ESTIMATIONS DANS L'ENQUÊTE EMPLOI EN CONTINU

J. BOSREDON[*] *et P. FEVRIER*[**]

[*] INSEE - Division Emploi

[**] INSEE - Unité "Méthodes Statistiques"

1. Introduction

La mise au point des méthodes d'estimation est un des éléments centraux de la refonte de l'enquête emploi i.e du passage à partir de 2002 d'une enquête annuelle à une enquête en continu sur l'année.

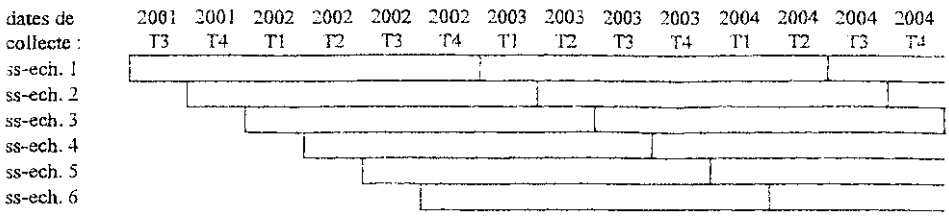
Il existe deux grandes approches en terme d'estimation pour les enquêtes répétées : l'approche dite « classique » et l'approche dite « série temporelle ». Dans la première, la séquence des paramètres d'intérêt (en général des moyennes ou des totaux sur la population) est supposée être une quantité fixe inconnue. Dans la seconde, les paramètres d'intérêt sont supposés être des quantités aléatoires dont on peut décrire l'évolution par une série temporelle.

Nous nous limiterons dans ce papier à l'estimation d'un niveau moyen mais les techniques présentées se généralisent à d'autres objectifs tels que ceux décrits par Kalton et Duncan (1987) : estimation des évolutions, mesure des différentes composantes de ces évolutions, calcul de grandeurs au niveau individuel...

Dans une première partie, nous décrirons les caractéristiques de l'enquête emploi en continu. La deuxième partie sera consacrée à une revue de la littérature concernant l'approche « classique ». La troisième partie se concentrera sur la littérature concernant l'approche série temporelle. Nous décrirons en particulier l'apport que représente le modèle espace d'état et son estimation par le filtre de Kalman. La dernière partie étudiera le problème de la désaisonnalisation et son inclusion dans les modèles précédents.

2. Description de l'enquête

L'enquête emploi en continu utilisera un schéma de rotation « simple » [1]: les ménages sont interrogés 6 trimestres consécutifs puis remplacés. On dispose ainsi de 6 sous-échantillons de type identique, et dont les plannings d'interrogation sont décalés d'un trimestre les uns par rapport aux autres.



Au cours de chaque trimestre, chaque ménage sera interrogé sur une période définie à l'avance, l'ensemble de ces périodes couvrant uniformément tout le trimestre. Ils seront réinterrogés exactement 13 semaines plus tard. Le nombre de ménages désignés chaque trimestre est de l'ordre de 40 000.

Ce mode d'organisation présente en fait de nombreux points communs avec celui qui prévaut pour l'enquête emploi annuelle comme le montre le tableau suivant. Cependant l'étalement de la collecte sur l'ensemble de l'année a un impact extrêmement important sur le type de résultats que l'on pourra produire, et sur les méthodes d'exploitation qui seront nécessaires.

	Enquête emploi annuelle	Enquête emploi en continu
Type d'enquête	aréolaire (aires de 20 ou 40 logts)	aréolaire (aires de 20 logts)
Mode de collecte	en visite	en visite et par téléphone
	sous Capi	sous Capi
Période de collecte	tout le mois de mars	toute l'année
Taille des échantillons	100 000 logements par an	50 000 logements par trimestre
Renouvellement	par tiers tous les ans	par sixième tous les trimestres

[1] Ce schéma est généralement appelé schéma de rotation à un niveau. Il existe d'autres possibilités dans lesquelles, par exemple, les ménages sont interrogés deux trimestres consécutifs, pas les deux trimestres suivants, puis de nouveau 2 trimestres consécutifs (schéma 2-2-2)

L'échantillon utilisé, comme précédemment, sera aréolaire. Il est déterminé une fois pour toute, pour une durée d'environ 8 ans, à partir des résultats du recensement de 1999. Cet échantillon a été stratifié par région et par tranche d'urbanisation (rural, unités urbaines de moins de 10 000 habitants, de 10 000 à 50 000 habitants, de 50 000 à 200 000 habitants, de plus de 200 000 habitants) pour permettre de publier plus facilement des résultats sur ces catégories et pour des raisons de précision. Ces deux objectifs ont été conciliés en calculant des tailles d'échantillon dans chaque strate

- qui optimisent la précision de l'estimateur national du taux de chômage

- sous la contrainte que la précision des taux de chômage régionaux reste supérieure à un seuil de l'ordre de 8 %[2]

- sous la contrainte budgétaire (celle-ci se traduisant en pratique par la limitation de l'échantillon interrogeable chaque trimestre à 40 000 ménages).

