de valeurs estimées de corrélations avec décalage pour la Retail Trade Survey, une étude spéciale a permis de le faire au moyen de micro-données (totaux de groupes aléatoires) de l'échantillon de l'enquête pour la période allant de janvier 1973 à mars 1975. À cette époque toutefois, le plan de sondage prévoyait quatre panels avec renouvellement (à l'exclusion du panel aréolaire). Faute de données plus récentes, nous avons "fait la moyenne" des corrélations aux décalages 4, 8, 12, 16, 20 et 24 pour e_t' et e_{t-1}'. (En fait, nous avons commencé par appliquer la transformation de Fisher, $.5 \log((1 + r)/(1 - r))$, pour rendre la distribution des corrélations transformées plus symétrique, puis nous avons transformé à nouveau les résultats ainsi obtenus.) Les résultats pertinents figurent dans le tableau 1. Ce tableau révèle une corrélation positive assez forte entre les erreurs d'échantillonnage et l'existence d'un mouvement saisonnier au décalage 12. Compte tenu de ces données, nous pouvons proposer le modèle suivant:

$$(1 - \phi^m B^m)(1 - \Phi B^{12}) e_t' = v_{1t}, \tag{5.1}$$

où $m = 4$ pour l'enquête à quatre panels. Ce modèle vaut aussi pour e'_{t-1}, sauf que $v_{2,t-1}$ est substitué à v_{1t}. (v_{1t} et $v_{2,t-1}$ représentent le bruit blanc avec comme variance σ_v^2.)

Une caractéristique particulièrement utile du modèle (5.1) est que s'il peut décrire, mois après mois, l'erreur d'échantillonnage dans chaque panel avec $m = 1$, alors pour n'importe quel nombre m (qui est un diviseur de 12) de panels indépendants qui participent successivement à l'enquête, e'_t est conforme au modèle (5.1). Grâce à cette caractéristique, nous pouvons nous servir des résultats du tableau 1 pour l'enquête à quatre panels afin d'estimer ϕ^4 et Φ et (en supposant que $\phi > 0$) de transformer ces valeurs en valeurs estimées de ϕ^3 et Φ, qui sont les paramètres du modèle pour l'enquête à trois panels. Nous avons donc déterminé ϕ^4 et Φ pour minimiser la somme des carrés des écarts entre les corrélations calculées à l'aide de (5.1) et celles calculées à l'aide de la transformation de Fisher. (Nous avons exclu les décalages 20 et 24 du calcul et attribué un poids de 0.5 au décalage 16 à cause du faible nombre de corrélations estimées ayant servi au calcul d'une moyenne pour ces décalages.) Nous avons ainsi obtenu les valeurs estimées $\hat{\phi}^3 = .685$ et $\hat{\Phi} = .723$ pour les restaurants et $\hat{\phi}^3 = .664$ et $\hat{\Phi} = .714$ pour les débits de boissons. Les corrélations calculées à l'aide de (5.1) pour $m = 4$ figurent dans le tableau 1 et peuvent être comparées à celles qui ont servi au calcul de moyennes. Nous pourrions envisager des méthodes d'estimation statistique plus formelles pour ϕ^3 et Φ et imaginer un test de validité de l'ajustement du modèle s'il était possible d'établir des estimations d'autocorrélation des erreurs d'échantillonnage à l'aide de micro-données plus récentes provenant de l'enquête à trois panels.

Nous posons aussi comme hypothèse que $\text{Corr}(e'_t, e'_{t-1-k}) = \rho\,\text{Corr}(e'_t, e'_{t-k})$ pour tous k. Pour appuyer cette hypothèse, signalons que la régression de e'_{t-1-k} par rapport à e'_{t-k} pour la population s'exprime par $\rho e'_{t-k} + \epsilon$; si ϵ et e'_t ne sont pas non corrélées, ϵ est, à tout le moins, sûrement faible puisque $\text{Var}(\epsilon) = (1 - \rho^2)\text{Var}(e'_t)$ et que ρ est très voisin de 1. Compte tenu de cette hypothèse, nous pouvons définir le modèle bidimensionnel suivant pour (e'_t, e'_{t-1}) à partir de l'équation (5.1):

$$(1 - \phi^3 B^3)(1 - \Phi B^{12}) \begin{bmatrix} e'_t \\ e'_{t-1} \end{bmatrix} = \begin{bmatrix} v_{1t} \\ v_{2,t-1} \end{bmatrix} \quad \text{Var} \begin{bmatrix} v_{1t} \\ v_{2,t-1} \end{bmatrix} = \sigma_v^2 \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}, \tag{5.2}$$

où $\rho = \text{Corr}(v_{1t}, v_{2,t-1}) = \text{Corr}(e'_t, e'_{t-1})$. On produit régulièrement des estimations de $\text{Corr}(e'_t, e'_{t-1})$; d'ailleurs, nous disposions de ces estimations pour les années 1982 à 1986. En faisant la moyenne de ces valeurs estimées (à l'aide de la transformation de Fisher), nous avons obtenu $\hat{\rho} = .985$ pour les restaurants et $\hat{\rho} = .986$ pour les débits de boissons.

