Analyse de l'évolution d'un paramètre dans les enquêtes répétées

Patrice BERTAIL,
Pierre COMBRIS

Cahiers d'économie et sociologie rurales, n° 48, 1998

Patrice BERTAIL*, Pierre COMBRIS*

Dynamic Analysis in Repeated Survey Sampling

Summary — This article presents a survey on the analysis and estimation of the dynamics of a parameter in repeated survey sampling, in particular in panels (with rotation or not). We first present the classical and model based approaches. We compute the exact form of the Horvitz-Thompson estimators for a two-period comparison in the classical approach and give the main tools for a model based analysis of the dynamic of a parameter of interest (in our case the consumption mean). We show that state-space models are perfectly adapted to take into account the dynamic of the sampling scheme and the proper dynamic of the parameters of interest. We illustrate this approach on the Secodip consumption rotating panel and study more particularly the consumption of oil. In particular we emphasize some tests to detect transient and permanent changes.

Key-words:
survey sampling, dynamic analysis, Horvitz Thompson estimator, times series, Kalman filter, consumption analysis

Analyse de l'évolution d'un paramète dans les enquêtes répétées

Résumé : Cet article a pour but de faire le point sur l'analyse et l'estimation de la dynamique de paramètres dans des enquêtes, en particulier dans les panels (avec ou sans rotations). Nous abordons rapidement l'approche classique et l'approche modèle. Nous donnons la forme exacte des estimateurs de Horvitz-Thompson pour une comparaison période par période, ainsi que les principaux outils de l'approche modèle. Nous montrons que, dans de très nombreuses situations, les modèles espace-état sont parfaitement adaptés à la prise en compte de l'erreur d'échantillonnage et de la dynamique des paramètres. Ils permettent en particulier d'étudier la dynamique d'un paramètre sur longue période. En guise d'illustration des méthodes évoquées, une modélisation sous forme espace-état du panel de consommation Secodip est proposée et appliquée à la consommation d'huile d'olive.

Mots-clés:
sondages, dynamique de paramètres, enquêtes répétées, estimateurs de Horvitz-Thompson, séries temporelles, filtre de Kalman, consommation

* *INRA, CORELA, Laboratoire de recherche sur la consommation, 65, boulevard de Brandebourg, 94205 Ivry-sur-Seine.*
e-mail: bertail@ivry.inra.fr ; combris@ivry.inra.fr

DE plus en plus couramment les organismes de sondage effectuent des enquêtes dans le temps. Cependant, du fait de la répétition des enquêtes, les méthodes traditionnelles de théorie des sondages ne peuvent s'appliquer directement lorsque l'on cherche à étudier la dynamique des paramètres.

Il convient de rappeler en préambule, qu'en théorie des sondages l'aléa provient uniquement de l'échantillonnage, c'est-à-dire de la manière dont est sélectionné un sous-échantillon dans une population (forcément) dénombrable. Par exemple la moyenne «empirique» de la consommation d'un produit sur une population totale de taille N, à $N^{-1} \sum C_i$, les quantités C_i étant supposées déterministes, est un paramètre et non une statistique, que l'on désire estimer à partir de n observations de la consommation sur un sous-échantillon. Deux types de problèmes apparaissent dans la dimension temporelle :

– **L'échantillonnage peut changer au cours du temps**. Les enquêtes peuvent par exemple être effectuées à intervalles irréguliers sur des populations différentes, la corrélation entre les erreurs d'échantillonnage à plusieurs dates peut être forte etc. Ce problème pose essentiellement la question pratique du choix «optimal», en un certain sens, de la procédure statistique adaptée à l'échantillonnage temporel : on parle alors d'approche classique dans le sens ou le principe du déterminisme des paramètres continue d'être accepté, même s'il est possible de les comparer d'une période à l'autre et d'en analyser la dynamique. Se pose alors essentiellement le problème de savoir comment sont construits les échantillons successifs, en termes statistiques, quelle est la loi jointe des plans de sondages.

– **L'échantillonnage et les caractéristiques des individus, *i.e.* les paramètres, peuvent changer au cours du temps** selon des principes réguliers. Une telle observation remet en cause la notion de déterminisme inhérente à la théorie des sondages. L'idée qu'il puisse exister un lien de causalité entre les paramètres aux différentes dates conduit à considérer les paramètres eux-mêmes comme des variables aléatoires liées entre elles par un (sur-) modèle de type causal : on parle alors d'approche modèle.

Dans cet article, nous nous intéresserons essentiellement à l'analyse de l'évolution d'un paramètre sur plusieurs périodes en mettant en évidence l'apport des méthodes classiques et de l'approche modèle. Cette analyse est en effet très différente selon que l'on considère les paramètres à chaque période comme des constantes (approche classique) ou que l'on admet l'existence d'un surmodèle sur les paramètres, qui en décrit la dynamique (approche modèle). Pour illustrer ce point, considérons par exemple le cas d'un produit dont la consommation moyenne estimée est $C_t = 0{,}5$ à la date t et $C_{t+1} = 0{,}5$, à la date $t + 1$: y a-t-il eu une aug-

mentation significative (ou un choc) de la consommation entre ces deux dates ? Cette question est loin d'être aussi simple qu'il n'y paraît. Si l'on considère les consommations aux dates t et $t + 1$ comme des caractéristiques du produit sur la population sans lien temporel, on conclura que la différence est non significative (puisque nulle), mais si l'on est conscient qu'il existe une dynamique propre à la consommation (par exemple à travers l'existence d'une saisonnalité), on pourra conclure à la présence d'un choc (par exemple, si compte tenu des observations passées, on s'attendait à avoir $C_{t+1} = 0$). Nous verrons dans cette optique comment l'approche modèle permet d'améliorer la précision d'un estimateur en utilisant l'information recueillie aux périodes précédentes.

L'analyse statistique qui découle de ces deux approches dépend bien entendu de la nature des données dont on dispose et plus exactement du plan de sondage utilisé. On en distingue en général trois types (voir aussi Dussaix, 1987b) qui, par combinaison, peuvent donner lieu à des formes de sondage plus complexes. Nous illustrerons ces situations par quelques exemples liés à l'analyse de la consommation des ménages. On trouvera dans Duncan et Kalton (1987) un panorama détaillé de leurs avantages et de leurs inconvénients respectifs selon les objectifs poursuivis.

– **Les enquêtes par panel** : l'observation d'un même ensemble d'individus sur une longue période permet d'avoir des informations précises sur les comportements individuels. Les panels sont donc parfaitement adaptés à une approche microéconomique de phénomènes présentant une forte hétérogénéité, dans la mesure où ils permettent d'étudier la dynamique des distributions et a fortiori des paramètres de ces distributions. Le problème essentiel des panels est lié à la question de la représentativité. Deux problèmes principaux se posent : le changement de structure de la population échantillonnée elle-même au cours du temps et les biais de sélection liés soit au plan de sondage (l'erreur d'échantillonnage se reproduit alors au cours du temps), soit à des phénomènes de censure (non-réponses ou perte de certains individus), voire des biais d'autosélection liés au phénomène de conditionnement ou de lassitude. Nous n'aborderons pas ce type de problèmes liés à l'évaluation des biais dans la suite (voir par exemple Kalton, Kasprzyk et McMillen, 1989).

– **Les enquêtes répétées dans le temps** sur échantillons indépendants : ce type d'enquête permet d'appréhender le comportement macroéconomique, *i.e.* le comportement moyen des variables. Il est parfois possible de reconstituer des sous-populations homogènes et donc des points moyens dont on étudie alors le comportement comme s'il s'agissait d'un vrai panel : ce suivi de « cohortes » a récemment donné lieu à une abondante littérature statistique (les problèmes, qui apparaissent alors, combinent dans une certaine mesure les inconvénients des panels et l'aspect moyen des phénomènes observés). Par ailleurs, ce type de données n'échappe pas aux problèmes de censure liés à la nature des zéros

observés qui nécessitent une modélisation particulière (voir par exemple le problème des infréquences d'achat dans l'enquête consommation des ménages de l'INSEE réalisée à période fixe, pendant une semaine, sur des échantillons indépendants). Bien que ce type d'enquête soit plus adapté à une approche classique, nous verrons comment il est possible de tenir compte de la dimension temporelle.

– **Les enquêtes sur échantillons tournants ou panels avec rotations**: ce type d'enquête proche du panel, dans lequel une partie de l'échantillon est abandonnée et remplacée à certaines périodes, permet d'éliminer certains des inconvénient des panels (biais d'autosélection, censure par abandon) par des renouvellements plus ou moins fréquents des individus (rotation). Dans la mise au point de telles enquêtes se posent les problèmes de la détermination du taux de renouvellement en fonction des objectifs choisis et du choix des estimateurs. Dans une optique temporelle, on sera plus intéressé par le suivi d'individus présents dans toutes les enquêtes, ce qui n'est possible que sur des courtes périodes si la rotation est rapide (c'est le cas par exemple de la rotation à deux niveaux où un même individu ne participe qu'à deux enquêtes successives). Ce type de renouvellement permet d'appréhender l'évolution des paramètres entre deux périodes mais se révèle inadapté pour une analyse de longue durée, si l'on n'a aucune information préalable sur le processus économique sous-jacent. Les données hebdomadaires sur la consommation du panel Secodip avec renouvellement annuel du quart du panel permettent un bon arbitrage entre durée d'observation des individus et taux de rotation.

Dans la première partie, nous donnons les éléments essentiels de l'analyse statistique des enquêtes dans l'approche classique. Nous verrons qu'elle permet essentiellement de tenir compte de la dynamique de l'erreur d'échantillonnage. On se référera aussi aux surveys de Cochran (1977), de Dussaix (1987b) et de Binder et Hidiroglou (1988) pour une description des procédures pouvant être mises en œuvre dans un tel contexte et de plus amples références. Nous donnons la forme explicite des estimateurs de Horvitz-Thompson (l'estimateur sans biais usuel de la théorie des sondages) dans un contexte très général.

Dans la seconde partie, nous donnons les méthodes essentielles de l'approche modèle, fortement liées au développement des techniques de séries temporelles. Nous montrerons comment ces modèles peuvent s'interpréter comme des modèles espace-état, dont la flexibilité permet d'incorporer un grand nombre de renseignements à la fois sur la façon dont sont recueillies les données (dynamique d'échantillonnage) et sur la dynamique du paramètre d'intérêt. Nous analyserons les différentes approches utilisées sur les trois types de données considérées dans l'introduction et en guise d'illustration nous appliquerons ces méthodes à l'analyse de l'évolution de la consommation d'huile d'olive à partir du panel avec rotation de Secodip.

ANALYSE STATISTIQUE DES ENQUÊTES : L'APPROCHE CLASSIQUE

L'approche classique en sondage met essentiellement l'accent sur l'erreur d'échantillonnage. Dans cette optique, le problème est de trouver des **estimateurs linéaires des observations, sans biais et optimaux** (au sens où ils sont de variance minimum). La construction de tels estimateurs repose essentiellement sur le théorème de Gauss-Markov (voir Gurney et Daly, 1965 ; Wolter, 1979 ; Jones, 1980).

1. Notations

On note dans la suite $E_t = (Y^{(t)}, \Pi^{(t)})$ le sondage effectué à la date t dans une population $Y^{(t)} = (Y_1,...,Y_{N_t})$ de taille N_t selon le plan de sondage $\Pi^{(t)}$, pour $t = 1,...,T$. La loi jointe des plans de sondages $\Pi^{(t)}$ est notée Π. Pour simplifier, on s'intéresse à l'estimation de la moyenne

$$\theta^{(t)} = N_t^{-1} \sum_{i=1}^{Nt} Y_i^{(t)},\ t = 1,...,T$$

mais les résultats peuvent facilement s'étendre à des paramètres non linéaires tels des fractiles ou des ratios : on notera néanmoins que, dans ces deux cas, si l'on conserve asymptotiquement la plupart des propriétés, l'absence de biais et l'optimalité à distance finie sont en général perdues.

On observe un échantillon $e^{(t)} = (y_1^{(t)},..., y_{n_t}^{(t)})$. Sur cet échantillon il est possible de fournir plusieurs estimations $\overline{y}_i^{(t)}$, $i = 1,..., k_t$, du paramètre $\theta^{(t)}$. Ces estimateurs $\overline{y}_i^{(t)}$, i = 1,..., k_t, de la moyenne, sont des estimateurs élémentaires à la date t. On note $S_{i,t}$ l'écart-type de l'estimateur élémentaire $\overline{y}_i^{(t)}$ et $\hat{S}_{i,t}$ un estimateur (au moins convergent) de cette quantité.

On peut par exemple supposer que $e^{(t)}$ est divisé en r_t sous-populations. Dans de très nombreuses applications, on a $r_t = 2$ et cette partition correspond à la sous-population M_t des individus qui sont dans l'échantillon aux dates $t-1$ et t (*« matched »*) et la sous-population U_t de ceux qui sont nouveaux à la date t (*« unmatched »*). Cette subdivision apparaît naturellement pour des panels avec rotation (voir par exemple annexe 1). Pour des enquêtes par panels sans rotation on a $U_t = \emptyset$, pour des enquêtes répétées indépendantes $M_t = \emptyset$.

Pour écarter toute ambiguïté (souvent présente dans de nombreux travaux sur les enquêtes répétées), on notera par la suite $\overline{x}_i^{(t-1)}$, $i = 1,...,$ r_t la valeur des estimateurs élémentaires à la date $t-1$ selon la partition effectuée à la date t. Les $\overline{x}_i^{(t-1)}$ sont en fait aussi des estimateurs élémen-

taires de la date $t-1$ et peuvent donc être inclus dans l'ensemble des $\overline{y}_i^{(t-1)}$ mais il n'y a aucune raison (pour des échantillons tournants) que $\overline{x}_i^{(t-1)} = \overline{y}_i^{(t-1)}$) (*i.e.* que $M_t = M_{t-1}$ et/ou $U_t = U_{t-1}$). Dans le cas $r_t = 2$, la construction d'estimateurs sans biais nécessite une bonne connaissance du processus de renouvellement de l'échantillon et donc de la loi jointe du plan de sondage Π qui a conduit à ce renouvellement. Les estimateurs de Horvitz-Thompson de la moyenne et leur variance dans ce cadre sont donnés en annexe 1.