A l'intérieur de chaque strate, on a eu recours à des procédures de découpage aboutissant au tirage d'aires dont la taille au recensement était de 20 logements en moyenne. La probabilité finale de tirage d'un logement donné existant au RP est uniforme dans chacune des strates. L'échantillon ainsi défini reste néanmoins un échantillon géographique qui inclut toutes les zones non-bâties. Au moment de l'enquête, l'ensemble des logements se situant à l'intérieur du territoire définissant les aires, qu'ils aient été construits antérieurement ou postérieurement au RP, seront interrogés.

Pour décrire les méthodes d'estimation applicable à l'enquête emploi en continu, nous considérons, dans la suite du papier, le cas plus général d'une enquête répétée à échantillon rotatif à un niveau. Nous cherchons à estimer à chaque date $t \leq T$ une caractéristique inconnue θ_t de la population à partir des observations individuelles $\left(Y_t^i\right)_{t=1\ldots T, i=1\ldots n}$ issues du sondage. n est le nombre d'unités interrogées à chaque date (n est égal à 40000 pour l'enquête emploi en continu et égal à 80000 dans l'enquête emploi annuelle).

Chacune de ces unités est interrogée durant L périodes consécutives et chaque échantillon est donc composé de L vagues (L est égal à 6 pour l'enquête emploi en continu et égal à 3 pour l'enquête emploi annuelle). Nous appellerons estimateurs élémentaires (notion introduite par Gurney et Daly, 1965), notés $\left(Y_{lt}\right)_{t=1\ldots T, l=1\ldots L}$, les estimateurs de θ_t supposés sans biais, construits pour chaque vague l.

[2] Ce chiffre est celui qui figure dans le règlement n°577/98 de l'Union européenne

On obtient alors :

$$Y_{lt} = \theta_t + \xi_{lt}$$

ξ_{lt} est l'erreur d'échantillonnage et on suppose que $E(\xi_{lt}) = 0,\ V(\xi_{lt}) = S^2$ qui ne dépend pas du temps. On supposera également que les Y_{lt} sont indépendants conditionnellement à θ_t, ce qui se justifie par la faible taille de l'échantillon par rapport à la population.

3. Approche « classique »

Dans l'approche « classique » (Jessen, 1942; Patterson, 1950; Gurney et Daly, 1965), la séquence des paramètres d'intérêt $\{\theta_t\}_{t\leq T}$ est supposée être une séquence fixe inconnue. Les variables aléatoires sont les variables Y_{lt} et on suppose qu'il existe une structure de corrélation entre ces variables pour un même individu, à deux dates différentes.

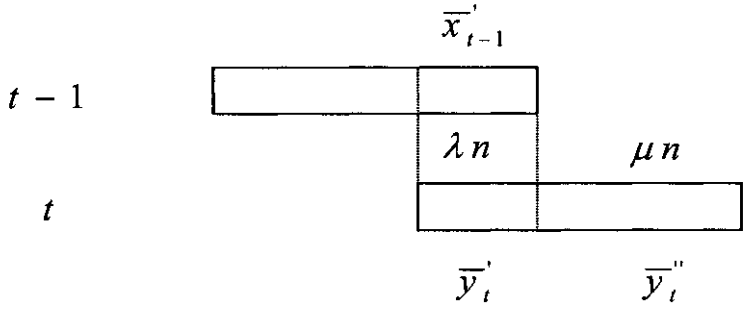
3.1 Le modèle de Patterson (1950)

Patterson spécifie la forme de la corrélation entre deux dates successives par :

$$Y_{lt} - \theta_t = \rho(Y_{(l-1)(t-1)} - \theta_{t-1}) + \eta_{lt}$$

Les η_{lt} sont supposés non corrélés, avec $E(\eta_{lt}) = 0,\ V(\eta_{lt}) = (1-\rho^2)S^2$.