Nous pouvons maintenant nous servir de (5.2) pour élaborer un modèle d'erreur d'échantillonnage pour l'estimateur composite de forme linéaire (Wolter 1979), qui est défini

$$\begin{aligned} Y'_t{}'' &= (1 - \beta) Y'_t + \beta(Y'_{t-1}{}' + Y'_t - Y'_{t-1}) \quad \text{(estimateur provisoire)}, \\ Y_{t-1} &= (1 - \alpha) Y'_{t-1} + \alpha Y'_{t-1}{}'' \qquad\qquad \text{(estimateur final)}. \end{aligned} \tag{5.3}$$

Dans l'enquête à trois panels, $\alpha = .8$ et $\beta = .75$. On voit facilement que les équations ci-dessus s'appliquent aussi bien aux erreurs d'échantillonnage (e étant substitué à Y). Ceci nous permet d'exprimer e_t en fonction de e'_t et de e'_{t-1} par l'équation suivante:

$$(1 - .75B) e_t = .2\, e'_t{}'' - .75\, e'_{t-1} + .8\, e'_t. \tag{5.4}$$

En combinant (5.2) et (5.4), nous obtenons

$$(1 - .75B)(1 - \phi^3 B^3)(1 - \Phi B^{12})\, e_t = .2\, v_{2t} - .75\, v_{2,t-1} + .8\, v_{1t}. \tag{5.5}$$

Tableau 2

Coefficients de variation (CV)[1] pour les estimations de ventes au détail

	Horvitz-Thompson	Composite final[2]	Extraction de signal[3]	
	CV	CV	min	max
Restaurants	.042	.025	.017	.023
Débits de boissons	.088	.052	.032	.038

[1] CV = (variance relative)$^{.5}$.
[2] Les valeurs pour l'estimateur composite final sont déterminées à l'aide des modèles (5.7a) et (5.7b).
[3] En fait, les valeurs pour l'extraction de signal varient dans le temps, la valeur la plus élevée se trouvant à la fin de la série et la valeur la plus faible, vers le milieu. Nous donnons ici la valeur la plus élevée (décembre 1986) et la plus faible (janvier 1982) pour les deux séries. Les variances de l'extraction de signal ne sont pas chronologiquement symétriques parce que les "rééchantillonnage" de janvier 1982 ne coincide pas parfaitement avec le milieu de la série.

Le membre de droite de l'équation ci-dessus représente un processus de moyennes mobiles du premier ordre (Box et Jenkins, 1976, p. 121), dont on peut déterminer les paramètres en connaissant les valeurs estimées de σ_v^2 et ρ. L'équation (5.5) donnerait donc un modèle ARMA pour e_t.

Au lieu de poursuivre dans la même voie, nous allons maintenant poser l'hypothèse plutôt audacieuse qu'il existe un modèle du même type pour $\log(u_t)$ dans l'expression $\log(Y_t) = \log(\theta_t) + \log(u_t)$, par conséquent,

$$(1 - .75B)(1 - \phi^3 B^3)(1 - \Phi B^{12}) \log(u_t) = (1 - \eta B)\, c_t. \tag{5.6}$$

Nous faisons cela parce que les estimations de la variance d'échantillonnage pour ces séries dépendent fortement du niveau de la série; les estimations de la variance relative, elles, sont beaucoup plus stables. Nous supposons de plus que les valeurs estimées de la variance relative et de ρ peuvent servir à calculer η et σ_c^2. Les valeurs estimées Y_t', Y_{t-1}', $\widehat{\text{Var}}(e_t')$ et $\widehat{\text{Var}}(e_{t-1}')$ sont connues pour les années 1982 à 1986 inclusivement. Une fois les estimations de la variance relative calculées, elles ont servi à un processus d'estimation s'inspirant de la méthode du maximum de vraisemblance appliquée à la distribution normale logarithmique - c.-à-d. calculer la moyenne des logarithmes des estimations de la variance relative, additionner cette moyenne et la moitié de la variance empirique des estimations de forme logarithmique, puis prendre la valeur exponentielle de cette somme. (On aurait obtenu des résultats semblables en faisant simplement la moyenne des estimations de la variance relative.) Var rel(Y_t') et Var rel(Y_{t-1}') ont été calculées de façon indépendante; on a ensuite fait la moyenne des deux pour obtenir une estimation unique de la variance relative qui est constante. Les résultats pertinents figurent dans la première colonne du tableau 2. Nous pouvons maintenant nous servir de ces résultats et des valeurs $\hat{\rho}$ calculées précédemment pour déterminer η et σ_c^2 dans le membre de droite de l'équation (5.6). Les modèles d'erreur d'échantillonnage sont donc

$$(1 - .75B)(1 - .685B^3)(1 - .723B^{12}) \log(u_t) = (1 + .130B)c_t \tag{5.7a}$$

$$\text{(restaurants)} \qquad \hat{\sigma}_c^2 = 1.948 \times 10^{-5}$$

$$(1 - .75B)(1 - .664B^3)(1 - .714B^{12})\log(u_t) = (1 + .134B)c_t \tag{5.7b}$$

$$\text{(débits de boissons)} \qquad \hat{\sigma}_c^2 = 9.301 \times 10^{-5}.$$

On peut employer la méthode de McLeod (1975,1977) pour déterminer Var(log(u_t)) dans ces modèles; Var(log(u_t)) est une estimation de la variance relative de l'estimateur composite final. Le tableau 2 donne les coefficients de variation correspondants (.025 pour les restaurants et .052 pour les débits de boissons). Ces valeurs sont très proches des estimations qui figurent dans les rapports mensuels sur le commerce de détail du Census Bureau et qui sont obtenues d'une manière plus directe.