2. Modèle classique

Le modèle statistique classique postule que, en espérance, les $\overline{y}_i^{(t)}$ sont des estimateurs sans biais (au moins asymptotiquement) de $\theta^{(t)}$. La théorie des sondages privilégie en effet le critère d'absence de biais et les estimateurs de Horvitz-Thompson (voir annexe 1) qui fournissent simplement des estimateurs sans biais des moyennes par pondérations des invidus par l'inverse de leur probabilité d'inclusion (probabilité qu'un individu soit tiré lors du sondage) (voir Gouriéroux, 1985 ; Särndall, Swensson, Wretman, 1992). On a :

$$\overline{y}_i^{(t)} = \theta^{(t)} + e_i^{(t)}$$

avec $E_\Pi e_i^{(t)} = 0$ soit sous forme matricielle :

$$\overline{Y} = A\theta + e \qquad (2.1)$$

avec

$$\overline{Y}' = (\overline{y}_1^{(1)}, ..., \overline{y}_{k_1}^{(1)}, ..., \overline{y}_1^{(t)}, ..., \overline{y}_{k_t}^{(t)}, ..., \overline{y}_1^{(T)}, ..., \overline{y}_{k_T}^{(T)}),$$

$$e' = (e_1^{(1)}, ...e_{k_1}^{(1)}, ..., e_1^{(t)}, ..e_{k_t}^{(t)}, ..., e_i^{(T)}, ...e_{k_T}^{(T)})$$

$$\theta' = (\theta^{(1)}, ..., \theta^{(T)})$$

et A est une matrice de taille $(\Sigma_{t=1}^{T} k_t, T)$,

$$A' = \begin{pmatrix} \overbrace{1\ 1\ 1}^{k_1} & 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0\ 0\ 0 & 1 & 1 & 0 & 0 & \cdots & 0 & 0 \\ \vdots\ \vdots\ \vdots & & & & & & & \\ 0\ 0\ 0 & \cdots & \cdots & \cdots & \cdots & 1 & 1 & 1 \\ & & & & & \underbrace{\qquad}_{k_T} & & \end{pmatrix}$$

Il s'agit d'un modèle quasi-linéaire des paramètres θ qui présente de fortes analogies avec l'analyse de la variance. Les estimateurs optimaux

dans ce type de modèle dépendent essentiellement de la matrice de variance covariance des résidus, *i.e.* de :

$$\Sigma = V_{\Pi}(e)$$

dont la structure peut être obtenue analytiquement, si le plan de sondage Π est connu (voir annexe 1). L'estimateur sans biais optimal est alors donné par le théorème de Gauss-Markov par

$$\hat{\theta} = (A'\Sigma^{-1}A)^{-1}\, A'\Sigma^{-1}\, Y = [\hat{\theta}^{(t)}]_{1 \le t \le T} \qquad (2.2)$$

de variance :

$$V_{\Pi}(\hat{\theta}) = (A'\Sigma^{-1}A)^{-1}.$$

Dans une perspective dynamique, on s'intéresse en général plutôt aux contrastes $\hat{\tau} = \hat{\theta}^{(t)} - \hat{\theta}^{(t-1)} = c'\hat{\theta}$, $t = 2,...,T$, avec $c' = (0,..., \underbrace{-1}_{t-1}, \underbrace{1}_{t} ,... 0)$, combinaison linéaire des paramètres. La matrice de variance covariance de cet estimateur se déduit simplement de celle de $\hat{\theta}$:

$$V(\hat{\tau}) = c' V_{\Pi}(\hat{\theta}) c = c'(A'\Sigma^{-1}A)^{-1}c\ . \qquad (2.3)$$

Afin de ne pas surcharger la présentation nous reportons en annexe 1 le calcul explicite de l'ensemble des quantités intervenant dans les calculs dans un cadre général puis pour des sondages classiques poissonniens (pouvant servir d'approximation à des sondages asymptotiquement indépendants). Nous montrons dans les exemples suivants comment cette approche s'applique à l'analyse de l'évolution d'un paramètre suivant les différents types de données listées dans l'introduction.

3. Exemples

Exemple 1 : *Enquêtes indépendantes répétées dans le temps*

On suppose $k_t = r_t = 1$ (un seul estimateur sur un seul échantillon à chaque période), dans ce cadre la matrice de variance covariance Σ est diagonale et l'estimateur optimal de θ n'est autre que $\overline{Y}$. On est simplement ramené à un problème à T-échantillons et les tests usuels (tests de comparaison de moyenne) s'appliquent, lorsque les hypothèses de normalité asymptotique dans les sondages sont satisfaites (cf. Rosen, 1972) et Sen (1988) pour une revue de la littérature sur le sujet). En particulier le test de l'hypothèse H_0 : $\theta^{(t)} = \theta^{(t-1)}$ est basé sur la statistique de Student

$$\frac{\overline{y}_1^{(t)} - \overline{y}_1^{(t-1)}}{\sqrt{\widehat{S}^2_{1,t} + \widehat{S}^2_{1,t-1}}}$$

asymptotiquement gaussienne. On notera donc que l'approche classique ignore ici complètement la composante temporelle du paramètre.

Exemple 2 : Panels

On fait souvent l'hypothèse que les erreurs d'échantillonnage sont liées d'une période sur l'autre par une corrélation ρ (voir Patterson, 1950), que l'on suppose connue. Ceci revient en fait à considérer qu'il existe une liaison sous-jacente de la forme

$$\bar{y}_i^{(t)} - \theta^{(t)} = \rho(\bar{x}_i^{(t-1)} - \theta^{(t-1)}) + \eta_i^{(t)},\ t = 2,...,T \qquad (2.4)$$

où $\eta_i^{(t)}$ est une suite de bruits blancs indépendants. Il convient de remarquer que ce modèle est un modèle dynamique sur les estimateurs et non sur les paramètres comme ce sera le cas dans la partie suivante. Nous donnons en annexe 1 l'expression de ρ dans un contexte très général ainsi qu'un estimateur simple de cette quantité pour des sondages poissonniens (utilisés pour approximer des sondages asymptotiquement indépendants). Dans ce cas, la corrélation des erreurs d'échantillonnage ρ s'interprète aussi comme une version pondérée de la corrélation entre $Y^{(t)}$ et $Y^{(t-1)}$. Il s'ensuit donc que dans les panels, l'introduction de la dynamique sur les erreurs d'échantillonnage conduit indirectement à prendre en compte la dynamique sur les paramètres.

On suppose là encore pour simplifier $r_t = k_t = 1$, $t = 1,...,T$. Dans ce cas, on a $\bar{x}_1^{(t-1)} = \bar{y}_1^{(t-1)}$. On rappelle que $S_{1,t}$ est l'écart-type de l'estimateur élémentaire $\bar{y}_1^{(t)}$. On en déduit la matrice de variance covariance du modèle classique

$$\Sigma = \begin{pmatrix} S_{1,1}^2 & \rho S_{1,1}S_{1,2} & & & \rho^{T-1}S_{1,1}S_{1,T} \\ \rho S_{1,1}S_{1,2} & S_{1,2}^2 & \rho S_{1,2}S_{1,3} & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \rho S_{1,T-1}S_{1,T} \\ & & & \rho S_{1,T-1}S_{1,T} & S_{1,T}^2 \end{pmatrix}$$

Si $\rho = 0$, on est ramené au cas précédent. Si l'on s'intéresse à l'ensemble des paramètres $\theta = [\theta^{(t)}]_{1\le t\le T}$, la matrice A n'est autre que l'identité et l'estimateur des moindres carrés généralisés n'est autre ici que l'estimateur des moindres carrés ordinaires : on en déduit donc une fois de plus que

$$\hat{\theta} = \bar{Y} = [\bar{y}_1^{(t)}]_{1\le t\le T}.$$

La variance d'une combinaison linéaire des paramètres se déduit de (2.3). (2.4) ne change donc rien à l'estimateur lui-même : seule la matrice de variance covariance de l'estimateur est différente et tient compte de ρ.

Ainsi le test de l'hypothèse H_0: $\theta^{(t)} = \theta^{(t-1)}$ est cette fois-ci basé sur la statistique

$$\frac{\overline{y}_1^{(t)} - \overline{y}_1^{(t-1)}}{\sqrt{\widehat{S}_{1,t}^2 + \widehat{S}_{1,t-1}^2 - 2\rho\widehat{S}_{1,t}\widehat{S}_{1,t-1}}}.$$

On notera que cette formulation ne permet pas de prendre en compte le cas $\rho = 1$ qui correspond à la forme non-stationnaire de 2.4.

Exemple 3 : Echantillons tournants (Wolter, 1979)

Supposons $r_t = 2$, $t = 1,...,T-1$ correspondant aux sous-populations M_{t+1} (population « *matched* » présente en t et $t+1$) et U_{t+1}. On indicera par 1 les estimateurs élémentaires calculés sur la sous-population M_t supposée de taille fixe n_1 et par 2 ceux calculés sur U_t, de taille fixe n_2. On suppose que la taille entre $t-1$ et t est fixe égale à $n = n_1 + n_2$.

On suppose que $\overline{y}_i^{(t)}$ et $\overline{x}_i^{(t-1)}$ satisfont une condition similaire à (2.4). On dispose alors à chaque période de quatre estimateurs élémentaires (sauf en T car U_{T+1} et M_{T+1} et donc les $\overline{x}_i^{(T)}$ ne sont pas définis). On peut alors en déduire les estimateurs optimaux à partir de la relation (2.2).

Par exemple, on s'intéresse à la comparaison de deux moyennes à deux périodes. On suppose ici afin de comprendre le rôle de ρ que les variances sont égales à $S_{i,t}^2 = S^2/n_i$ (hypothèse forte qui est satisfaite si le sondage est à probabilités égales avec remise et si la structure de la variable Y n'évolue dans le temps que par translation). L'estimateur optimal de $\theta^{(t)} - \theta^{(t-1)}$ est alors

$$\widehat{\theta}^{(t)} - \widehat{\theta}^{(t-1)} = \overline{y}_2^{(t)} - \overline{x}_2^{(t-1)} + \frac{n_1}{n - \rho n_2}((\overline{y}_1^{(t)} - \overline{x}_1^{(t-1)}) - (\overline{y}_2^{(t)} - \overline{x}_2^{(t-1)})) \quad (2.5)$$

de variance

$$V = \frac{2S^2}{n - \rho n_2}(1 - \rho).$$

L'estimateur (2.5) s'interprète comme la différence entre les moyennes intertemporelles sur les deux populations indépendantes, corrigée des écarts de cette différence observés sur les deux sous-populations M_t et U_t. La correction est d'autant plus forte que la corrélation ρ entre les erreurs d'échantillonnage des plans de sondage est forte : ainsi en cas limite, si $\rho = 1$, l'estimateur se base uniquement sur la sous-population M_t.

Dans tous les cas, le test H_0 : $\theta^{(t)} = \theta^{(t-1)}$ est donc simplement basé sur

$$\frac{\hat{\theta}^{(t)} - \hat{\theta}^{(t-1)}}{\hat{V}^{1/2}}$$

où $\hat{V}$ est une estimateur de V. On en déduit que si $\rho > 0$, alors le choix optimal pour n_2 (en terme de minimisation de la variance de l'estimateur est $n_2 = 0$, *i.e.* le panel est préférable à l'échantillon tournant : on est alors ramené aux estimateurs de l'exemple 2. Si $\rho < 0$ alors il faut prendre $n_2 = n$, *i.e.* les enquêtes répétées (exemple 1) sont préférables. Il convient cependant de prendre garde à l'interprétation de ces résultats : le coefficient de corrélation ρ s'interprète comme *la corrélation entre les erreurs d'échantillonnage* dont la valeur est liée au coefficient de corrélation entre les variables $Y^{(t-1)}$ et $Y^{(t)}$. Ces résultats signifient simplement que en prenant des différences entre des erreurs très corrélées (positivement), on fait une erreur plus faible...

Dans le cas général, la mise en œuvre de ces procédures statistiques simples nécessite de disposer d'estimateurs des variances des estimateurs élémentaires mais aussi de ρ. En théorie ρ est entièrement défini par la loi jointe Π, des plans de sondage aux différentes dates et il est possible d'en donner une expression analytique et un estimateur simple (voir annexe 1). Cependant en pratique, il est souvent fréquent que l'on connaisse les lois marginales $\Pi^{(t)}$ du plan de sondage mais que ρ soit inconnu.

Plus généralement, il est nécessaire de disposer d'un estimateur de la matrice de variance covariance Σ qui est en général inconnue. Lorsque T est grand, il est possible de recourir aux moindres carrés quasi-généralisés : *i.e.* d'estimer d'abord le modèle par les moindres carrés ordinaires et d'estimer Σ à partir de résidus estimés. Ceci nécessite cependant d'avoir un grand nombre d'estimateurs élémentaires à chaque période. Ce type de technique d'estimation de la variance s'apparente à des techniques de type jackknife et bootstrap ou de sous-échantillonnage (Bertail, 1997). Leur mise en œuvre s'avère assez simple lorsque le plan de sondage est uniforme, à probabilités égales, mais peut devenir beaucoup plus complexe lorsque les plans de sondage sont à probabilités inégales (voir Bertail et Combris, 1997).

Si l'on suppose que la structure (2.4) est vérifiée avec $\Theta = \theta^{(i)} = \ldots = \theta^{(T)}$ et que T est suffisamment grand alors un estimateur de ρ est donné par

$$\hat{\rho} = \frac{\sum_{t=2}^{T}(\bar{y}_1^{(t)} - \bar{\bar{y}})(\bar{y}_1^{(t-1)} - \bar{\bar{y}})}{\sum_{t=1}^{T}(\bar{y}_1^{(t)} - \bar{\bar{y}})^2} \qquad (2.6)$$

et il est possible de remplacer les expressions dans (2.5) pour obtenir un estimateur convergent. Ce résultat permet par exemple de tester la stabilité d'un paramètre sur une longue période.