Il note $\bar{x}'_{t-1}$ la moyenne des $Y_{l(t-1)}$ sur le sous-échantillon constitué par les L-1 vagues interrogées aux dates $t-1$ et t, $\bar{y}'_t$ la moyenne des Y_{lt} sur le même ensemble et $\bar{y}''_t = Y_{1t}$ la moyenne des Y_t^i sur le sous-échantillon constitué par les unités qui entrent dans l'échantillon à la date t i.e la vague 1. Il note également $\mu = \frac{1}{L}$ la proportion d'unités interrogées pour la première fois et $\lambda = 1 - \mu = \frac{L-1}{L}$ la proportion d'unités interrogées à deux dates consécutives.



Patterson montre alors que $\theta_{T|T}$, le meilleur estimateur linéaire sans biais de la variable θ_T à la date T, conditionnellement aux observations $\{Y_{lt}\}_{l=1...L,t=1...T}$, est défini récursivement grâce à la formule suivante :

$$\theta_{t|t} = (1-\varphi_t)\left\{\overline{y}'_t + \rho(\theta_{t-1|t-1} - \overline{x}'_{t-1})\right\} + \varphi_t \overline{y}''_t$$

$$1-\varphi_t = \frac{\lambda}{1-(\mu-\lambda)\rho^2 - \lambda\rho^2(1-\varphi_{t-1})}$$

$$V(\theta_{t|t}) = \frac{\varphi_t S^2}{\mu n}$$

L'estimateur $\theta_{t|t}$ est la somme pondérée de deux estimateurs de θ_t :

- $\overline{y}''_t$ qui correspond à l'estimation sur les unités qui rentrent dans l'échantillon

- $\overline{y}'_t + \rho(\theta_{t-1|t-1} - \overline{x}'_{t-1})$ qui correspond à l'estimation sur les autres unités. Cette estimation consiste à corriger $\overline{y}'_t$ par le biais de représentativité de l'échantillon à la date précédente $\theta_{t-1|t-1} - \overline{x}'_{t-1}$, en utilisant la structure de corrélation.

3.2 La généralisation de Gurney et Daly (1965)

Gurney et Daly généralise le résultat précédent car ils ne spécifient pas la forme de la corrélation entre Y_{lt} et $Y_{(l-1)(t-1)}$.

- l'estimateur optimal

Le modèle $Y_{lt} = \theta_t + \xi_{lt}$ s'écrit vectoriellement en notant $Y^T = (Y_{lt})_{l=1,...,L;t=1,..,T}$, $\xi^T = (\xi_{lt})_{l=1,...,L;t=1,..,T}$ et $\theta^T = (\theta_t)_{t=1,..,T}$:

$$Y^T = A\theta^T + \xi^T$$

où A est une matrice de 0 et de 1 bien choisie. On note Σ la matrice de variance covariance de ξ^T.

L'application des moindres carrés généralisés permettent d'obtenir $\theta^{T|T}$, le meilleur estimateur linéaire sans biais de θ^T :

$$\theta^{T|T} = (A'\Sigma^{-1}A)^{-1}A'\Sigma^{-1}Y^T$$
$$V(\theta^{T|T}) = (A'\Sigma^{-1}A)^{-1}$$

Lorsque Σ est spécifiée pour correspondre aux hypothèses du modèle précédent ($Y_{lt} - \theta_t = \rho(Y_{(l-1)(t-1)} - \theta_{t-1}) + \eta_{lt}$), on retrouve le résultat de Patterson. Dans le cas général, le calcul de ces estimateurs optimaux peuvent poser des problèmes lorsque le nombre de périodes est élevé en raison de l'instabilité des procédures numériques d'inversion de matrices de grande taille. Pour remédier à ce problème, des estimateurs dits composites ont été développés.

- les estimateurs composites

Ces estimateurs utilisent les corrélations entre les dates t et t-1 sur la partie commune de l'échantillon pour améliorer l'estimateur à la date t.

L'estimateur AK est défini par Gurney et Daly de la façon suivante :

$$\theta_{t|t} = \frac{1}{L}\left[(1-K+A)Y_{1t} + (1-K-\frac{1}{L-1}A)\sum_{l=2}^{L} Y_{lt}\right] + K\left\{\theta_{t-1|t-1} + \bar{y}_t^{'} - \bar{x}_{t-1}^{'}\right\}$$

Une version simplifiée, l'estimateur K, s'écrit :

$$\theta_{t|t} = (1-K)\bar{y}_t + K\left\{\theta_{t-1|t-1} + \bar{y}_t' - \bar{x}_{t-1}'\right\}$$

Le calcul des variances de l'estimateur K dans le cadre de l'enquête emploi en continu est donné par Caron et Ravalet (2000).