5.2 Modélisation de séries chronologiques et extraction de signal

Les figures 1a et 1b montrent le graphique de la série des estimations composites finales Y_t pour les restaurants et les débits de boissons respectivement. Afin d'élaborer des modèles pour θ_t, nous allons commencer par modéliser directement la série Y_t. Les deux séries des ventes au détail reflètent des tendances ainsi qu'un fort mouvement saisonnier, l'amplitude des fluctuations saisonnières augmentant avec le niveau de la série. Cela fait entrevoir l'opportunité de recourir au calcul logarithmique et la nécessité d'appliquer des différences aux séries, ce qui est très courant dans le cas des séries économiques. Après un examen des autocorrélations empiriques pour log(Y_t) et ses différences, nous avons pensé que l'opérateur de différence $(1 - B)(1 - B^{12})$ serait approprié pour les deux séries. On sait que les séries touchant le commerce de détail sont influencées par la variation des jours commerciaux; il est possible de modéliser cette variation en introduisant dans le modèle sept variables de régression: X_{1t} = nombre de lundis dans le mois t, ..., X_{7t} = nombre de dimanches dans le mois t. D'après Bell et Hillmer (1983), il serait plus commode d'utiliser les variables $T_{1t} = X_{1t} - X_{7t}$ (nombre de lundis − nombre de dimanches), ..., $T_{6t} = X_{6t} - X_{7t}$ (nombre de samedis − nombre de dimanches), $T_{7t} = \sum_1^7 X_{it}$ (longueur du mois t). Afin de définir les structures ARMA, nous avons analysé les autocorrélations et les autocorrélations partielles des résidus de la régression de $(1 - B)(1 - B^{12})\log(Y_t)$ par rapport à $(1 - B)(1 - B^{12})T_{it}$, $i = 1, \ldots, 7$. Cette analyse nous a amenés à définir un modèle ARMMI $(0,1,2)(0,1,1)_{12}$ pour les restaurants et un modèle ARMMI $(0,1,3)(0,1,1)_{12}$ pour les débits de boissons. Nous avons donc les modèles estimés suivants:

$$(1 - B)(1 - B^{12})\left[\log(Y_t) - \sum_i \beta_i T_{it}\right] = (1 - .25B - .22B^2)(1 - .79B^{12})\, a_t$$

(restaurants) $\qquad \hat{\sigma}_a^2 = .000230$ (5.8a)

$$(1 - B)(1 - B^{12})\left[\log(Y_t) - \sum_i \beta_i T_{it}\right] = (1 - .21B - .15B^2 + .03B^3)(1 - .56B^{12})\, a_t$$

(débits de boissons) $\qquad \hat{\sigma}_a^2 = .000587.$ (5.8b)

Pour des raisons de concision, nous allons omettre les valeurs estimées des paramètres de la variation des jours commerciaux. Bien que les paramètres de moyenne mobile pour les décalages 2 et 3 dans l'équation (5.8b) soient peu élevés, nous allons les conserver car les équations (5.8a) et (5.8b) ne vont servir qu'à la première étape de la modélisation de log(θ_t) pour les deux séries.

En prenant des modèles du type (5.8a) et (5.8b) pour log(θ_t) et les modèles (5.7a) et (5.7b) pour log(u_t), nous avons pu estimer les paramètres des modèles pour log(θ_t). La valeur estimée des paramètres de moyenne mobile saisonnière est très proche de 1 pour les deux séries (.985 pour les restaurants et .992 pour les débits de boissons), ce qui suppose un mouvement saisonnier quasi déterministe, que l'on peut modéliser en supprimant le facteur $(1 - B^{12})$ de part et d'autre de l'équation du modèle de θ_t et en lui substituant une constante de tendance

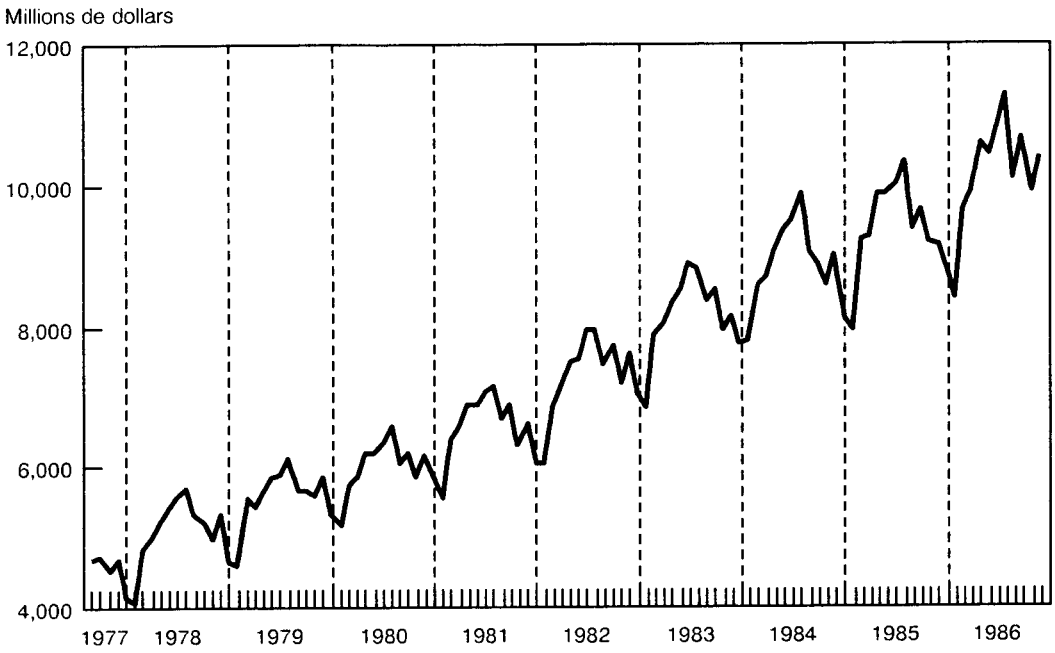

Figure 1.a Ventes au détail des restaurants – estimations composites (non étalonnées)