La question qui se pose pour les échantillons tournants est de savoir comment choisir la loi jointe du plan de sondage à chaque étape de manière à avoir la plus grande précision possible : ceci détermine comment en pratique renouveller l'échantillon. Ce problème très délicat est étudié par exemple par Särndal, Swensson, Wretman (1992), chapitre 9, dans un cadre relativement simple mais il n'existe pas de solutions ad-hoc, la solution optimale dépendant fortement des objectifs fixés, qui peuvent générer des solutions incompatibles.

Une façon simple et très générale de procéder lorsqu'on dispose d'estimateurs et de paramètres complexes est de considérer une combinaison des estimateurs construits sur les sous-populations d'intérêt, par exemple sous la forme composite (voir Jones, 1980)

$$\widehat{\theta} = \sum_{t=1}^{T} \lambda_t \overline{y}_1^{(t)} + \mu_t \overline{y}_2^{(t)} + \delta_t \overline{x}_1^{(t-1)} + \gamma_t \overline{x}_2^{(t-1)} \qquad (2.7)$$

où les $\overline{y}_i^{(t)}$ et les $\overline{x}_i^{(t-1)}$ désignent plus généralement des estimateurs élémentaires « naturels » des paramètres (pas forcément sans biais) correspondant à la partition U_t, M_t. L'étape suivante est d'écrire la condition d'absence de biais (au moins asymptotique) par rapport au paramètre considéré

$$E_\Pi \hat{\theta} = \theta.$$

Ceci contraint les pondérations à satisfaire des contraintes du type

$$h(\{\lambda_t, \mu_t, \delta_t, \gamma_t\}_{1 \le t \le T}) = 0 \qquad (2.8)$$

On cherche alors quelles sont les valeurs de λ_t, μ_t, δ_t, γ_t qui minimisent la variance de $\hat{\theta}$ sous la contrainte (2.8). L'estimateur (2.5) peut s'obtenir directement par cette méthode.

Illustration : le panel (avec rotation) Secodip

Les panels de ménages de Secodip sont constitués sur la base d'un échantillonnage aléatoire stratifié à deux niveaux (régions, habitat) à partir des données du recensement général de la population. Le plan de sondage est reconstruit périodiquement pour tenir compte des résultats du recensement le plus récent. Les panélistes retournent des relevés hebdomadaires d'achat. Pendant une période de mise en route d'une durée minimum de trois semaines les relevés font l'objet d'un contrôle systématique mais ne sont pas pris en compte dans les résultats statistiques périodiques. Les panels sont partiellement renouvelés à la fin de chaque trimestre ce qui conduit chaque année au remplacement d'un quart de l'échantillon. Un premier problème intéressant est la comparaison des moyennes des consommations par tête entre deux années successives.

Pour chaque ménage, on dispose chaque année de l'observation de la consommation totale CT_t, $t = 1,...,T$ où t indice les années. Etant donné

le faible nombre d'années observées dont nous disposons actuellement, il semble difficile d'adopter une approche de type modèle sans spécifier la dynamique de la consommation annuelle . On peut appliquer ici les relations obtenues dans l'exemple 3 sur un panel avec rotation. L'annexe 1 donne la forme générale des estimateurs sans faire d'hypothèses fortes sur les plans de sondages utilisés et le comportement de la consommation d'une période sur l'autre. On notera que l'utilisation de cet estimateur nécessite de connaître les probabilités d'inclusion d'ordre 1 et 2 du plan de sondage initial (*i.e.* la pondération des individus) ainsi que celles qui déterminent le renouvellement, à moins de supposer l'indépendance asymptotique du sondage, *i.e.* le caractère asymptotiquement poissonnien du plan de sondage. En utilisant cette méthode on peut montrer par exemple sur le cas de la consommation d'huile d'olive qu'il n'y a pas de variation significative de la consommation: nous verrons que cette conclusion est remise en cause par l'approche modèle.

Notons ici que les pondérations fournies par Secodip ne correspondent pas exactement aux inverses des probabilités d'inclusion telles qu'elles ont été introduites jusqu'à présent mais résultent d'un calage sur marge destiné à corriger les biais d'échantillonnage. Pour ne pas alourdir la présentation, nous ne tiendrons pas compte de cette correction supplémentaire, qui pose des problèmes théoriques certains dans l'étude de la dynamique et qui pourrait faire l'objet de traitements plus sophistiqués.

L'APPROCHE MODÈLE: VERS UNE MODÉLISATION DE LA DYNAMIQUE DES PARAMÈTRES

Les exemples 2 et 3 donnés dans le paragraphe précédent montrent clairement que faire simplement une hypothèse de corrélation entre les erreurs d'échantillonnage n'est pas suffisant pour pouvoir comprendre la dynamique du paramètre. Ceci a conduit à introduire des modèles stochastiques sur le paramètre, très proches de ceux utilisés en séries temporelles, et donc à considérer le paramètre lui-même comme une variable aléatoire. L'idée de fond de l'approche modèle est que l'information passée notée $\Im_{t-1}$ ou passée et présente notée $\Im_t$ conditionne la valeur du paramètre à la date t. Dans la plupart des cas, cette information est simplement la valeur des estimateurs élémentaires des paramètres aux périodes précédentes. Par exemple, dans le cas de l'estimation d'une moyenne on a

$$\Im_t = \{\bar{y}_i^{(t')}, \bar{x}_i^{(t'-1)}, i = 1,2,\ t' \leq t\}.$$

Le modèle tient donc compte du fait que l'information recueillie aux périodes précédentes apporte de l'information sur la valeur présente du paramètre. Dans cette optique, proche d'une optique de type bayésien, le

paramètre θ est lui-même une variable aléatoire. Les estimateurs obtenus sont donc maintenant des estimateurs sans biais de $E\theta$ (voir Dussaix, 1987b). L'estimateur linéaire sans biais optimal (en terme de risque bayésien) est

$$\widehat{\theta}_t^{(t)} = E(\theta^{(t)}|\Im_t)$$

i.e. la meilleure prédiction de $\theta^{(t)}$ connaissant $\Im_t$. De façon analogue, le meilleur prédicteur linéaire sans biais de $\theta^{(t+1)}$ ne disposant que des informations à la date t est

$$\widehat{\theta}_t^{(t+1)} = E(\theta^{(t+1)}|\Im_t) .$$

La quantité

$$i_{t+1} = \widehat{\theta}_{t+1}^{(t+1)} - \widehat{\theta}_t^{(t+1)}$$

est appelée innovation du processus à la date $t + 1$: c'est la différence entre l'estimateur optimal en $t + 1$ et la valeur prédite en $t + 1$ avec l'information passée $\Im_t$. Si le modèle sous-jacent est vérifié alors on a

$$Ei_{t+1} = 0.$$

Il peut arriver qu'à une date $t + 1$ un choc important (au sens de contamination de la loi du paramètre) affecte la variable $\theta^{(t+1)}$ de sorte que le modèle n'est plus vérifié et l'innovation du processus ne satisfait pas $Ei_{t+1} = 0$. Tester la présence d'un choc *i.e.*, H_o: $Ei_{t+1} = 0$, revient alors simplement à comparer i_{t+1} à 0.

1. L'approche série temporelle

Modélisation de la dynamique des paramètres

Dans ce type de modèle, la dynamique du paramètre est décrite simplement par une équation récursive selon une optique de série temporelle. Par exemple, Blight et Scott (1973) ont introduit le modèle suivant qui tient compte à la fois des dynamiques des paramètres et de l'erreur d'échantillonnage :

$$\theta^{(t)} - \mu = \alpha(\theta^{(t-1)} - \mu) + \epsilon_t \, , \epsilon_t \; i.i.d. \; E\epsilon_t = 0, \; V(\epsilon_t) = \sigma_\epsilon^2 \qquad (3.1)$$

$$\overline{y}_1^{(t)} - \theta^{(t)} = \rho(\overline{x}_1^{(t-1)} - \theta^{(t-1)}) + \eta_1^{(t)} \, , E\eta_1^{(t)} = 0, \; V\eta_1^{(t)} = \sigma_{\eta,1}^2 \qquad (3.2)$$

$$\overline{y}_2^{(t)} = \theta^{(t)} + \eta_2^{(t)} \, , E\eta_2^{(t)} = 0, \; V\eta_2^{(t)} = \sigma_{\eta,2}^2 \qquad (3.3)$$

L'équation (3.1) décrit le comportement du paramètre d'intérêt sous la forme d'un modèle autorégressif d'ordre 1 (AR(1)). Il est bien sûr possible de complexifier cette structure simple pour obtenir un modèle plus flexible, par exemple en modélisant le comportement de $\theta^{(t)}$ sous la

forme d'un ARMA(p,q) (autorégressif moyenne mobile d'ordre (p,q)) de la forme

$$\Phi_p(B)(\theta^{(t)} - \mu) = \Psi q(B)\varepsilon_t, \tag{3.4}$$

où $\Phi_p(B)$ et $\Psi_q(B)$ sont des polynômes retard respectivement d'ordre p et q. Cette équation peut s'interpréter dans le cadre des sondages comme un modèle de surpopulation (ou modèle bayesien) sur le paramètre.

L'équation (3.2) similaire à (2.4) décrit le mécanisme d'erreur d'échantillonnage associé à l'estimateur élémentaire construit sur la population M_t. Là encore l'hypothèse AR(1) faite par Blight et Scott (1973) peut s'avérer peu flexible et des modélisations de type autorégressif moyenne-mobile (ARMA(p,q)) ont aussi été proposées (voir par exemple Scott et Smith, 1977). Lorsque le paramètre présente des caractéristiques de non-stationnarité (ce qui est souvent le cas pour des paramètres économiques), des modèles autorégressifs moyenne-mobile intégrés ARIMA(p,d,q) peuvent s'avérer plus intéressants. Enfin l'équation (3.3) décrit l'erreur d'échantillonnage relative à l'estimateur élémentaire construit sur U_t.

Enquêtes indépendantes

Cette approche a l'avantage sur l'approche classique d'intégrer la dynamique du paramètre même si les enquêtes répétées sont indépendantes (auquel cas le modèle se réduit aux équations (3.1) (voire (3.4) pour le cas plus général) et (3.3).

On peut montrer (voir Scott et Smith, 1974) que, sous (3.1), l'estimateur optimal s'obtient de manière récursive par réactualisation de sa prédiction

$$\hat{\theta}_t^{(t)} = (1 - \lambda_t)\bar{y}_2^{(t)} + \lambda_t \hat{\theta}_{t-1}^{(t)}$$

avec

$$\lambda_t = \frac{Var(\bar{y}_2^{(t)})}{Var(\bar{y}_2^{(t)}) + Var(\theta^{(t)}|\Im_{t-1})} \tag{3.5}$$

En effet, pour comprendre de façon intuitive ce résultat, on peut remarquer que, si aucune information n'était disponible à la date t, l'estimateur sans biais optimal (en terme de variance minimale) serait $\hat{\theta}_{t-1}^{(t)}$, la prédiction de $\theta^{(t)}$ compte tenu de l'information disponible aux dates précédentes $\Im_{t-1}$. Ayant collecté l'information en t et calculé $\bar{y}_2^{(t)}$, il est naturel de supposer qu'un meilleur estimateur de $\theta^{(t)}$ est une combinaison linéaire de ces deux valeurs (les échantillons étant indépendants, $\bar{y}_2^{(t)}$ est indépendant de $\hat{\theta}_{t-1}^{(t)}$). La pondération optimale est celle qui minimise la variance de l'estimateur, *i.e.*

$$(1 - \lambda_t)^2 Var(\bar{y}_2^{(t)}) + \lambda_t^2 Var(\theta^{(t)}|\Im_{t-1})$$

Le minimum en λ_t de cette quantité est réalisé pour la valeur de λ_t donnée en (3.5). On a alors l'interprétation suivante de cette formule de réactualisation: si la prédiction de $\theta^{(t)}$ sachant le passé $\Im_{t-1}$ est très bonne alors on a $Var(\theta^{(t)}|\Im_{t-1}) \cong 0$ et donc $\lambda_t \cong 1$ de sorte que la connaissance du passé suffit à construire un estimateur très précis de $\theta^{(t)}$. Ceci sera par exemple le cas si le paramètre $\theta^{(t)} = \mu$ est pratiquement constant au cours du temps. Inversement, si le comportement de $\theta^{(t)}$ est difficilement prédictible (ceci correspondant au cas $Var(\theta^{(t)}|\Im_{t-1}) \cong \infty$), le meilleur estimateur de $\theta^{(t)}$ est donné par l'information à la date t. En pratique, on se situe entre ces pôles et la valeur optimale de λ_t dépend essentiellement de la dynamique sous-jacente à $\theta^{(t)}$. Une estimation précise de $Var(\theta^{(t)}|\Im_{t-1})$ passe par l'estimation préliminaire des paramètres du modèle de surpopulation.

Pour tester la présence d'un choc, il suffit de remarquer que

$$\widehat{\theta}_t^{(t)} - \widehat{\theta}_{t-1}^{(t)} = (1 - \lambda_t)(\overline{y}_2^{(t)} - \widehat{\theta}_{t-1}^{(t)}) \ ,$$

de sorte que le test est simplement basé sur la statistique de test

$$\frac{(\overline{y}_2^{(t)} - \widehat{\theta}_{t-1}^{(t)})}{\sqrt{Var(\overline{y}_2^{(t)}) + Var(\theta^{(t)}|\Im_{t-1})}} \ ,$$

i.e. la comparaison de l'estimation sur l'échantillon observé en t avec sa prédiction.

Panels (avec ou sans rotation): analyse de la consommation dans le panel Secodip

Tenir compte de la dynamique du paramètre dans un panel avec ou sans rotation depend très fortement de la manière dont sont obtenues les données et de la forme du modèle retenu pour les paramètres. Aussi nous nous contenterons d'illustrer ici ce type de modèle avec l'exemple particulier du panel Secodip. Ce panel se compose d'environ 3500 ménages observés hedomadairement sur deux ans.