L'estimateur K s'interprète facilement : c'est une moyenne de l'estimateur naturel $\bar{y}_t$ et d'un estimateur un peu plus complexe basé sur un estimateur à la date précédente $\theta_{t-1|t-1}$ auquel on ajoute un estimateur de l'évolution : $\bar{y}_t' - \bar{x}_{t-1}'$.

L'estimateur AK s'interprète de la même façon. La seule différence réside dans la surpondération de l'estimateur sur la vague entrante Y_{1t} par rapport aux autres estimateurs Y_{lt}.

Ces estimateurs sont simples à mettre en oeuvre et ont une bonne efficacité par rapport à l'estimateur optimal : Gurney et Daly ont montré que l'estimateur AK avait une variance proche de la variance de l'estimateur optimal. Kumar et Lee (1983) ont utilisé des données de l'enquête emploi canadienne pour calculer ces différents estimateurs et ont confirmé les bonnes propriétés de l'estimateur AK.

4. Approche « série temporelle »

Dans l'approche « série temporelle » (Blight et Scott, 1973; Scott et Smith, 1974; Scott, Smith et Jones, 1977; Jones, 1980, Binder et Hidiroglou, 1988; Binder et Dick, 1989), les paramètres d'intérêt θ_t sont supposés être des quantités aléatoires. La séquence $\{\theta_t\}$ est donc considérée comme une série temporelle que l'on peut modéliser.

4.1 Le modèle de Blight et Scott

Ce modèle est le même que celui de Patterson :

$$Y_{lt} = \theta_t + \xi_{lt}$$
$$Y_{lt} - \theta_t = \rho(Y_{(l-1)(t-1)} - \theta_{t-1}) + \eta_{lt}$$

auquel on ajoute une équation qui décrit l'évolution du paramètre θ_t :

$$\theta_t = \alpha\theta_{t-1} + \varepsilon_t$$

Les ε_t sont supposés non corrélés et indépendants des η_{lt}, avec $E(\varepsilon_t) = 0,\ V(\varepsilon_t) = \sigma^2$.

Les auteurs montrent que l'estimateur optimal est défini récursivement de la manière suivante :

$$\theta_{t|t} = \Delta_t^{-1}\left(\frac{\rho^2}{w_t^{'}} + \frac{1}{V(\theta_{t-1|t-1})} + \frac{\alpha^2}{\sigma^2}\right)\frac{\bar{y}_t^{''}}{w_t^{''}}$$
$$+ \Delta_t^{-1}\left(\frac{\rho}{w_t^{'}} + \frac{\alpha}{\sigma^2}\right)\frac{\theta_{t-1|t-1}}{V(\theta_{t-1|t-1})}$$
$$+ \Delta_t^{-1}\left(\frac{1}{V(\theta_{t-1|t-1})} + \frac{\alpha(\alpha-\rho)}{\sigma^2}\right)\left(\frac{\bar{y}_t^{'} - \rho\bar{x}_{t-1}^{'}}{w_t^{'}}\right)$$

$$V(\theta_{t|t}) = \Delta_t^{-1}\left(\frac{\rho^2}{w_t^{'}} + \frac{1}{V(\theta_{t-1|t-1})} + \frac{\alpha^2}{\sigma^2}\right)$$

avec

$$\Delta_t = \left(\frac{\rho^2}{w_t^{'}} + \frac{1}{V(\theta_{t-1|t-1})} + \frac{\alpha^2}{\sigma^2}\right)\left(\frac{1}{w_t^{'}} + \frac{1}{w_t^{''}} + \frac{1}{\sigma^2}\right) - \left(\frac{\rho}{w_t^{'}} + \frac{\alpha}{\sigma^2}\right)^2$$

$$w_t^{'} = \frac{(1-\rho^2)S^2}{\lambda n}$$

$$w_t^{''} = \frac{S^2}{\mu n}$$

$$\theta_{1|1} = \bar{y}_1^{''}$$

$$V(\theta_{1|1}) = \frac{S^2}{n}$$

- *Le cas particulier* $\sigma^2 = \infty$

Ce cas limite revient à dire que le modèle n'apporte aucune information sur θ_t et on retrouve les résultats de Patterson i.e les résultats de l'approche classique.