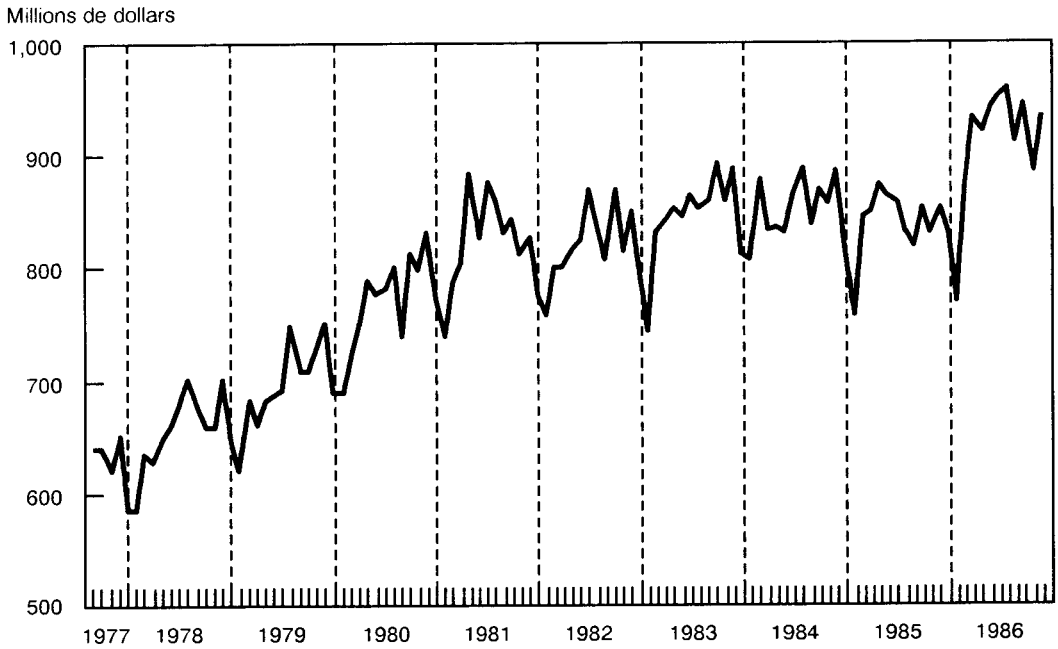

Figure 1.b Ventes au détail des débits de boissons – estimations composites (non étalonnées)

ainsi qu'une fonction de régression saisonnière de la forme $\sum_1^{11} \gamma_i M_{it}$, où M_{1t} est 1 en janvier, -1 en décembre et 0 aux autres mois, ..., M_{11t} est 1 en novembre, -1 en décembre et 0 aux autres mois, . . ., M_{11t} est 1 en novembre, -1 en décembre et 0 aux autres mois (Bell 1987). On obtient ainsi les modèles estimés suivants:

$$(1 - B)\left[\log(\theta_t) - \sum_i \hat{\beta}_i T_{it} - \sum_i \hat{\gamma}_i M_{it}\right] = .00762 + (1 - .20B - .29B^2)b_t$$

(restaurants) $\qquad \hat{\sigma}_b^2 = .000139$ (5.9a)

$$(1 - B)\left[\log(\theta_t) - \sum_i \hat{\beta}_i T_{it} - \sum_i \hat{\gamma}_i M_{it}\right] = .00352 + (1 - .18B - .09B^2 - .42B^3)b_t$$

(débits de boissons) $\qquad \hat{\sigma}_b^2 = .000244.$ (5.9b)

Là encore, nous omettons les valeurs estimées des paramètres de régression. Nous ne calculons pas non plus d'erreurs types pour les paramètres ARMA; compte tenu des modèles que nous utilisons ici, ce genre de calcul mérite de faire l'objet de recherches. Toutefois, l'hypothèse peu réaliste selon laquelle le modèle d'erreurs d'échantillonnage est connu rend la tâche particulièrement difficile dans notre contexte. En examinant les résidus normalisés obtenus au moyen du filtre de Kalman, ainsi que les autocorrélations correspondantes, on ne trouve aucune déficience majeure dans les modèles ajustés pour l'une ou l'autre série.