Le processus économique observé, la consommation d'un produit sur une semaine notée $\theta^{(t)}$ (pour conserver les notations des parties précédentes) est stationnaire pour un grand nombre de produits de consommation courante. On supposera que le processus admet une représentation AR(1) de la forme

$$\theta^{(t)} - \mu = \phi\,(\theta^{(t-1)} - \mu) + \varepsilon_t \qquad (3.6)$$

soit en terme d'opérateur retard

$$(1 - \phi B)(\theta^{(t)} - \mu) = \varepsilon_t, \qquad (3.7)$$

ce qui suppose entre autres que le type de produit considéré n'a pas un caractère saisonnier annuel auquel cas il conviendrait d'introduire l'opé-

rateur de différentiation $(1 - B^{12})$. ϕ de module inférieur à 1 s'interprète comme la corrélation à l'ordre 1 entre les variations de consommation. Dans cette écriture μ s'interprète comme la moyenne.

La modélisation de l'erreur d'échantillonnage est plus délicate: en effet il s'agit de tenir compte du fait que, sur la première année indicée $1,\ldots,T_1$, on dispose en fait d'un panel complet noté P. Sur ce sous-ensemble, les erreurs d'échantillonnage sont fortement corrélées. Le renouvellement s'opère à la date $T_1 + 1$: on observe alors un nouveau panel complet composé des deux populations M (« *matched* ») (3/4 de l'échantillon total) et U (« *unmatched* ») suivies pendant T_2 semaines.

La partition M et U permet de définir des estimateurs $\bar{y}_1^{(t)}$ et $\bar{y}_2^{(t)}$, $t = T_1 + 1,\ldots,T_2 + T_1$ qui sont liés par une condition du type

$$\begin{aligned} \bar{y}_1^{(t)} - \theta^{(t)} &= \rho_1(\bar{y}_1^{(t-1)} - \theta^{(t-1)}) + \epsilon_{1,t},\ t = T_1 + 2,\ldots,T_2 + T_1 \\ \bar{y}_2^{(t)} - \theta^{(t)} &= \rho_2(\bar{y}_2^{(t-1)} - \theta^{(t-1)}) + \epsilon_{2,t},\ t = T_1 + 2,\ldots,T_2 + T_1 \end{aligned} \tag{3.8}$$

$$\bar{y}_2^{(T_1+1)} - \theta^{(T_1+1)} = \epsilon_{2,T_1+1} \tag{3.9}$$

où $\varepsilon_{1,t}$ et $\varepsilon_{2,t}$ définissent une suite de v.a. indépendantes de variances respectives $\sigma^2_{\varepsilon 1}$ et $\sigma^2_{\varepsilon 2}$. En moyenne, les estimateurs sont convergents de $\theta^{(t)}$. (3.8) traduit le fait que les erreurs d'échantillonnage sont corrélées d'une période sur l'autre. (3.9) modélise l'erreur d'échantillonnage au moment du renouvellement. ρ_1 dépend essentiellement de la structure de la population de M, du plan de sondage initial et du tirage conditionnel à P. ρ_2 dépend essentiellement de la façon dont l'échantillon U a été constitué.

Rétrospectivement, la partition M et U permet de définir des estimateurs $\bar{x}_1^{(t)}$ et $\bar{x}_2^{(t)}$, $t = 1,\ldots,T_1$ calculé sur M et $P - M$, qui sont liés par une condition du type

$$\bar{x}_1^{(t)} - \theta^{(t)} = \rho_1(\bar{x}_1^{(t-1)} - \theta^{(t-1)}) + \epsilon_{1,t},\ t = 2,\ldots,T_1 \tag{3.10}$$

$$\bar{y}_1^{(T_1+1)} - \theta^{(T_1+1)} = \rho_1(\bar{x}_1^{(T_1)} - \theta^{(T_1)}) + \epsilon_{1,T_1} \tag{3.11}$$

$$\bar{x}_2^{(t)} - \theta^{(t)} = \rho_3(\bar{x}_2^{(t-1)} - \theta^{(t-1)}) + \epsilon_{3,t},\ t = 2,\ldots,T_1 \tag{3.12}$$

$$\bar{x}_1^{(1)} - \theta^{(1)} = \epsilon_{1,1} \tag{3.13}$$

$$\bar{x}_2^{(1)} - \theta^{(1)} = \epsilon_{3,1} \tag{3.14}$$

(3.10) et (3.11) traduisent l'hypothèse que, pour la population M suivie sur les deux années, la corrélation des erreurs d'échantillonnage est la même y compris au moment du renouvellement. (3.12) modélise le comportement de l'erreur d'échantillonnage sur $P - M$ la première année. (3.13) et (3.14) définissent les conditions initiales. Vu la structure du modèle, il est toujours possible de poser

$$\overline{y}_1^{(t)} = \overline{x}_1^{(t)}, t = 2,\ldots,T_1$$

auquel cas les équations (3.8), (3.10) et (3.11) se réduisent à

$$\overline{y}_1^{(t)} - \theta^{(t)} = \rho_1(\overline{y}_1^{(t-1)} - \theta^{(t-1)}) + \epsilon_{1,t}, \quad t = 2, \ldots, T_2 + T_1 \quad . \tag{3.15}$$

L'analyse et l'estimation de ce type de modèle sont grandement simplifiées par l'introduction de modèles espace-état.

2. Modèles espace-état (ou état-mesure): intégration des dynamiques des paramètres et des échantillonnages

Modèles espace-état et estimation par filtre de Kalman

Les modèles proposés précédemment peuvent se voir comme des cas particuliers de modèles espace-état (connus aussi sous le nom de modèles état-mesure) décrivant le comportement joint du « paramètre » $\theta^{(t)}$ et des estimateurs calculés sur des sous-échantillons. Les modèles espace-état se réduisent à deux équations :

– la première, dite équation d'état, décrit la dynamique et donc l'état du système, *i.e.* des variables d'intérêt X_t à chaque t (que ces variables ou paramètres soient observés ou non)

$$X_t = F_t X_{t-1} + V_t \tag{3.16}$$

– la seconde, dite équation d'espace ou de mesure, décrit le lien entre ce qui est effectivement observé (les Y_t) et les variables d'intérêt

$$Y_t = G_t X_t + e_t, \tag{3.17}$$

où, dans le cadre des sondages, Y_t est un vecteur d'estimateurs élémentaires et X_t est un vecteur contenant un vecteur de paramètres $\theta^{(t)}$ (pouvant être pris à plusieurs dates $u \leq t$) : voir Rao, Srinath et Quenneville (1989). L'équation (3.16) décrit l'évolution du système en termes des matrices connues F_t et des variables résiduelles V_t (en général supposées gaussiennes). L'équation d'espace (3.17) décrit ce qui est observé (avec une erreur possible sur la variable elle-même). G_t désigne une suite de matrices déterministes (dans la plupart des applications ne dépendant pas de t) et e_t une suite de variables aléatoires d'espérance nulle (voir en annexe 2, une description plus précise du modèle espace-état et des hypothèses requises).

De façon générale, dans ce type de modèle on observe les valeurs initiales des processus et la variable d'espace Y_t mais la variable d'état X_t qui décrit la dynamique du système n'est pas entièrement observée. Le problème est d'estimer complètement ou partiellement X_t (dans le cadre des sondages uniquement la composante $\theta^{(t)}$) compte tenu de l'information collectée à la date t (voir par exemple Brockwell and Davies (1991) pour une introduction aux modèles espace-état et aux filtres de Kalman).

Nous rappelons que l'estimation linéaire de X_t en fonction de $\Im_{t-1} = \{Y_0, Y_1,...,Y_{t-1}\}$ est appelée **problème de prédiction** et revient au calcul de

$$\hat{X}^{(t)}_{t-1} \triangleq E(X_t|\Im_{t-1})$$

où $E(X|Z)$ désigne la projection de X sur l'espace engendré par Z, pour la métrique associée au produit scalaire espérance $E(.)$. Dans le modèle gaussien, *i.e.* lorsque le vecteur des résidus (V_t', e_t') est gaussien, $E(X|Z)$ coïncide avec l'espérance conditionnelle. La matrice de variance covariance de l'erreur de prédiction est alors définie par

$$\Omega_t \triangleq E(X_t - \hat{X}^{(t)}_{t-1})(X_t - \hat{X}^{(t)}_{t-1})'.$$

L'estimation linéaire de X_t en fonction de $\Im_t = \{Y_0, Y_1,...,Y_t\}$ est appelée **problème de filtrage** et revient au calcul de

$$\hat{X}^{(t)}_t \triangleq E(X_t|\Im_t).$$

Cette technique est donc parfaitement adaptée à l'estimation d'un paramètre lorsqu'on désire tenir compte de toute l'information passée disponible (contenue dans les données mais aussi apportée par le modèle) à un instant t. Il est alors clair que la comparaison de la valeur prédite avec les échantillon passés $\hat{\theta}^{(t)}_{t-1}$ avec la valeur calculée sur l'échantillon présent peut permettre de mettre en évidence des chocs sur le paramètre d'intérêt.

L'estimation linéaire de X_t en fonction de $\Im_n = \{Y_0, Y_1,...,Y_n\}$, $n > t$ (i.e. toute l'information disponible) est appelée **problème de lissage** et revient dans le cas du modèle gaussien au calcul de $\hat{X}^{(t)}_n \triangleq E(X_t|\Im_n)$. Ainsi connaître la dynamique d'un paramètre sur l'ensemble d'une période en supposant le modèle sous-jacent exact se ramène typiquement à un problème de lissage avec $n = T$.

Le modèle espace-état permet un très grand nombre possible de configurations, ce qui rend son utilisation très flexible : voir Rao, Srinath et Quenneville (1989) pour une utilisation dans les sondages en environnement stationnaire. Par ailleurs, on dispose pour ce type de modèle d'un ensemble d'outils statistiques standard, les filtres de Kalman, issus du contrôle de systèmes linéaires (Kalman, 1960)) qui permet de façon systématique de calculer les valeurs de $\hat{\theta}^{(t)}_t$, $\hat{\theta}^{(t)}_{t-1}$, $\hat{\theta}^{(t)}_T$, et leur variance $Var(\hat{\theta}^{(t)}|\Im_{t-1})$, $Var(\hat{\theta}^{(t)}|\Im_t)$ ou encore $Var(\hat{\theta}^{(t)}|\Im_T)$ par des formules de remise à jour. L'annexe 2 présente quelques rappels fondamentaux sur les procédures à mettre en œuvre.

L'estimation des paramètres se fait en supposant la normalité des résidus et en écrivant la vraisemblance du modèle (mais on peut aussi considérer cette méthode comme du pseudo-maximum de vraisemblance permettant d'obtenir des estimateurs asymptotiquement convergents et gaussiens). L'écriture sous forme de modèles espace-état permet aussi de tenir compte du fait que les enquêtes ne sont pas toujours faites à des dates régulières ou encore (ce qui revient en fait au même) que certaines enquêtes n'ont pas pu être réalisées aux dates requises : on traite alors les valeurs manquantes comme des paramètres et le filtre de Kalman permet de prévoir ces valeurs : voir Brockwell et Davies (1991), chapitre 9, pour ce type d'utilisation. L'estimation des variances peut se faire soit de manière traditionnelle à partir du calcul des dérivées de la vraisemblance, soit en utilisant des techniques de type bootstrap, par exemple de sous-échantillonnage lorsque le modèle est stationnaire (les panels pouvant s'interpréter comme des champs aléatoires discrets il est en particulier possible d'utiliser les résultats de Politis et Romano (1994) et de Bertail (1997)). Il est clair cependant que l'utilisation de ces modèles suppose que l'on a suffisamment de dates T d'observation pour que les estimateurs obtenus aient un sens à distance finie. Par ailleurs, leur principal inconvénient est (comme pour toute modélisation linéaire paramétrique) son manque de robustesse. En effet, un fort choc transitoire sur le paramètre peut considérablement modifier les estimateurs : si ce choc n'est pas détecté puis intégré au modèle *a priori* les estimateurs (en particulier les prédictions) seront en général de mauvaise qualité.

Construction et estimation du modèle espace-état associé au panel Secodip

Les équations (3.7), (3.8), (3.9), (3.12), (3.13), (3.14), (3.15) définissent un modèle espace-état (que l'on peut voir comme un modèle avec données manquantes : les valeurs de $\overline{x}_2^{(t)}$, calculées sur $P - M$ la première année, sont manquantes ensuite. Inversement, les valeurs de $\overline{y}_2^{(t)}$ calculées sur U sont manquantes la première année). On pourra aisément vérifier que la forme du modèle espace-état qui s'en déduit est la suivante ($I_{(A)}$ désigne l'indicatrice de la période A)

$$X_t = F_t X_{t-1} + V_t$$
$$Y_t = G_t X_t$$

avec

$$X_t' = (\theta^{(t)}, \overline{y}_1^{(t)}, \overline{y}_2^{(t)}, \overline{x}_1^{(t)}, \mu)$$

$$Y_t' = \begin{cases} (\overline{y}_1^{(t)}, \overline{x}_2^{(t)}), & t = 1, \ldots, T_1 \\ (\overline{y}_1^{(t)}, \overline{y}_2^{(t)}), & t = T_1 + 1, \ldots, T_1 + T_2 \end{cases}$$

$$F_t = \begin{pmatrix} \phi & 0 & 0 & 0 & 1-\phi \\ \phi - \rho_1 & \rho_1 & 0 & 0 & 1-\phi \\ \phi - \rho_2 & 0 & \rho_2 & 0 & 1-\phi \\ \phi - \rho_3 & 0 & 0 & \rho_3 & 1-\phi \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

$$G_t = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}, \; t = 1, \ldots, T_1$$

$$G_t = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix}, \; t = T_1 + 1, \ldots, T_1 + T_2$$

et

$$V_t = \begin{pmatrix} \epsilon_t \\ \epsilon_t + \epsilon_{1,t} \\ \epsilon_t + \epsilon_{2,t} \\ \epsilon_t + \epsilon_{3,t} \end{pmatrix}.$$

Cette écriture dite canonique du modèle espace-état permet d'estimer les paramètres structurels du modèle, d'estimer et de prévoir (ou lisser) les valeurs de la consommation hebdomadaire. On notera qu'il n'existe pas une écriture unique d'un modèle espace-état mais qu'il existe des modèles de dimension minimale (forme canonique) qui simplifient les procédures d'estimation.