- *Le cas particulier* $\rho = 0$ *ou* $\lambda = 0$

Ce cas limite correspond en fait au modèle de Scott et Smith (1974) dans lequel l'échantillon est renouvelé entièrement à chaque période. L'estimateur optimal et la variance s'écrive alors très simplement :

$$\theta_{t|t} = (1-\pi)\bar{y}_t + \pi\alpha\theta_{t-1|t-1}$$

$$V(\theta_{t|t}) = (1-\pi)S^2$$

$$\text{avec } \pi = \frac{1}{1 + \dfrac{V(\theta_t | Y_{t-1})}{S^2}}$$

L'estimateur est la somme pondérée de l'estimateur naturel $\bar{y}_t$ et de l'estimateur construit grâce au modèle à partir de l'estimateur optimal de la période précédente $\alpha\theta_{t-1|t-1}$.

La réduction de variance est d'autant plus grande que π est proche de 1. C'est le cas lorsque $\dfrac{V(\theta_t | Y_{t-1})}{S^2}$ est proche de zéro i.e lorsque la variance induite par le modèle sur θ_t est petite par rapport à la variance due à l'échantillonnage. La prise en compte d'un modèle expliquant bien l'évolution de θ_t permet donc des gains d'efficacité importants.

Ce résultat reste vrai dans le cas général (ρ et λ quelconque) et permet de comprendre en quoi l'approche « série temporelle » permet des gains d'efficacité par rapport à l'approche « classique ».

4.2 Les modèles état-mesure et l'utilisation du filtre de Kalman

Le calcul de l'estimateur optimal dans le modèle de Blight et Scott peut sembler compliqué et peu exploitable. De plus il n'est pas forcément très facilement généralisable en l'état. Pour remédier à ces problèmes, « l'astuce » consiste à voir cet estimateur comme une application du calcul d'un estimateur optimal par le filtre de Kalman (Kalman, 1960) dans un modèle état-mesure (voir annexe).

En effet, le modèle de Blight et Scott peut être modélisé par les équations suivantes :

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

$$\begin{bmatrix} \bar{y}_t'' \\ \bar{y}_t' - \rho \bar{x}_t' \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & -\rho \end{bmatrix} \begin{bmatrix} \theta_t \\ \theta_{t-1} \end{bmatrix} + \begin{bmatrix} \xi_{1t} \\ \dfrac{1}{L-1} \displaystyle\sum_{l=2}^{L} \eta_{lt} \end{bmatrix}$$

C'est donc un modèle état-mesure et les équations obtenues par le filtre de Kalman correspondent bien à celles données dans la partie précédente.

Cette modélisation a deux avantages : d'une part les matrices à inverser sont de taille raisonnables : le calcul de l'estimateur optimal n'est plus un problème; d'autre part, cette méthode d'estimation permet d'utiliser des modèles plus riches et donc plus complexes.

On peut en particulier modéliser la série des θ_t et la série des erreurs e_t par des modèles ARMA (Binder et Hidiroglou, 1988; Binder et Dick, 1989). En effet, un modèle ARMA(p,q) s'écrit de manière générale :

$$\theta_t - \alpha_1 \theta_{t-1} - \cdots - \alpha_p \theta_{t-p} = \varepsilon_t - \beta_1 \varepsilon_{t-1} - \cdots - \beta_q \varepsilon_{t-q}$$

mais on peut le réécrire comme un modèle état-mesure sous la forme générale donnée par Harvey et Phillips (1979) :

$$\begin{cases} Z_{t+1} = AZ_t + B\varepsilon_{t+1} \\ \theta_t = CZ_t \end{cases}$$

avec

$$A = \begin{bmatrix} -\alpha_1 & 1 & 0 & \cdots & 0 \\ -\alpha_2 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ -\alpha_{r-1} & 0 & 0 & \cdots & 1 \\ -\alpha_r & 0 & 0 & \cdots & 0 \end{bmatrix}, \quad \mathrm{B} = \begin{bmatrix} 1 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{r-1} \end{bmatrix}$$

$$C = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix}, \; r = Max(p, q+1), \; \alpha_i = 0 \text{ si } i > p, \; \beta_i = 0 \text{ si } i > q,$$

la j^{ème} composante de Z_{t+1} s'écrit :

$$Z_{t+1}^j = -\alpha_j \theta_t - \cdots - \alpha_r \theta_{t-r+j} + \beta_{j-1}\varepsilon_{t+1} + \cdots + \beta_{r-1}\varepsilon_{t-r+j+1} \quad \text{(par convention } \beta_0 = 1)$$