Nous nous sommes servis des modèles estimés (5.7a,b) et (5.9a,b) pour établir des estimations d'extraction de signal de $\log(\theta_t)$, que nous avons ensuite retransformées pour obtenir les valeurs estimées de θ_t. Les résultats pertinents pour les deux séries sont reproduits dans les figures 2a et 2b, abstraction faite des effets saisonniers estimés et des effets estimés de la variation des jours commerciaux. Notons que l'extraction de signal produit des estimations très peu différentes des estimations composites pour ce qui est des restaurants, l'erreur d'échantillonnage étant d'ailleurs assez faible dans leur cas (variance relative faible); en revanche, la même extraction produit des estimations très différentes des estimations composites dans le cas des débits de boissons, pour lesquels l'erreur d'échantillonnage est beaucoup plus forte (variance relative plus élevée). Nous avons aussi calculé des variances d'extraction de signal pour $\log(\theta_t)$; il s'agit des variances relatives des valeurs estimées de θ_t. Le tableau 2 montre que, dans le cas des restaurants, l'extraction de signal produit des CV de 8 à 32% moins élevés – selon le mois auquel correspond l'estimation dans la série – que ceux des estimateurs composites finals (qui sont déjà peu élevés) et que dans le cas des débits de boissons, elle produit des CV de 27 à 38% moins élevés. Comme nous l'avons déjà mentionné, ces résultats sont optimistes car ils supposent que les véritables modèles de composantes sont ceux qui ont été estimés. Afin de voir en partie de quoi il retourne, nous allons maintenant étudier la sensibilité des résultats pour les débits de boissons par rapport à une variation des paramètres des modèles.

5.3 Analyse de sensibilité pour les ventes des débits de boissons

Dans cette section, nous nous penchons sur la sensibilité des résultats par rapport à des variations dans le modèle d'erreurs d'échantillonnage, étant donné que celui-ci a été défini avec moins de renseignements que le modèle du signal. Notre méthode consiste à faire varier des paramètres du modèle d'erreurs d'échantillonnage, puis à réestimer le modèle du signal et à refaire l'extraction de signal. Bien qu'il serait préférable de disposer de mesures statistiques plus formelles pour rendre compte de l'erreur d'extraction de signal attribuable à l'erreur de modèle (ce qui n'est pas possible dans l'état actuel de la théorie et des logiciels), cette méthode devrait à tout le moins indiquer sous quels rapports les résultats de l'extraction de signal sont sensibles à une variation de paramètres et sous quels rapports ils ne le sont pas.

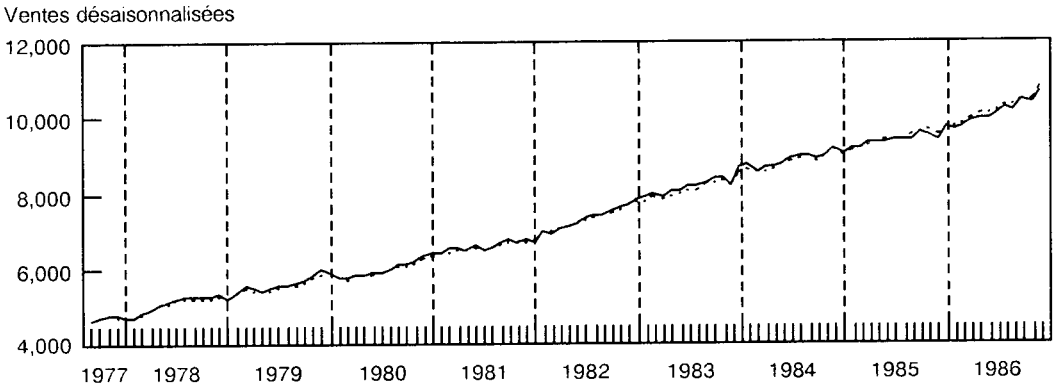

Figure 2.a Restaurants: estimations composites (continu) et estimations de l'extraction de signal (pointillé)

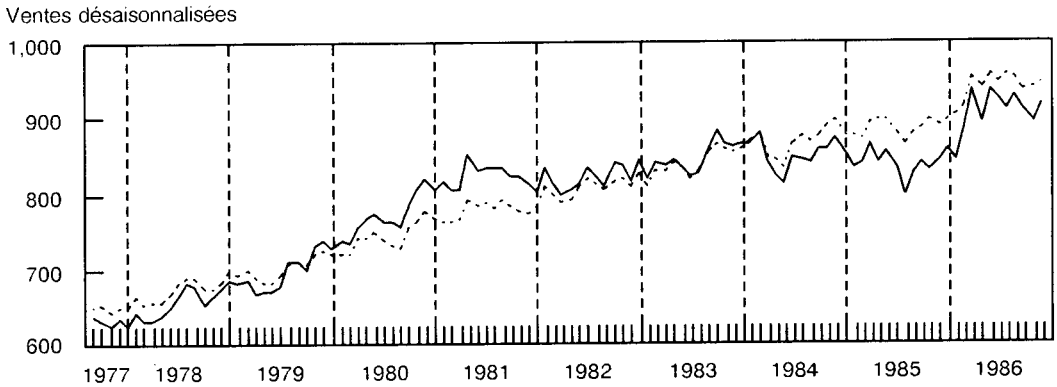

Figure 2.b Débits de boissons: estimations composites (continu) et estimations de l'extraction de signal (pointillé)

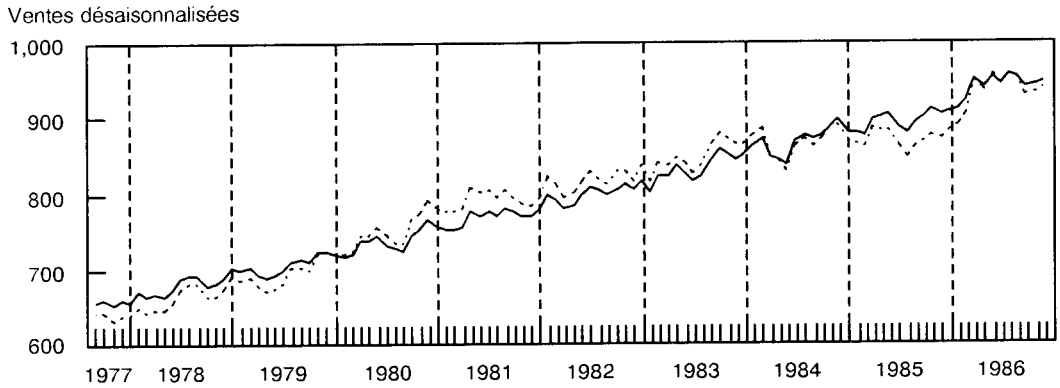

Figure 2.c Débits de boissons: autre série d'estimations de l'extraction de signal