En guise d'illustration, nous avons appliqué ces méthodes à la consommation d'huile d'olive en 1990-1991 en utilisant le panel Secodip. Les estimations ont été conduites à l'aide du module de Gauss, TSM *(Times Series Model)*.

i) Estimation du modèle espace-état

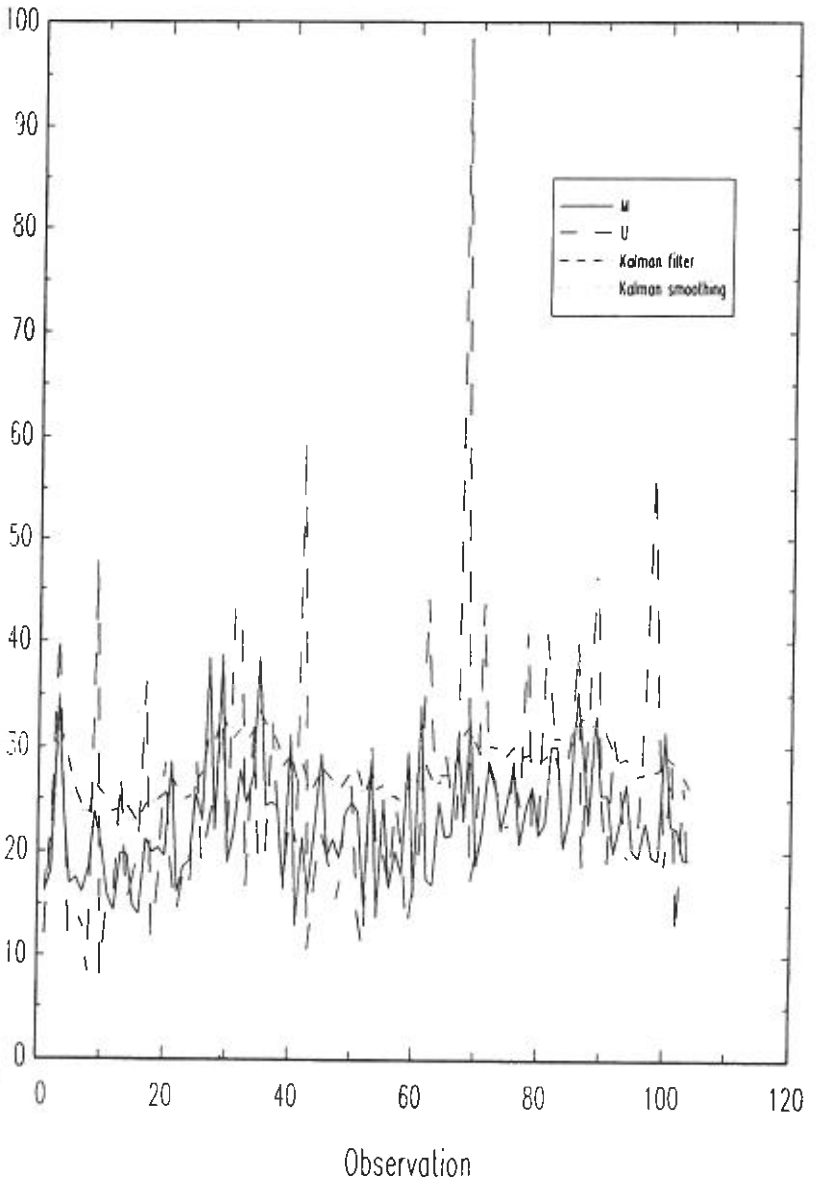
Le tableau suivant donne les estimateurs des paramètres du modèle espace-état.

Paramètres	*Estimation*	*Std*	*P – value*
μ	0,030	7,1*e*-4	0,00
Φ	0,834	0,119	0,00
ρ_1	-0,206	0,098	0,03
ρ_2	-0,029	0,090	0,74
ρ_3	-0,181	0,109	0,09
σ^2	0,0018	7,9*e*-4	0,02
σ_1^2	0,0047	4,7*e*-4	0,00
σ_2^2	0,0131	13,2*e*-4	0,00
σ_3^2	0,0098	10,2*e*-4	0,00

On notera la significativité du coefficient d'autocorrélation sur des données « *matched* » qui indique une dynamique propre de la consommation d'huile d'olive. En revanche ce coefficient est non significatif à 5 %

sur les données *« unmatched »*. Ceci s'explique sans doute pas le faible nombre de données disponibles sur chacune des deux sous-périodes. La significativité des variances σ_i^2, i = 1,2,3 montre que l'on peut ignorer les erreurs dues directement à l'échantillonnage. Par contre l'écart-type σ_2 est à la limite de la significativité (à 1% non significatif), ce qui signifie que l'on pourrait presque considérer l'évolution de la consommation d'huile d'olive comme déterministe.

Graphique 1. **Données de l'échantillon tournant (consommation moyenne d'huile d'olive en centilitre par ménage par semaine). Lissage et prédiction de la consommation moyenne hebdomadaire**

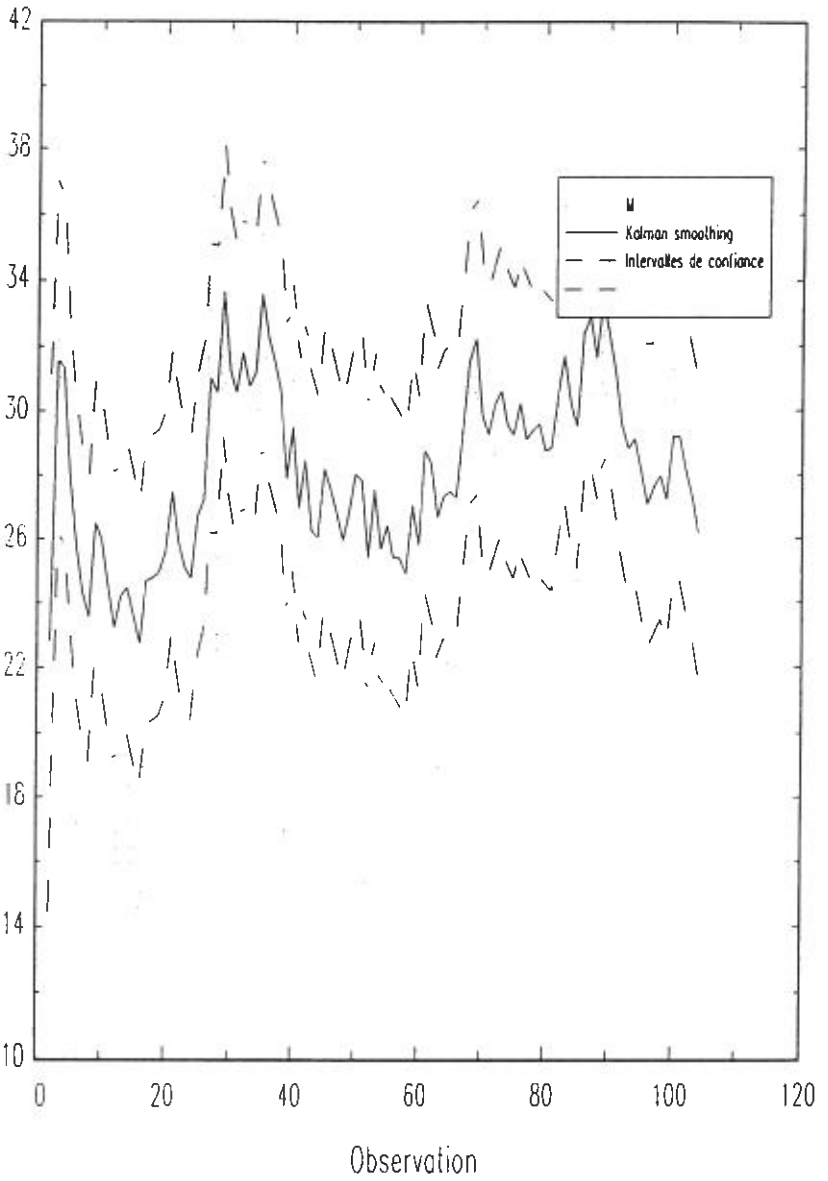


On en déduit alors les valeurs lissées ($E(\theta^{(i)}|\Im_{T1+T2})$) et prévues ($E(\theta^{(i)}|\Im_{t-1})$) de la série des $\theta^{(i)}$ (voir graphique 1) compte tenu de toute

l'information disponible. Ces quantités décrivent la tendance globale de consommation hebdomadaire.

Le graphique 2 suivant donne plus précisément l'estimation de la consommation sur l'ensemble de la période ainsi que les intervalles de confiance pour les quantités consommées compte tenu de toute l'information disponible.

Graphique 2. **Prédiction optimale de la consommation hebdomadaire sur 2 ans et intervalle de confiance**



La méthode permet donc d'avoir une idée de la tendance (et de l'intervalle de confiance associé qui donne l'ordre de grandeur des fluctua-

tions possibles de la série) à partir de toute l'information disponible. On notera que si l'on avait simplement conservé les indivus disponibles sur toute la période (données « matched ») on aurait sensiblement sous-estimé le niveau de la tendance.

Mise en évidence d'un choc significatif

Pour simplifier, on considère l'estimation d'une moyenne $\theta^{(t)}$ à partir d'estimateurs élémentaires construits sur M_t et U_t. La variable $\theta^{(t)}$ aura fortement augmenté (par rapport à ce que l'on attendait) si sa prédiction, compte tenu de l'information passée $\Im_{t-1}$, est très différente par rapport à sa valeur estimée compte tenu de toute l'information disponible $\Im_t$. On est donc amené à comparer $E(\theta^{(t)}|\Im_t)$ et $E(\theta^{(t)}|\Im_{t-1})$ *i.e.* la valeur lissée avec la valeur prédite. En utilisant l'équation de récurrence liée au filtrage, on peut montrer (voir annexe 3) que le test de significativité se ramène à la statistique de test

$$\frac{1}{2}\begin{pmatrix} \overline{y}_1^{(t)} - E(\overline{y}_1^{(t)}|\Im_{t-1}) \\ \overline{z}^{(t)} - E(\overline{z}^{(t)}|\Im_{t-1}) \end{pmatrix}' \Gamma^{-1} \begin{pmatrix} \overline{y}_1^{(t)} - E(\overline{y}_1^{(t)}|\Im_{t-1}) \\ \overline{z}^{(t)} - E(\overline{z}^{(t)}|\Im_{t-1}) \end{pmatrix}$$

où l'on a posé

$$\Gamma = VAR\begin{pmatrix} \overline{y}_1^{(t)} - E(\overline{y}_1^{(t)}|\Im_{t-1}) \\ \overline{z}^{(t)} - E(\overline{z}^{(t)}|\Im_{t-1}) \end{pmatrix}$$

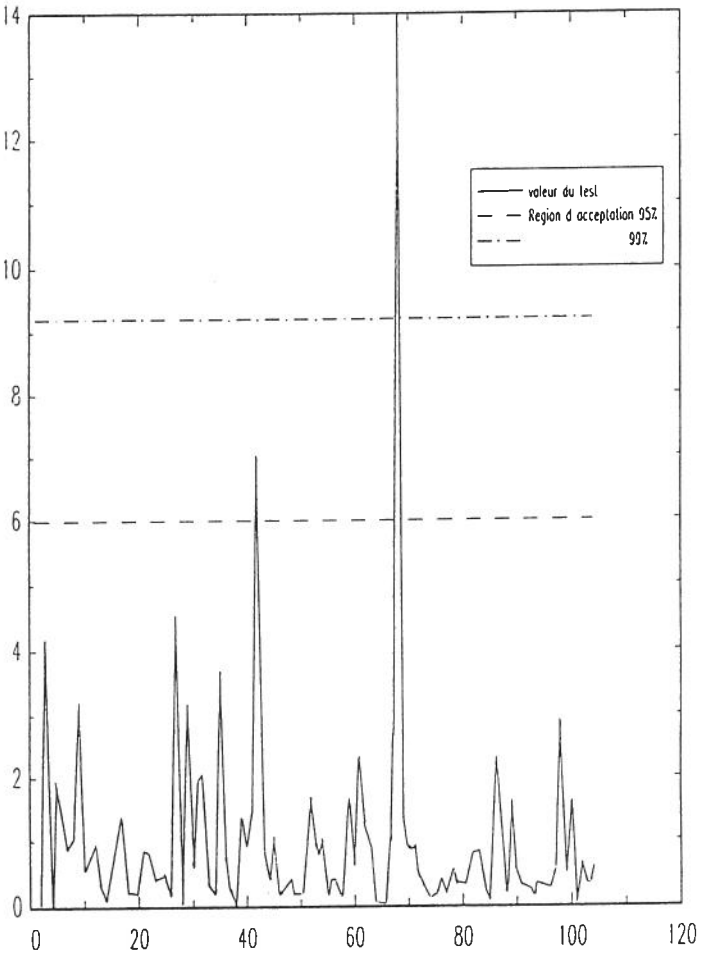
$$\overline{z}^{(t)} = \begin{cases} \overline{x}_2^{(t)} \text{ pour } t = 1, \ldots, T_1 \\ \overline{y}_2^{(t)} \text{ pour } t = T_1 + 1, \ldots, T_1 + T_2 \end{cases}$$

pour les erreurs de prédiction de $\overline{y}_1^{(t)}$ et $\overline{z}^{(t)}$. L'interprétation de cette statistique est simple : on compare en tenant compte de leur corrélation les innovations du processus des moyennes par rapport à 0. Cette statistique de test suit asymptotiquement une loi du χ^2 (2). Les prédictions $E(\overline{y}_1^{(t)}|\Im_{t-1})$ et $E(\overline{z}^{(t)}|\Im_{t-1})$ s'expriment comme des pondérations des valeurs passées $\overline{y}_1^{(u)}$ et $\overline{z}^{(u)}$, $u < t$. Toutes les quantités présentes dans cette expression sont données par des procédures d'estimation classiques par maximum de vraisemblance disponibles dans certains logiciels de séries temporelles, par exemple GAUSS.