En combinant deux représentations de ce type, on peut écrire le modèle général suivant :

$$\begin{aligned} Y_{lt} &= \theta_t + \xi_{lt} \\ \theta_t &\approx ARMA(p,q) \\ \xi_{lt} &\approx ARMA(m,n) \end{aligned}$$

sous forme d'un modèle état-mesure et par conséquent calculer l'estimateur optimal par le filtre de Kalman. Notons que le modèle de Blight et Scott est un cas particulier de ce modèle avec $\theta_t \approx ARMA(1,0)$ et $\xi_{lt} \approx ARMA(1,0)$.

5. Le problème de la désaisonnalisation

La modélisation sous forme de modèle état-mesure permet également d'enrichir les modèles en incluant facilement les problèmes liés à la saisonnalité des séries. En général, la désaisonnalisation est traitée soit par l'application de filtres *ad hoc* (c'est le cas de la méthode X11 Arima par exemple), soit par une approche modélisée (c'est le cas de la méthode Tramo-Seats par exemple). Nous nous restreignons dans ce papier à l'étude de l'approche modélisée.

On peut distinguer deux grandes approches modélisées de la saisonnalité : les modèles dits « structurels » (Engle, 1978; Harvey et Todd, 1983; Harvey, 1989) et

les modèles à composantes ARIMA (Box, Hillmer et Tiao, 1978; Burman, 1980; Hillmer et Tiao, 1982; Bell et Hillmer, 1984).

5.1 Les modèles « structurels »

Le modèle le plus simple est le suivant :

$$\theta_t = T_t + S_t + I_t$$
$$T_t = T_{t-1} + R_{t-1} + \nu_t^T$$
$$R_t = R_{t-1} + \nu_t^R$$
$$\sum_{j=0}^{S-1} S_{t-j} = \nu_t^S$$

où (I_t), (ν_t^T), (ν_t^R) et (ν_t^S) sont des bruits blancs indépendants de variance respective σ_I^2, $\sigma_{\nu^T}^2$, $\sigma_{\nu^R}^2$ et $\sigma_{\nu^S}^2$.

La première équation postule que la variable se décompose en un trend, une composante saisonnière et un irrégulier. La deuxième équation décrit le trend comme un trend linéaire local, tandis que la dernière équation modélise la variation saisonnière (une saisonnalité constante est obtenue dans le cas $\sigma_{\nu^S}^2 = 0$).

Ces modèles admettent une représentation état-mesure et peuvent donc être facilement intégrés dans les modélisations précédentes, lorsque l'on tient compte de l'erreur d'échantillonnage puisque seuls les Y_{It} et non les θ_t sont observés. Le filtre de Kalman permet alors de calculer l'estimateur optimal à chaque date.

Par exemple, pour S=4, une représentation espace d'états peut être la suivante :

$$Z_t = (T_t, R_t, S_t, S_{t-2}, S_{t-2})$$

$$\begin{cases} Z_{t+1} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & -1 & -1 & -1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix} Z_t + \begin{bmatrix} \nu_{t+1}^T \\ \nu_{t+1}^R \\ \nu_{t+1}^S \\ 0 \\ 0 \end{bmatrix} \\ \theta_t = (1 \quad 0 \quad 1 \quad 0 \quad 0) Z_t + I_t \end{cases}$$

Pfefferman (1991) a ainsi réalisé des estimations sur le nombre d'heures hebdomadaires et annuelles travaillées à partir de l'enquête emploi israélienne. Cette enquête est une enquête trimestrielle. Chaque trimestre, l'échantillon est composé de quatre panels de 3000 logements : trois panels ont déjà été interrogés par le passé, un panel est nouveau. Un panel est interrogé deux fois, puis laissé pendant deux trimestres et à nouveau interrogé deux fois.

Sa modélisation lui permet également d'inclure la prise en compte du biais de rotation, phénomène connu dans les enquêtes répétées (Bailar, 1975). Il montre que le modèle est bien approprié aux données. De plus, les résultats obtenus montrent une diminution de la variance d'un tiers par rapport à l'estimateur de Patterson. Enfin, son modèle lui permet d'obtenir directement la décomposition des variables en tendance, effet saisonnier et irrégulier.