En comparant les modèles (5.8b) et (5.9b), nous pouvons avoir une idée de la sensibilité du modèle du signal par rapport à une variation de σ_c^2, la variance de résidus du modèle d'erreurs d'échantillonnage, puisque (5.8b) correspond à $\sigma_c^2 = 0$ et (5.9b), à $\sigma_c^2 = 9.3 \times 10^{-5}$. Nous observons les variations les plus notables dans la valeur estimée de σ_b^2, ce qui n'est pas étonnant, et dans la valeur estimée du paramètre de moyenne mobile saisonnière, η_{12} par exemple, qui était essentiellement égal à 1 lorsqu'on a formulé le modèle (5.9b). La réestimation du modèle du signal pour d'autres valeurs de σ_c^2 a donné $\hat{\eta}_{12} \geq .99$ pour $\sigma_c^2 \geq 3.0 \times 10^{-5}$. Compte tenu de cela, et pour simplifier la présentation des résultats, nous supposons que $\eta_{12} = 1$ et utilisons un modèle du signal qui, comme (5.9b), comprend des variables indicatrices saisonnières.

La figure 2c donne les estimations d'extraction de signal $\hat{\theta}_t$ (désaisonnalisées et corrigées en fonction de la variation des jours commerciaux) qui correspondent aux modèles d'erreurs d'échantillonnage pour lesquels $(\phi^3,\Phi) = (.564,.614)$ et $(.764,.814)$ et où $\rho = .986$ et $\text{Var}(\log(u_t)) = .00776$ (la variance relative des estimations d'Horvitz-Thompson) sont fixes. Ces estimations comprennent les valeurs extrêmes de $\hat{\theta}_t$ pour l'analyse de sensibilité. La nature des diverses estimations $\hat{\theta}_t$ que nous avons produites semble correspondre à peu près à la valeur de $\text{CV}_{56} = [\text{Var}(\log(\hat{\theta}_{56}) - \log(\theta_{56})]^{1/2}$, qui est le coefficient de variation de l'extraction de signal observé au milieu de la série. (CV_{56} est très proche de la valeur minimum, qui est observée à $t = 53$ - voir tableau 2.) Moins CV_{56} est élevé, plus $\hat{\theta}_t$ est lisse. CV_{56} est 2.78%, 3.28% et 3.70% pour des valeurs (ϕ^3,Φ) de (.564,.614), (.664,.714) et (.764,.814) respectivement. D'autres estimations $\hat{\theta}_t$ que nous avons produites sont encore plus près des estimations de l'extraction de signal des figures 2b ou 2c, l'écart le plus mince étant observé à CV_{56}.

Nous allons maintenant étudier la sensibilité de CV_{56} par rapport à une variation des paramètres du modèle d'erreurs d'échantillonnage, à commencer par ρ. Le seul paramètre de (5.7b) qui soit touché par une variation de ρ est η. Le tableau 3 donne les valeurs de η considérées et les valeurs de ρ correspondantes ainsi que les valeurs CV_{56} qui en découlent. Nous constatons que CV_{56} est assez sensible aux variations de ρ, plus particulièrement aux hausses: CV_{56} est 6% plus élevé pour $\rho = 1$ (3.49) que pour $\rho = .985$ (3.28), qui est la valeur utilisée pour (5.7b).

Tableau 3

Sensibilité de CV_{56}[1] par rapport à une variation de η (variation de ρ) pour les débits de boissons

η	.00	−.05	−.10	−.15	−.20	−.25
ρ	.9375	.9642	.9792	.9888	.9953	1.000
CV_{56}	3.03	3.12	3.21	3.31	3.40	3.49

[1] CV_{56} est le coefficient de variation de l'extraction de signal pour $t = 56$ (milieu de la série), exprimé en pourcentage, c.-à.-d. racine carrée de $\text{Var}(\log(\hat{\theta}_t) - \log(\theta_t))$ multipliée par 100.

Tableau 4

Sensibilité de CV_{56} par rapport à une variation de $\text{Var}(\log(u_t))$[1] (variation de σ_c^2) pour les débits de boissons

$\text{Var}(\log(u_t))$	.00676	.00726	.00776	.00826	.00876
CV(HT)[2]	8.22	8.52	8.81	9.09	9.36
$\sigma_c^2 \times 10^5$	8.16	8.76	9.30	9.97	10.57
CV_{56}	3.16	3.23	3.28	3.35	3.40

[1] $\text{Var}(\log(u_t))$ est la variance relative des estimateurs d'Horvitz-Thompson.

[2] CV(HT) est le coefficient de variation des estimateurs d'Horvitz-Thompson, exprimé en pourcentage, c.-à.-d. racine carrée de $\text{Var}(\log(u_t))$ multipliée par 100.