Revenons pour finir à notre exemple illustratif. A la vue du graphique 1, il semble que la consommation d'huile ait fortement augmenté à la 16e semaine de la seconde période. Le graphique suivant donne la valeur du test précédent pour l'ensemble des valeurs. Le quantile du χ^2 à 95 % est 5,99 (9,21 à 99 %).

L'application de la méthode décrite précédemment, *i.e.* la comparaison de $E(\theta^{(t)}|\Im_t)$ et $E(\theta^{(t)}|\Im_{t-1})$, met en évidence un choc important à la 16e semaine de la seconde année et un moindre à la 42e de la première année. Il convient donc de réestimer le modèle espace-état en tenant compte de ces divers changements structurels.

Graphique 3.
Détection de chocs sur la consommation



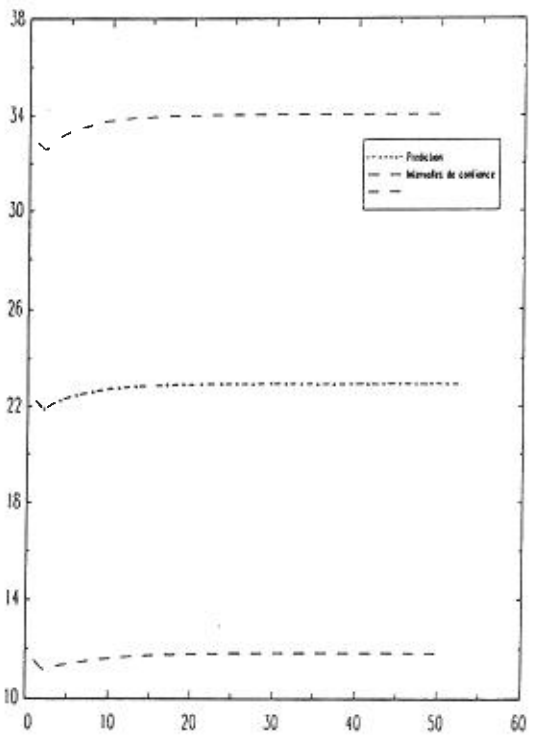
Test : présence d'un choc transitoire

L'estimation du modèle espace-état avec introduction d'une variable dummy à la 16e semaine de la deuxième année (δ_2) conduit aux estimations suivantes

Paramètres	*Estimation*	*Std*	*P – value*
μ	0,030	7,1e-4	0,00
ϕ	0,834	0,119	0,00
ρ_1	-0,225	0,112	0,05
ρ_2	-0,010	0,120	0,92
ρ_3	-0,153	0,175	0,38
σ^2	1,8e-3	0,8e-3	0,02
σ_1^2	4,7e-3	0,5e-3	0,00
σ_2^2	13,1e-3	1,3e-3	0,00
σ_3^2	9,8e-3	1,0e-3	0,00
δ_2	-1,1e-3	4,2e-3	0,78

On notera par ailleurs la relative stabilité des autres coefficients. Cette nouvelle estimation permet de prédire plus rigoureusement le comportement de la consommation sur l'année suivante :

Graphique 4.
Prédiction de la consommation sur l'année suivante



La stabilité des prédictions s'explique par le faible nombre de retards intervenant dans le modèle. Il est clair sur cet exemple que le principal intérêt des méthodes mises en œuvre ne réside pas dans les prédictions (souvent très frustes si le modèle mis en œuvre est simple et si l'on ne dispose pas de beaucoup de données). La construction de l'intervalle de confiance associé a au moins le mérite de fournir une estimation précise des écarts nécessaires pour qu'une variation de la consommation puisse être considérée comme significative.

CONCLUSION

L'objet de cet article est de rappeler les principales méthodes relatives à l'exploitation des données d'enquêtes répétées (enquêtes indépendantes, panels, panels rotatifs) dans l'optique de la théorie des sondages. Il s'agit essentiellement d'une approche d'inférence descriptive, qui n'a pas de rapport direct avec l'économétrie des données de panels, laquelle cherche à modéliser des relations temporelles entre variables et qui est

donc de type analytique. Il convient cependant de rappeler qu'ignorer un plan de sondage dans une simple régression peut avoir des effets catastrophiques sur les estimations si le plan de sondage est lié (par exemple à cause d'une stratification selon certains critères *a priori*) aux variables (exogènes ou endogènes) que l'on cherche à modéliser (on pourra à ce sujet consulter Gouriéroux, 1987, chapitre 6). Ce problème est souvent évacué en pratique en ignorant (parfois abusivement) l'aspect plan de sondage de l'enquête. Notre article se situe donc largement en amont de tels problèmes. Nous avons voulu surtout mettre l'accent sur la différence entre approche classique et approche modèle en montrant dans quelle mesure le statut des paramètres est différent entre ces deux approches. Les méthodes d'estimation relèvent essentiellement de la théories des sondages dans la première optique et plutôt des séries temporelles dans la seconde. L'approche modèle sous la forme espace-état qui nous paraît la plus intéressante dans le cadre de l'analyse de la consommation permet de prendre en compte non seulement la dynamique des erreurs d'échantillonnage mais aussi la dynamique des paramètres lorsqu'on en suppute une. Il convient alors de spécifier la forme fonctionnelle de cette dynamique : les résultats ne sont donc valides que dans le cadre de cette modélisation et ne peuvent pas être comparés avec ceux de l'approche classique. Nous avons illustré ces méthodes sur le cas de l'analyse de l'évolution de la consommation de l'huile d'olive à partir des enquêtes Secodip. Cette illustration montre comment il est possible à partir d'un panel avec rotation d'estimer la tendance de la consommation, de détecter des changements structurels et de prévoir (parfois de façon sommaire) la consommation future.

BIBLIOGRAPHIE

BAILAR (B. A.), 1989 — Information needs, surveys and measurement errors, *in : Panels Surveys*, D. KASPRZYK, G. DUNCAN, G. KALTON, M. P. SINGH (Eds.), N. Y. Wiley.

BERTAIL (P.), 1997 — Second order properties of an extrapolated bootstrap without replacement under weak assumptions : the I.I.D. and strong mixing cases, *Bernouilli*, 3, pp. 149-179.

BERTAIL (P.), COMBRIS (P.), 1997 — Bootstrap généralisé d'un sondage, une application aux données Secodip, *Annales d'Economie et de Statistiques*, 46, pp. 49-83.

BINDER (D. A.), HIDIROGLOU (M. A.), 1988 — Sampling in time, *in : Handbook of Statistics*, vol. 6, P. R. KRISHNAIA, C. R. RAO (Eds.), pp. 187-211.

BLIGHT (B. J. N.), SCOTT (A. J.), 1973 — A stochastic model for repeated surveys, *J.R.S.S.* (B), 35, pp. 61-68.

BROCKWELL (P. J.), DAVIES (R. A.), 1991 — *Times series: theory and methods*, 2nd Edition, Springer Series in Statistics, New York, Springer-Verlag.

COCHRAN (W. G.), 1977 — *Sampling Techniques*, Toronto, Wiley.

DUNCAN (G. J.), KALTON (G.), 1987 — Issues of design and analysis of surveys across time, *Int. Stat. Rev.*, 55, pp. 97-117.

DUSSAIX (A. M.), 1987a — Modèles de surpopulation, *in: Les Sondages*, chap. 4, J.J. DROESBEKE, B. FICHET, P. TASSI (Eds.), Economica.

DUSSAIX (A. M.), 1987b — Enquêtes dans le temps, *in: Les Sondages*, chap. 9, J.J. DROESBEKE, B. FICHET, P. TASSI (Eds.), Economica.

GOURIEROUX (C.), 1987 — Effets d'un sondage: cas du χ^2 et de la régression, *in: Les Sondages*, chap. 6, J.J. DROESBEKE, B. FICHET, P. TASSI (Eds.), Economica.

GURNEY (M.), DALY (J. F.), 1965 — A multivariate approach to estimation in periodic sample surveys, *Proceedings of The American Statistical Association*, pp. 247-257.

JONES (R. G.), 1980 — Best linear unbiased estimators for repeated surveys, *J.R.S.S.* (B), 42, pp. 221-226.

KALMAN (R. E.), 1960 — A new approach to linear filtering and prediction problems, *Transactions ASME Journal of Basic Engeeneering*, 82.

KALTON (G.), KASPRZYK (D.) et MCMILLEN (D. B.), 1989 — Nonsampling errors in panel surveys, *in: Panels Surveys*, D. KASPRZYK, G. DUNCAN, G. KALTON, M. P. SINGH (Eds.), N. Y. Wiley.

PATTERSON (H. D.),1950 — Sampling on successive occasions with partial replacement of units, *J.R.S.S.* (B), 12, pp. 241-255.

POLITIS (D.), ROMANO (J. P.), 1994 — Large sample confidence regions based on subsamples under minimal assumptions, *Ann. Statist.*, 22, pp. 2031-2050.

RAO (J. N. K.), SRINATH (K. P.), QUENNEVILLE (B.), 1989 — Estimation of level and change using current preliminary data, *in: Panels Surveys*, D. KASPRZYK, G. DUNCAN, G. KALTON, M. P. SINGH (Eds.), N. Y. Wiley.

ROSEN (B.), 1972 — Asymptotic theory for successive sampling with varying probabilities without replacement, *Ann. Math. Statist.*, 43, pp. 373-397.

SÄARNDAL (C. A.), SWENSSON (B.), WRETMAN (J.), 1992 — *Model Assisted Survey Sampling*, Springer Series in Statistics, New York, Springer-Verlag.

SCOTT (A. J.), SMITH (T. M. F.), 1977 — The application of time series methods to the analysis of repeated surveys, *Int. Stat. Rev.*, 45, pp. 13-28.

SEN (P. K.), 1988 — Asymptotics in finite sampling, in : *Handbook of Statistics*, vol. 6, P. R. KRISHNAIAH, C. R. RAO (Ed.), pp. 291-331.

SMITH (T. M. F.), 1978 — Principles and problems in the analysis of repeated surveys, *in: Survey Sampling and Measurement*, New York, Academic Press, pp. 201-206.

TAM (S. M.), 1987 — Analysis of repeated surveys using a dynamic linear model, *Int. Stat. Rev.*, 55, pp. 63-73.

WOLTER (K. M.), 1979 — Composite estimation in finite populations, *J.A.S.A.*, 74, pp. 604-613.

ANNEXE 1

Estimateurs élémentaires, plans de sondage et forme de la corrélation

Nous reprenons ici l'exemple 3 lorsque la structure de la population varie d'une année sur l'autre et que l'on considère un échantillon avec rotation sur deux périodes. On supposera néanmoins que la taille de la population totale ne change pas d'une période sur l'autre.

A la date $t-1$, l'échantillon $e^{(t-1)}$ est tiré selon un plan de sondage $\Pi^{(t-1)}$. On note $\pi_k^{(t-1)}$ la probabilité d'inclusion de l'individu k

$$\pi_k^{(t-1)} = \Pr\{k \in e^{(t-1)}\}$$

dans l'échantillon et $\pi_{j,k}^{(t-1)}$ la probabilité d'inclusion d'ordre 2 des individus j et k:

$$\pi_{j,k}^{(t-1)} = \Pr\{(j, k) \in e^{(t-1)}\}.$$

$\bar{e}^{(t-1)}$ désigne le complémentaire de $e^{(t-1)}$ dans la population totale. A la deuxième période un nouvel échantillon composé de deux sous-populations M_t et U_t est tiré. Le plan de sondage à la deuxième étape est donné par les plans conditionnels respectivement à $e^{(t-1)}$ et $\bar{e}^{(t-1)}$. On note $\pi_{k,m}^{(t)}$ (resp. $\pi_{j,k,m}^{(t)}$) la probabilité d'inclusion d'un individu $k \in e^{(t-1)}$ (resp. j et k) dans M_t. De la même façon on note $\pi_{k,u}^{(t)}$ (resp. $\pi_{j,k,u}^{(t)}$) la probabilité d'inclusion d'un individu $k \notin e^{(t-1)}$ (resp. j et k) dans U_t. On supposera que les plans de sondage respectivement dans $e^{(t-1)}$ et $\bar{e}^{(t-1)}$ sont indépendants. Les estimateurs élémentaires de Horvitz-Thompson s'en déduisent aisément à partir des relations sur les probabilités d'inclusion pour les plans de sondage en deux étapes (voir Särndall, Swensson, Wretman, 1992):

$$\bar{y}_2^{(t)} = N^{-1} \sum_{k \in U_t} \frac{Y_k^{(t)}}{(1 - \pi_k^{(t-1)})\pi_{k,u}^{(t)}}$$

$$\bar{y}_2^{(t)} = N^{-1} \sum_{k \in U_t} \frac{Y_k^{(t)}}{(1 - \pi_k^{(t-1)})\pi_{k,u}^{(t)}}$$

de variances respectives

$$S_{1,t}^2 = N^{-2} \sum_{k=1}^{N} \left(\frac{1}{\pi_k^{(t-1)}\pi_{k,m}^{(t)}} - 1\right) Y_k^{(t)2}$$

$$+N^{-2}\sum_{k=1}^{N}\sum_{\substack{l=1\\l\neq k}}^{N}\left(\frac{\pi_{k,l}^{(t-1)}\pi_{k,l,m}^{(t)}}{\pi_k^{(t-1)}\pi_{k,m}^{(t)}\pi_l^{(t-1)}\pi_{l,m}^{(t)}}-1\right)Y_k^{(t)}Y_l^{(t)}$$

et

$$\begin{aligned}S_{2,t}^2 &= N^{-2}\sum_{k=1}^{N}\left(\frac{1}{(1-\pi_k^{(t-1)})\pi_{k,u}^{(t)}}-1\right)Y_k^{(t)2}\\ &\quad+N^{-2}\sum_{k=1}^{N}\sum_{\substack{l=1\\l\neq k}}^{N}\left(\frac{(1-\pi_k^{(t-1)}-\pi_l^{(t-1)}+\pi_{k,l}^{(t-1)})\pi_{k,l,u}^{(t)}}{(1-\pi_k^{(t-1)})\pi_{k,u}^{(t)}(1-\pi_l^{(t-1)})\pi_{l,u}^{(t)}}-1\right)Y_k^{(t)}Y_l^{(t)}\end{aligned}$$