Tiller (1992) utilise également ce type de modélisation pour estimer le taux de chômage à un niveau local (Etat du Massachusetts). Les données sont issues de l'enquête emploi américaine (US Current Population Survey). C'est une enquête mensuelle, chaque échantillon mensuel étant composé de huit panels. Chaque panel est alors interrogé quatre fois, laissé pendant huit mois et réinterrogé quatre fois (4in-8out-4in).

Tiller justifie son approche par la possibilité d'obtenir des estimations précises à des niveaux locaux en utilisant une approche modélisée. Les tests qu'il pratique conduise à accepter le modèle lorsqu'il prend bien en compte l'erreur d'échantillonnage i.e lorsqu'il prend en compte l'équation de mesure $Y_{lt} = \theta_t + \xi_{lt}$. Il obtient une réduction de 50% de la variance des estimateurs par rapport à l'approche classique.

5.2 Les modèles à composantes ARIMA

Dans cette modélisation, on suppose que chaque composante suit un modèle ARIMA :

$$\theta_t = T_t + S_t + I_t$$
$$(1-L)^d \varphi_T(L) T_t = \Theta_T(L) \nu_t^T$$
$$(1 + L + \cdots + L^{S-1}) S_t = \Theta_S(L) \nu_t^S$$
$$\varphi_I(L) I_t = \Theta_I(L) \nu_t^I$$

où (ν_t^T), (ν_t^S) et (ν_t^I) sont des bruits blancs indépendants de variance respective $\sigma^2_{\nu^T}$, $\sigma^2_{\nu^S}$ et $\sigma^2_{\nu^I}$.

La première équation postule que la variable se décompose en un trend, une composante saisonnière et un irrégulier. La deuxième (resp. troisième et quatrième) équation modélise le trend stochastique (resp. la composante saisonnière et l'irrégulier) à l'aide de modèles ARIMA.

Nous avons vu dans la partie précédente qu'un modèle ARMA (et il en est de même pour un modèle ARIMA) admettait une représentation état-mesure. Ainsi chaque composante peut être modélisée par une représentation du type :

$$\begin{cases} Z_{t+1}^T = A_T Z_t^T + B_T \varepsilon_{t+1}^T \\ T_t = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix} Z_t^T \end{cases}$$

$$\begin{cases} Z_{t+1}^S = A_S Z_t^S + B_S \varepsilon_{t+1}^S \\ S_t = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix} Z_t^S \end{cases}$$

$$\begin{cases} Z_{t+1}^I = A_I Z_t^I + B_I \varepsilon_{t+1}^I \\ I_t = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix} Z_t^I \end{cases}$$

Ceci permet d'en déduire une représentation état-mesure pour θ_t :

$$\begin{cases} Z_{t+1} = \begin{bmatrix} Z_{t+1}^T \\ Z_{t+1}^S \\ Z_{t+1}^I \end{bmatrix} = \begin{bmatrix} A_T & 0 & 0 \\ 0 & A_S & 0 \\ 0 & 0 & A_I \end{bmatrix} Z_t + \begin{bmatrix} B_T & 0 & 0 \\ 0 & B_S & 0 \\ 0 & 0 & B_I \end{bmatrix} \begin{bmatrix} \varepsilon_{t+1}^T \\ \varepsilon_{t+1}^S \\ \varepsilon_{t+1}^I \end{bmatrix} \\ \\ \theta_t = \begin{bmatrix} 1 & 0 & \cdots & 0 & 1 & 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \end{bmatrix} Z_t \end{cases}$$

On peut alors obtenir une représentation plus générale en prenant en compte les erreurs d'échantillonnage dans l'équation de mesure $Y_{lt} = \theta_t + \xi_{lt}$. Le filtre de Kalman permet encore une fois de calculer l'estimateur optimal.

C'est cette modélisation qu'ont retenu Binder et Dick (1990) pour étudier le taux de chômage grâce à l'enquête emploi canadienne (Canadian Labour Force Survey). C'est une enquête mensuelle composée de six panels. Chaque panel est interrogé six fois consécutivement.

Cette étude, réalisée à un niveau local (province de la Nouvelle Ecosse) leur permet d'obtenir un modèle qui reproduit correctement les séries. Les auteurs montrent l'importance de la prise en compte de l'erreur d'échantillonnage dans l'estimation. Malheureusement, ils ne comparent pas leurs résultats avec des estimations de l'approche « classique ».

6. Conclusion

L'intégration de la dynamique agrégée, la représentation sous forme de modèle état-mesure et son estimation par le filtre de Kalman permet d'obtenir des estimateurs plus précis que dans l'approche classique. Ce gain de précision nécessite toutefois l'utilisation d'un modèle pour décrire la grandeur à mesurer et il faut donc être prudent quant à la spécification de ce modèle.