Tableau 5

Sensibilité des résultats par rapport à une variation de (ϕ^3,Φ) pour les débits de boissons

		(i) Valeurs de $\sigma_c^2 \times 10^5$ pour une paire de valeurs (ϕ^3,Φ) donnée				
		ϕ^3				
		.564	.614	.664	.714	.764
	.614	16.90	14.70	12.36	9.98	7.64
	.664	15.03	13.00	10.87	8.72	6.62
Φ	.714	13.04	11.23	9.30	7.44	5.60
	.764	10.96	9.40	7.78	6.15	4.58
	.814	8.79	7.51	6.17	4.85	3.58
		(ii) Valeurs de CV_{56} pour une paire de valeurs (ϕ^3,Φ) donnée				
		ϕ^3				
		.564	.614	.664	.714	.764
	.614	2.78	2.88	2.99	3.12	3.27
	.664	2.95	3.04	3.14	3.26	3.38
Φ	.714	3.10	3.19	3.28	3.39	3.50
	.764	3.24	3.33	3.42	3.51	3.60
	.814	3.36	3.45	3.54	3.62	3.70

Nous étudions ensuite la sensibilité de CV_{56} par rapport à une variation de $\text{Var}(\log(u_t))$. Le seul paramètre du modèle d'erreurs d'échantillonnage qui peut être touché par une telle variation est σ_c^2. Le tableau 4 donne les valeurs de $\text{Var}(\log(u_t))$, et de sa racine carrée (CV(HT)) ainsi que les valeurs correspondantes de σ_c^2 et les valeurs CV_{56} qui en découlent. CV_{56} est moins sensible aux variations que dans le tableau 3.

Enfin, nous examinons la sensibilité de CV_{56} par rapport à ϕ^3 et à Φ. Si nous posons $\text{Var}(\log(u_t))$ égale à .00776 et que nous modifions (ϕ^3,Φ), σ_c^2 variera aussi. Le tableau 5 donne les séries de valeurs utilisées pour (ϕ^3,Φ) ainsi que les valeurs de σ_c^2 et de CV_{56} qui en découlent. Notons que σ_c^2 est plus sensible ici que dans le tableau 4 et que CV_{56} augmente sensiblement lorsque ϕ^3 et Φ augmentent.

En conclusion, nous pouvons dire qu'une variation modérée des paramètres du modèle d'erreurs d'échantillonnage influe relativement peu sur $\hat{\theta}_t$ – les plus fortes variations observées chez $\hat{\theta}_t$ étant de l'ordre de 2% – mais a des effets plus notables sur les variances de l'extraction de signal, la variation des valeurs de CV_{56} pouvant aller jusqu'à 17%. Cela porte à croire qu'en ce qui concerne cet exemple, la chose dont il faut se préoccuper le plus lorsqu'on ne connaît pas les paramètres du modèle d'erreurs d'échantillonnage est probablement l'effet que cela peut avoir sur la variance de l'extraction de signal de même que les mesures permettant d'évaluer la différence de qualité par rapport aux estimations composites. Néanmoins, pour tous les cas étudiés dans l'analyse de sensibilité, nous avons observé une amélioration notable de la variance dans le cas des estimations d'extraction de signal.

5.4 Conclusions

L'exemple des débits de boissons illustre bien les avantages que l'on peut tirer de l'application de l'approche chronologique à l'estimation dans les enquêtes à passages répétés. Les deux exemples donnent aussi un aperçu du caractère délicat et complexe de la modélisation de séries chronologiques que l'on peut être appelé à faire. Nous considérons ces résultats comme provisoires pour plusieurs raisons. Premièrement, nous avons déjà souligné le caractère "optimiste" des variances d'extraction de signal, qui ne tiennent pas compte de l'erreur

d'estimation des paramètres. Deuxièmement, nous ne savons pas exactement pourquoi les estimations de l'extraction de signal se maintiennent au-dessus ou au-dessous des estimations composites pendant de longues périodes. (Cela se voit clairement dans la figure 2b, et aussi dans la figure 2a.) Dans le cas des débits de boissons, ce phénomène ressortait constamment dans l'analyse de sensibilité et ne semble donc pas attribuable à une variabilité des paramètres du modèle d'erreurs d'échantillonnage. Nous nous attachons présentement à découvrir si ce phénomène ne serait pas imputable à une mauvaise spécification du modèle d'erreurs d'échantillonnage ou du modèle du signal. En fait, Bell et Wilcox (1990) affirment que les corrélations de e'_t et e'_{t-1} aux décalages qui ne sont pas des multiples de 3 ne sont pas nécessairement nulles comme le supposait le modèle.

REMERCIEMENTS

Cet article s'inspire en partie d'un travail de recherche subventionné par la National Science Foundation sous le numéro SES 84-01460 (''On-Site Research to Improve the Government-Generated Social Science Data Base'') et par le U.S. Bureau of the Census et l'Université du Kansas en vertu des conventions sur la statistique n° 87-14 et 88-27. Les opinions, conclusions ou recommandations exprimées dans cet article sont celles des auteurs et ne reflètent pas nécessairement la position de la National Science Foundation, du Census Bureau ou de l'Université du Kansas. Nous tenons à remercier Phillip Kott et David Findley pour leurs précieux commentaires concernant les résultats relatifs à la convergence, Abdelwahed Trabelsi pour le soutien qu'il nous a offert à titre d'associé de l'ASA/NSF/Census Research, de même qu'un arbitre anonyme qui, par ses commentaires, a contribué à améliorer sensiblement cet article. Nous adressons aussi des remerciements à Ruth Detlefsen, Michael Shimberg et Carol Veum, de la Business Division du Census Bureau, pour nous avoir fourni des données de la Retail Trade Survey et nous avoir renseigné sur cette enquête. Enfin, nous tenons à remercier de façon particulière James Bozik, Mark Otto et Marian Pugh, qui ont consacré beaucoup d'effort à l'élaboration du logiciel qui a servi à cette étude. Toute erreur qui pourrait s'être glissée dans l'analyse des exemples est la responsabilité des auteurs.