estimées sans biais par

$$\begin{aligned}\widehat{S}_{1,t}^2 &= N^{-2}\sum_{k\in M_t}\left(\frac{1-\pi_k^{(t-1)}\pi_{k,m}^{(t)}}{\left(\pi_k^{(t-1)}\pi_{k,m}^{(t)}\right)^2}\right)Y_k^{(t)2}\\ &\quad+N^{-2}\sum_{k\in M_t}\sum_{\substack{l\in M_t\\l\neq k}}\left(\frac{\pi_{k,l}^{(t-1)}\pi_{k,l,m}^{(t)}-\pi_k^{(t-1)}\pi_{k,m}^{(t)}\pi_l^{(t-1)}\pi_{l,m}^{(t)}}{\pi_{k,l}^{(t-1)}\pi_{k,l,m}^{(t)}\pi_k^{(t-1)}\pi_{k,m}^{(t)}\pi_l^{(t-1)}\pi_{l,m}^{(t)}}\right)Y_k^{(t)}Y_l^{(t)}\end{aligned}$$

et

$$\widehat{S}_{2,t}^2 = N^{-2}\sum_{k\in U_t}\left(\frac{1-(1-\pi_k^{(t-1)})\pi_{k,u}^{(t)}}{(1-\pi_k^{(t-1)})^2\pi_{k,u}^{(t)\,2}}\right)Y_k^{(t)2}$$

$$+N^{-2}\sum_{k\in U_t}\sum_{\substack{l\in U_t\\l\neq k}}\left(\frac{(1-\pi_k^{(t-1)}-\pi_l^{(t-1)}+\pi_{k,l}^{(t-1)})\pi_{k,l,u}^{(t)}-(1-\pi_k^{(t-1)})\pi_{k,u}^{(t)}(1-\pi_l^{(t-1)})\pi_{l,u}^{(t)}}{(1-\pi_k^{(t-1)}-\pi_l^{(t-1)}+\pi_{k,l}^{(t-1)})\pi_{k,l,u}^{(t)}(1-\pi_k^{(t-1)})\pi_{k,u}^{(t)}(1-\pi_l^{(t-1)})\pi_{l,u}^{(t)}}\right)Y_k^{(t)}Y_l^{(t)}.$$

On obtient de même

$$\overline{x}_1^{(t-1)} = N^{-1}\sum_{k\in M_t}\frac{Y_k^{(t-1)}}{\pi_k^{(t-1)}\pi_{k,m}^{(t)}}$$

$$\overline{x}_2^{(t-1)} = N^{-1}\sum_{k\in e_t-M_t}\frac{Y_k^{(t-1)}}{\pi_k^{(t-1)}(1-\pi_{k,m}^{(t)})}$$

de variances respectives

$$\begin{aligned}S_{1,t-1}^2 &= N^{-2}\sum_{k=1}^{N}\left(\frac{1}{\pi_k^{(t-1)}\pi_{k,m}^{(t)}}-1\right)Y_k^{(t-1)2}\\ &\quad+N^{-2}\sum_{k=1}^{N}\sum_{\substack{l=1\\l\neq k}}^{N}\left(\frac{\pi_{k,l}^{(t-1)}\pi_{k,l,m}^{(t)}}{\pi_k^{(t-1)}\pi_{k,m}^{(t)}\pi_l^{(t-1)}\pi_{l,m}^{(t)}}-1\right)Y_k^{(t-1)}Y_l^{(t-1)}\end{aligned}$$

et

$$S^2_{2,t-1} = N^{-2}\sum_{k=1}^{N}\left(\frac{1}{(1-\pi^{(t-1)}_{k,m})\pi^{(t)}_k}-1\right)Y_k^{(t-1)2} + N^{-2}\sum_{k=1}^{N}\sum_{\substack{l=1\\ l\neq k}}^{N}\left(\frac{(1-\pi^{(t)}_{k,m}-\pi^{(t)}_{l,m}+\pi^{(t)}_{k,l,m})\pi^{(t-1)}_{k,l}}{(1-\pi^{(t)}_{k,m})\pi^{(t-1)}_k(1-\pi^{(t)}_{l,m})\pi^{(t-1)}_l}-1\right)Y_k^{(t-1)}Y_l^{(t-1)}$$

$$\widehat{S}^2_{1,t} = N^{-2}\sum_{k\in M_t}\left(\frac{1-\pi^{(t-1)}_k\pi^{(t)}_{k,m}}{\left(\pi^{(t-1)}_k\pi^{(t)}_{k,m}\right)^2}\right)Y_k^{(t-1)2} + N^{-2}\sum_{k\in M_t}\sum_{\substack{l\in M_t\\ l\neq k}}\left(\frac{1}{\pi^{(t-1)}_k\pi^{(t)}_{k,m}\pi^{(t-1)}_l\pi^{(t)}_{l,m}}-\frac{1}{\pi^{(t-1)}_{k,l}\pi^{(t)}_{k,l,m}}\right)Y_k^{(t-1)}Y_l^{(t-1)}$$

et

$$\widehat{S}^2_{2,t} = N^{-2}\sum_{k\in e_t-M_t}\left(\frac{1-(1-\pi^{(t-1)}_{k,m})\pi^{(t)}_k}{(1-\pi^{(t-1)}_{k,m})^2\pi^{(t)2}_k}\right)Y_k^{(t-1)2}$$

$$+N^{-2}\sum_{k\in e_t-M_t}\sum_{\substack{l\in e_t-M_t\\ l\neq k}}\left(\frac{1}{(1-\pi^{(t)}_{k,m})\pi^{(t-1)}_k(1-\pi^{(t)}_{l,m})\pi^{(t-1)}_l} - \frac{1}{(1-\pi^{(t)}_{k,m}-\pi^{(t)}_{l,m}+\pi^{(t)}_{k,l,m})\pi^{(t-1)}_{k,l}}\right)Y_k^{(t-1)}Y_l^{(t-1)}\ .$$

$$cy_{1,2} = cov_\Pi(\overline{y}^{(t)}_1,\overline{y}^{(t)}_2) = N^{-2}\sum_{k=1}^{N}\sum_{\substack{l=1\\ l\neq k}}^{N}\left(\frac{\pi^{(t-1)}_k\pi^{(t-1)}_l-\pi^{(t-1)}_{k,l}}{\pi^{(t-1)}_k(1-\pi^{(t-1)}_l)}\right)Y_k^{(t)}Y_l^{(t)}$$

$$cx_{1,2} = cov_\Pi(\overline{x}^{(t-1)}_1,\overline{x}^{(t-1)}_2) = N^{-2}\sum_{k=1}^{N}\sum_{\substack{l=1\\ l\neq k}}^{N}\left(\frac{\pi^{(t-1)}_{k,l}(\pi^{(t)}_{k,m}-\pi^{(t)}_{k,l,m})}{\pi^{(t-1)}_k\pi^{(t-1)}_l\pi^{(t)}_{k,m}(1-\pi^{(t)}_{l,m})}-1\right)Y_k^{(t-1)}Y_l^{(t-1)}$$

$$cxy_{1,1} = cov_\Pi(\overline{x}^{(t-1)}_1,\overline{y}^{(t)}_1) = N^{-2}\sum_{k=1}^{N}\sum_{l=1}^{N}\left(\frac{\pi^{(t-1)}_{k,l}\pi^{(t)}_{k,l,m}}{\pi^{(t-1)}_k\pi^{(t-1)}_l\pi^{(t)}_{k,m}\pi^{(t)}_{l,m}}-1\right)Y_k^{(t-1)}Y_l^{(t)}$$

$$cxy_{1,2} = cov_\Pi(\overline{x}^{(t-1)}_1,\overline{y}^{(t)}_2) = N^{-2}\sum_{k=1}^{N}\sum_{\substack{l=1\\ l\neq k}}^{N}\frac{\pi^{(t-1)}_k\pi^{(t-1)}_l-\pi^{(t-1)}_{k,l}}{\pi^{(t-1)}_k(1-\pi^{(t-1)}_l)}Y_k^{(t-1)}Y_l^{(t)}$$

$$cxy_{2,1} = cov_\Pi(\bar{x}_2^{(t-1)}, \bar{y}_1^{(t)}) = N^{-2} \sum_{k=1}^{N} \sum_{\substack{l=1 \\ l \neq k}}^{N} \left(\frac{\pi_{k,l}^{(t-1)}(\pi_{l,m}^{(t)} - \pi_{k,l,m}^{(t)})}{\pi_k^{(t-1)}(1-\pi_{k,m}^{(t)})\pi_l^{(t-1)}\pi_{l,m}^{(t)}} - 1 \right) Y_k^{(t)} Y_l^{(t-1)}$$

$$cxy_{2,2} = cov_\Pi(\bar{x}_2^{(t-1)}, \bar{y}_2^{(t)}) = N^{-2} \sum_{k=1}^{N} \sum_{\substack{l=1 \\ l \neq k}}^{N} \frac{\pi_k^{(t-1)}\pi_l^{(t-1)} - \pi_{k,l}^{(t-1)}}{\pi_k^{(t-1)}(1-\pi_l^{(t-1)})} Y_k^{(t-1)} Y_l^{(t)}$$

avec la convention que $\pi_{k,k}^{(t-1)} = \pi_k^{(t-1)}$ et $\pi_{k,k,m}^{(t)} = \pi_{k,k}^{(t)}$ pour les redoublements. On notera qu'aucune de ces covariances ne dépend du plan de sondage (conditionnel à $\bar{e}^{(t)}$) utilisé pour obtenir U_t. Seul l'estimateur $\bar{y}_2^{(t)}$ et sa variance en dépendent.

Chacune de ces quantités est estimée sans biais par

$$\widehat{cy_{1,2}} = N^{-2} \sum_{k \in M_t} \sum_{l \in U_t} \left(\frac{\pi_k^{(t-1)}\pi_l^{(t-1)} - \pi_{k,l}^{(t-1)}}{(\pi_k^{(t-1)} - \pi_{k,l}^{(t-1)})\pi_{k,m}^{(t)}\pi_{l,u}^{(t)}\pi_k^{(t-1)}(1-\pi_l^{(t-1)})} \right) Y_k^{(t)} Y_l^{(t)}$$

$$\widehat{cx_{1,2}} = N^{-2} \sum_{k \in M_t} \sum_{l \in e_t - M_t} \left(\frac{1}{\pi_k^{(t-1)}\pi_l^{(t-1)}\pi_{k,m}^{(t)}(1-\pi_{l,m}^{(t)})} - \frac{1}{\pi_{k,l}^{(t-1)}(\pi_{k,m}^{(t)} - \pi_{k,l,m}^{(t)})} \right) Y_k^{(t-1)} Y_l^{(t-1)}$$

$$\begin{aligned} \widehat{cxy_{1,1}} &= N^{-2} \sum_{k \in M_t} \left(\frac{1-\pi_k^{(t-1)}\pi_{k,m}^{(t)}}{\pi_k^{(t-1)2}\pi_{k,m}^{(t)2}} \right) Y_k^{(t-1)} Y_k^{(t)} \\ &\quad + N^{-2} \sum_{k \in M_t} \sum_{\substack{l \in M_t \\ l \neq k}} \left(\frac{1}{\pi_k^{(t-1)}\pi_l^{(t-1)}\pi_{k,m}^{(t)}\pi_{l,m}^{(t)}} - \frac{1}{\pi_{k,l}^{(t-1)}\pi_{k,l,m}^{(t)}} \right) Y_k^{(t-1)} Y_l^{(t)} \end{aligned}$$

$$\widehat{cxy_{1,2}} = N^{-2} \sum_{k \in M_t} \sum_{l \in U_t} \frac{\pi_k^{(t-1)}\pi_l^{(t-1)} - \pi_{k,l}^{(t-1)}}{(\pi_k^{(t-1)} - \pi_{k,l}^{(t-1)})\pi_{k,m}^{(t)}\pi_{l,u}^{(t)}\pi_k^{(t-1)}(1-\pi_l^{(t-1)})} Y_k^{(t-1)} Y_l^{(t)}$$

$$\widehat{cxy_{2,1}} = N^{-2} \sum_{k \in e_t - M_t} \sum_{l \in M_t} \left(\frac{1}{\pi_k^{(t-1)}(1-\pi_{k,m}^{(t)})\pi_l^{(t-1)}\pi_{l,m}^{(t)}} - \frac{1}{\pi_{k,l}^{(t-1)}(\pi_{l,m}^{(t)} - \pi_{k,l,m}^{(t)})} \right) Y_k^{(t)} Y_l^{(t-1)}$$

$$\widehat{cxy_{2,2}} = N^{-2} \sum_{k \in e_t - M_t} \sum_{l \in U_t} \frac{\pi_k^{(t-1)}\pi_l^{(t-1)} - \pi_{k,l}^{(t-1)}}{(\pi_k^{(t-1)} - \pi_{k,l}^{(t-1)})(1-\pi_{k,m}^{(t)})\pi_{l,u}^{(t)}\pi_k^{(t-1)}(1-\pi_l^{(t-1)})} Y_k^{(t-1)} Y_l^{(t)} .$$

Le modèle sur deux périodes s'écrit

$$\begin{pmatrix} \bar{x}_1^{(t-1)} \\ \bar{x}_2^{(t-1)} \\ \bar{y}_1^{(t)} \\ \bar{y}_2^{(t)} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \theta^{(t-1)} \\ \theta^{(t)} \end{pmatrix} + \begin{pmatrix} \bar{\eta}_1^{(t-1)} \\ \bar{\eta}_2^{(t-1)} \\ \bar{\eta}_1^{(t)} \\ \bar{\eta}_1^{(t)} \end{pmatrix} = A\theta + \eta$$

avec

$$Var_{\Pi}(\eta) = \begin{pmatrix} S^2_{1,t-1} & cx_{1,2} & cxy_{1,1} & cxy_{1,2} \\ cx_{1,2} & S^2_{2,t-1} & cxy_{2,1} & cxy_{2,2} \\ cxy_{1,1} & cxy_{2,1} & S^2_{1,t} & cy_{1,2} \\ cxy_{1,2} & cxy_{2,2} & cy_{1,2} & S^2_{2,t} \end{pmatrix}$$