BIBLIOGRAPHIE

BELL, W.R. (1984). Signal extraction for nonstationary time series. *Annals of Statistics*, 12, 646-664.

BELL, W.R. (1987). A Note on overdifferencing and the equivalence of seasonal time series models with monthly means and models with (0,1,1)12 seasonal parts when $\Theta = 1$. *Journal of Business and Economic Statistics*, 5, 383-387.

BELL, W.R., et HILLMER, S.C. (1983). Modeling time series with calendar variation. *Journal of the American Statistical Association*, 78, 526-534.

BELL, W.R., et HILLMER, S.C. (1989). Modeling time series subject to sampling error. Research Report 89/01, Statistical Research Division, Bureau of the Census.

BELL, W.R., et HILLMER, S.C. (1990). A Matrix approach to signal extraction for nonstationary time series models. Soumis pour publication.

BELL, W.R., et WILCOX, D.W. (1990). The effect of sampling error on the time series behavior of consumption data. Paper presented at the CRDE/Journal of Econometrics Conference on Seasonality in Econometric Models, Montréal, Canada, mai 1990.

BINDER, D.A., et DICK, J.P. (1986). Modelling and estimation for repeated surveys. Statistique Canada Rapport technique. Division des méthodes d'enquêtes sociales.

BINDER, D.A., et DICK, J.P. (1989). Implications of survey designs for estimating seasonal ARIMA models. Article présenté à la réunion annuel de la American Statistical Association, Washington, D.C.

BOX, G.E.P., et JENKINS, G.M. (1976). *Time Series Analysis: Forecasting Control*. San Francisco: Holden Day.

CHUNG, K.L. (1968). *A Course in Probability Theory*. New York: Harcourt, Brace World, Inc.

ELTINGE, J.L., et FULLER, W.A. (1989). Time series random component models for sample surveys. Article présenté à la réunion d'hiver de la American Statistical Association, San Diego, CA.

FULLER, W.A. (1976). *Introduction to Statistical Time Series*. New York: Wiley.

FULLER, W.A., et ISAKI, C.T. (1981). Survey design under superpopulation models. *Current Topics in Survey Sampling*, (éds. D. Krewski, R. Platek, et J.N.K. Rao). New York: Academic Press, 199-226.

GARRETT, J.K., DETLEFSEN, R.E., et VEUM, C.S. (1987). Recent sample revisions and related enhancements for business surveys of the U.S. Bureau of the Census. *Proceedings of the Business and Economic Statistics Section, American Statistical Association*, 141-149.

GRANGER, C.W.J., et NEWBOLD, P. (1976). Forecasting transformed series. *Journal of the Royal Statistical Society*, Séries B, 38, 189-203.

HAUSMAN, J.A., et WATSON, M.W. (1985). Errors in variables and seasonal adjustment procedures. *Journal of the American Statistical Association*, 80, 531-540.

HILLMER, S.C., et TRABELSI, A. (1987). Benchmarking of economic time series. *Journal of the American Statistical Association*, 82, 1064-1071.

ISAKI, C.T., WOLTER, K.M., STURDEVANT, T.R., MONSOUR, N.J., et TRAGER, M.L. (1976). Sample redesign of the Census Bureau's monthly business surveys. *Proceedings of the Business and Economic Statistics Section, American Statistical Association*, 90-98.

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

McLEOD, I. (1975). Derivation of the theoretical autocovariance function of autoregressive-moving average time series. *Applied Statistics*, 24, 255-256.

McLEOD, I. (1977). Correction to derivation of the theoretical autocovariance function of autoregressive-moving average time series. *Applied Statistics*, 26,194.

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

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

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.

RAO, J.N.K., SRINATH, K.P., et QUENNEVILLE, B. (1989). Optimal estimation of level and change using current preliminary data. *Panel Surveys*, (éds. Daniel Kasprzyk, Greg Duncan, Graham Kalton et M.P. Singh). New York: Wiley, 457-479.

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. *Revue Internationale de Statistique*, 45, 13-28.

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

TAM, S.M. (1987). Analysis of repeated surveys using a dynamic linear model. *Revue Internationale de Statistique*, 55, 63-73.

TRABELSI, A., et HILLMER, S.C. (1990). Benchmarking time series with reliable benchmarks. *Applied Statistics*, 39, à paraître.

WOLTER, K.M. (1979). Composite estimation in finite populations. *Journal of the American Statistical Association*, 74, 604-613.

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

WOLTER, K.M., ISAKI, C.T., STURDEVANT, T.R., MONSOUR, N.J., et MAYES, F.M. (1976). Sample selection and estimation aspects of the Census Bureau's monthly business surveys. *Proceedings of the Business and Economic Statistics Section, American Statistical Association*, 99-109.

WOLTER, K.M., et MONSOUR, N.J. (1981). On the problem of variance estimation for a deseasonalized series. *Current Topics in Survey Sampling*, (éds. D. Krewski, R. Platek et J.N.K. Rao). New York: Academic Press, 199-226.