Il suffit alors d'appliquer les moindres carrés généralisés pour obtenir la forme de l'estimateur optimal. Un estimateur est donné en remplaçant les variances et les covariances par leurs estimateurs respectifs : on remarquera cependant que la propriété d'absence de biais de l'estimateur ainsi obtenu est perdue à distance finie dans cette opération

Cas particulier 1 : Sondages poissoniens

Si les sondages qui conduisent respectivement à $e^{(t)}$ et à M_t sont poissoniens, on a alors

$$\pi_k^{(t-1)}\,\pi_l^{(t-1)} - \pi_{k,l}^{(t-1)} = 0$$
$$\pi_{k,m}^{(t)}\,\pi_{l,m}^{(t)} - \pi_{k,l,m}^{(t)} = 0$$

(cette condition est souvent asymptotiquement vérifiée pour de nombreux types de sondages à probabilités inégales) et la variance de η devient

$$Var_{\Pi}(\eta) = \begin{pmatrix} S^2_{1,t-1} & 0 & cxy_{1,1} & 0 \\ 0 & S^2_{2,t-1} & 0 & 0 \\ cxy_{1,1} & 0 & S^2_{1,t} & 0 \\ 0 & 0 & 0 & S^2_{2,t} \end{pmatrix}$$

Les expressions se simplifient en

$$S^2_{1,t} = N^{-2}\sum_{k=1}^{N}\left(\frac{1}{\pi_k^{(t-1)}\pi_{k,m}^{(t)}} - 1\right)Y_k^{(t)2}$$

$$S^2_{2,t} = N^{-2}\sum_{k=1}^{N}\left(\frac{1}{(1-\pi_k^{(t-1)})\pi_{k,u}^{(t)}} - 1\right)Y_k^{(t)2}$$

$$S^2_{1,t-1} = N^{-2}\sum_{k=1}^{N}\left(\frac{1}{\pi_k^{(t-1)}\pi_{k,m}^{(t)}} - 1\right)Y_k^{(t-1)2}$$

et

$$S^2_{2,t-1} = N^{-2}\sum_{k=1}^{N}\left(\frac{1}{(1-\pi_{k,m}^{(t-1)})\pi_k^{(t)}} - 1\right)Y_k^{(t-1)2}$$

$$cxy_{1,1} = cov(\overline{x}_1^{(t-1)}, \overline{y}_1^{(t)}) = N^{-2}\sum_{k=1}^{N}\left(\frac{1}{\pi_k^{(t-1)}\pi_{k,m}^{(t)}} - 1\right)Y_k^{(t-1)}Y_k^{(t-1)}\,.$$

Dans ce cas, le coefficient de corrélation est donné par

$$\rho = \frac{\sum_{k=1}^{N} w_k Y_k^{(t-1)} Y_k^{(t)}}{\left(\sum_{k=1}^{N} w_k Y_k^{(t)2}\right)^{1/2} \left(\sum_{k=1}^{N} w_k Y_k^{(t-1)2}\right)^{1/2}}$$

avec

$$w_k = \left(\pi_k^{(t-1)^{-1}} \pi_{k,m}^{(t)^{-1}} - 1\right).$$

ρ s'interprète donc ici comme une version pondérée du coefficient de corrélation entre les valeurs à la date t et les valeurs à la date $t-1$. Sous des conditions de régularité ad-hoc, un estimateur asymptotiquement convergent de ρ est donné par

$$\hat{\rho} = \frac{\sum_{k \in M_t} \tilde{w}_k Y_k^{(t-1)} Y_k^{(t)}}{\left(\sum_{k \in M_t} \tilde{w}_k Y_k^{(t)2}\right)^{1/2} \left(\sum_{k \in M_t} \tilde{w}_k Y_k^{(t-1)2}\right)^{1/2}}$$

avec

$$\tilde{w}_k = \pi_k^{(t-1)^{-1}} \pi_{k,m}^{(t)^{-1}} \left(\pi_k^{(t-1)^{-1}} \pi_{k,m}^{(t)^{-1}} - 1\right).$$

Cas particulier 2: Panels

Dans le cas du panel, on ne considère que les estimateurs indicés par 1 et on a par définition

$$\pi_{k,m}^{(t)} = \pi_{k,l,m}^{(t)} = 1$$

les expressions des estimateurs s'en déduisent immédiatement et ne dépendent que de $\pi_k^{(t-1)}$ et $\pi_{k,l}^{(t-1)}$.

ANNEXE 2

Modèles espace-état (état-mesure) et filtres de Kalman

Nous rappelons ici quelques éléments sur les filtres de Kalman. Nous renvoyons à Brockwell et Davies (1991), dont nous avons emprunté les notations, pour plus de précisions. On s'intéresse au modèle défini par

Equation d'espace (ou : de mesure) $Y_t = G_t X_t + e_t, \quad t = 1,\ldots,T$

Equation d'état : $X_t = F_t X_{t-1} + V_t, t = 1,\ldots,T$

avec les hypothèses suivantes :

H_1 : $F_1,\ldots, F_T$ est une suite de matrices déterministes.

H_2 : $G_1,\ldots, G_T$ est une suite de matrices déterministes.

H_3 : $u_t = (X_t', (V_t', e_t')')$ $t = 1,\ldots,T$ est une suite de variables aléatoires de moments d'ordre au moins deux, orthogonales (au sens où $cov(u_t, u_{t+h}) = 0$ pour $h > 0$) avec $EV_t = 0$ et $Ee_t = 0$.

H_4 : $EV_tV_t' = Q_t$, $Ee_t e_t' = R_t$, $EV_t e_t' = S_t$ où Q_t, R_t et S_t désignent des matrices déterministes.

H_5 : Pour tout t, X_t et $(V_t', e_t')'$ sont non corrélées.

On dira que le modèle est gaussien si l'hypothèse suivante est en outre vérifiée.

H_6 : (V_t', e_t') est un vecteur gaussien dont la matrice de variance-covariance est donnée par H_4

Nous nous intéressons ici essentiellement aux problèmes de prédiction, de filtrage et de lissage et donnons ici quelques résultats fondamentaux.

Théorème 1 (Prédiction)

Sous $H_1 - H_5$, $\hat{X}_{t-1}^{(t)}$ et Ω_t sont déterminées de manière unique par les conditions initiales

$$\begin{aligned} \hat{X}_0^{(1)} &= E(X_0|Y_0) \\ \Pi_1 &= E(X_1X_1') \\ \Psi_1 &= E(\hat{X}_0^{(1)}\hat{X}_0^{(1)\prime}) \\ \Omega_1 &= \Pi_1 - \Psi_1 \end{aligned}$$

et l'équation de remise à jour

$$\hat{X}_t^{(t+1)} = F_t\hat{X}_{t-1}^{(t)} + \Theta_t\Delta_t^{-1}(Y_t - G_t\hat{X}_{t-1}^{(t)})$$

où Θ_t, Δ_t et Ω_t sont définies par les équations de récursion

$$\begin{aligned}\Delta_t &= G_t\Omega_t G_t' + R_t \\ \Theta_t &= F_t\Omega_t F_t' + S_t \\ \Omega_{t+1} &= \Pi_{t+1} - \Psi_{t+1}\end{aligned}$$

avec

$$\begin{aligned}\Pi_{t+1} &= F_t\Pi_t F_t' + Q_t \\ \Psi_{t+1} &= F_t\Psi_t F_t' + \Theta_t\Delta_t^{-1}\Theta_t'\end{aligned}$$

où Δ_t^{-1} désigne une inverse généralisée de Δ_t.

On prend en général, $Y_0 = 1$, de sorte que $\hat{X}_0^{(t)}$ n'est autre que $E(X_0)$.

Théorème 2 (Filtrage)

L'estimateur optimal de X_t connaissant $\Im_t$, à savoir $\hat{X}_t^{(t)}$ et la matrice de variance covariance de l'erreur associée

$$\Omega_t^{(t)} = E(X_t - \hat{X}_t^{(t)})(X_t - \hat{X}_t^{(t)})'$$

sont donnés par les relations de mise à jour

$$\hat{X}_t^{(t)} = \hat{X}_{t-1}^{(t)} + \Omega_t G_t'\Delta_t^{-1}(Y_t - G_t\hat{X}_{t-1}^{(t)})$$

$$\Omega_t^{(t)} = \Omega_t - \Omega_t G_t'\Delta_t^{-1}G_t\Omega_t' .$$

Théorème 3 (Lissage)

L'estimateur optimal de X_t connaissant $\Im_n$ pour $n > t$, à savoir $\hat{X}_n^{(t)}$ et la matrice de variance-covariance de l'erreur associée

$$\Omega_n^{(t)} = E(X_t - \hat{X}_n^{(t)})(X_t - \hat{X}_n^{(t)})'$$

sont donnés par les relations de mise à jour

$$\hat{X}_n^{(t)} = \hat{X}_{n-1}^{(t)} + \Omega_{t,n} G_n'\Delta_n^{-1}(Y_n - G_n\hat{X}_{n-1}^{(t)})$$

$$\Omega_n^{(t)} = \Omega_{n-1}^{(t)} - \Omega_{t,n} G_n'\Delta_n^{-1}G_n\Omega_{t,n}'$$

avec

$$\Omega_{t,n} = \Omega_{t,n-1} - \Omega_{t,n} G_n'\Delta_n^{-1}G_n\Omega_{t,n}'$$

et la condition initiale

$$\Omega_{t,t} = \Omega_t^{(t)} = \Omega_t.$$

ANNEXE 3

Test de présence d'un choc

En reprenant les notations de l'annexe 2 et le modèle espace-état vu dans la dernière partie, on a d'après l'équation de filtrage (cf. annexe 2)

$$\begin{aligned} Diff_t &\doteq E(\theta^{(t)}|\Im_t) - E(\theta^{(t)}|\Im_{t-1}) = e'_\Theta(X_t^{(t)} - X_{t-1}^{(t)}) \\ &= e'_\Theta \Omega_t G'_t \Delta_t^{-1} I_t \end{aligned}$$

où

$$I_t = (Y_t - G_t X_{t-1}^{(t)})$$

est appelée l'innovation et où e'_Θ désigne le vecteur de projection sur la 1re composante

$$e'_\Theta = (1, 0, 0, 0)$$

Comme, dans notre cas, $R_t = Var(e_t) = 0$ on en déduit que

$$\Delta_t = G_t \Omega_t G'_t$$

et donc

$$Diff_t = e'_\Theta \Omega_t G'_t (G_t \Omega_t G'_t)^{-1} I_t .$$

Sa variance est donnée par

$$e'_\Theta \Omega_t G'_t (G_t \Omega_t G'_t)^{-1} Var(I_t)(G_t \Omega_t G'_t)^{-1} G_t \Omega_t e_\Theta .$$

Mais on a (voir Brockwell and Davies, 1991, p. 477)

$$Var(I_t) = \Delta_t = G_t \Omega_t G'_t$$

on en déduit par un calcul classique de trace

$$\begin{aligned} Var(Diff_t) &= e'_\Theta \Omega_t G'_t (G_t \Omega_t G'_t)^{-1} G_t \Omega_t e_\Theta \\ &= Tr\left(\Omega_t G'_t (G_t \Omega_t G'_t)^{-1} G_t \Omega_t e_\Theta e'_\Theta\right) \\ &= \omega_t^2 Tr\left(\Omega_t G'_t (G_t \Omega_t G'_t)^{-1} G_t\right) \\ &= \omega_t^2 Tr\left((G_t \Omega_t G'_t)^{-1} G_t \Omega_t G'_t\right) \\ &= \omega_t^2 Tr(I_2) = 2\omega_t^2 \end{aligned}$$

où

$$\omega_t^2 = Var(\theta^{(t)}|\Im_{t-1}) .$$

D'autre part, on a

$$\begin{aligned} Diff_t^2 &= e'_\Theta \Omega_t G'_t (G_t \Omega_t G'_t)^{-1} I_t I'_t (G_t \Omega_t G'_t)^{-1} G_t \Omega_t e_\Theta \\ &= tr(I_t I'_t (G_t \Omega_t G'_t)^{-1} G_t \Omega_t e_\Theta e'_\Theta \Omega_t G'_t (G_t \Omega_t G'_t)^{-1}) \\ &= \omega_t^2 tr(I_t I'_t (G_t \Omega_t G'_t)^{-1}) \\ &= \omega_t^2 tr(I'_t (G_t \Omega_t G'_t)^{-1} I_t) \end{aligned}$$

d'où l'on déduit

$$Diff_t^2 / Var(Diff) = I'_t (G_t \Omega_t G'_t)^{-1} I_t / 2 \qquad (7.1)$$

statistique qui suit asymptotiquement une loi du χ^2 (1) (car Diff. est asymptotiquement gaussienne). Dans le cas particulier qui nous intéresse, en posant

$$\overline{z}^{(t)} = \begin{cases} \overline{x}_2^{(t)} \text{ pour } t = 1, \ldots, T_1 \\ \overline{y}_2^{(t)} \text{ pour } t = T_1 + 1, \ldots, T_1 + T_2 \end{cases}$$

et

$$er_1^{(t)} = \overline{y}_1^{(t)} - E(\overline{y}_1^{(t)} | \Im_{t-1})$$

$$er_2^{(t)} = \overline{z}^{(t)} - E(\overline{z}^{(t)} | \Im_{t-1})$$

pour les erreurs de prédiction de $\overline{y}_1^{(t)}$ et $\overline{z}^{(t)}$, la statistique se ramène à

$$\frac{1}{2} \begin{pmatrix} er_1^{(t)} \\ er_2^{(t)} \end{pmatrix}' \begin{pmatrix} Var(er_1^{(t)}) & cov(er_1^{(t)}, er_2^{(t)}) \\ cov(er_1^{(t)}, er_2^{(t)}) & Var(er_2^{(t)}) \end{pmatrix}^{-1} \begin{pmatrix} er_1^{(t)} \\ er_2^{(t)} \end{pmatrix}.$$