L'avantage de l'approche série temporelle réside de manière au moins aussi importante dans la possibilité d'intégrer dans le modèle de nombreux éléments tels que le biais de renouvellement, la saisonnalité, des variables explicatives... Cette approche permet également d'obtenir des estimations à un niveau local, comme l'ont montré l'étude de Tiller et celle de Binder et Dick. Il est également possible d'introduire dans le modèle des contraintes sur des grandeurs liées par une équation comptable (population active, emploi et chômage par exemple) pour que les estimations soient cohérentes.

Notons enfin que Harvey et Chung (2000) utilise cette approche pour obtenir des estimateurs mensuels du taux de chômage à partir de données trimestrielles, issues de l'enquête emploi anglaise. De plus, en étudiant un modèle bivarié, ils réussissent à incorporer des données administratives pour améliorer la précision des estimateurs issus du sondage. Ces résultats sont d'une grande utilité pour le cas français puisque la taille de l'enquête n'est adaptée en principe qu'à une estimation trimestrielle, alors que de nombreux utilisateurs sont intéressés par des estimations mensuelles. De plus, l'utilisation de données administratives provenant de l'ANPE semble indispensable pour remédier au faible nombre de points qui seront disponibles dans l'enquête en continu en 2003.

Cette approche théorique doit bien sûr être validée par une approche empirique. Il faut dans un premier temps estimer un modèle de ce type sur les données administratives pour lesquelles il n'y a pas le problème de l'erreur d'échantillonnage. Ensuite il serait utile de construire un modèle bivarié pour décrire les données administratives et les données de l'enquête annuelle pour laquelle il existe un grand nombre d'observations. Il faut enfin construire un modèle en incorporant les données issues de l'enquête en continu et tenant compte de l'erreur d'échantillonnage.

Bibliographie

Bailar, B.A. (1975) The effects of rotation group bias on estimates from panel surveys. *Journal of the American Statistical Association* **70**, 23-29.

Bell, W.R. and Hillmer, S.C. (1984) Issues involved with the seasonal adjustment of economic time series. *Journal of Business and Economic Statistics* **2**, 291-320.

Binder, D.A. and Dick, J.P. (1989) Modelling and estimation for repeated surveys. *Survey Methodology* **15**, 29-45.

Binder, D.A. and Dick, J.P. (1990) Analysis of seasonal ARIMA models. *Survey Methodology* **16**, 239-253.

Binder, D.A. and Hidiroglu, M.A. (1988) Sampling in time. In *Handbook of Statistics* (eds P.R. Krishnaiah and C.R. Rao), vol 6, pp. 187-211. Amsterdam : Elsevier Science.

Blight, B.J.N. and Scott, A.J. (1973) A stochastic model for repeated surveys. *Journal of the Royal Statistical Society, Series B* **35**, 61-68.

Box, G.E.P., Hillmer, S.C. and Tiao, G.C. (1978) Analysis and modelling of seasonal time Series. In Zellner, A. (ed), *Seasonal Analysis of Economic Time Series*. Washington, DC : US Department of Commerce, Bureau of the Census, 309-334.

Burman, J.P. (1980) Seasonal adjustment by signal extraction. *Journal of the Royal Statistical Society, Series A* **143**, 321-337.

Caron, N. et Ravalet, P. (2000) Estimation dans les enquêtes répétées : application à l'enquête emploi en continu. *Série des documents de travail Méthodologie Statistique* **5**.

Duncan, G.J. and Kalton, G. (1987) Issues of design and analysis of surveys across time. *International Statistical Review* **5**, 97-117.

Engle, R.F. (1978) Estimating structural models of seasonality. In Zellner, A. (ed), *Seasonal Analysis of Economic Time Series*. Washington, DC : US Department of Commerce, Bureau of the Census, 281-297.

Gourieroux, C. et Monfort, A. (1990) *Séries temporelles et modèles dynamiques*. Economica, Paris.

Gurney, M and Daly, J.F. (1965) A multivariate approach to estimation in periodic sample surveys. *Proceedings of the American Statistical Association, Section on Survey on Research Methods*, 247-257.

Harvey, A.C. (1989) *Forcasting, Structural Time Series Models and the Kalman Filter*. Cambridge : Cambridge University Press.

Harvey, A.C. and Chung, C-H. (2000) Estimating the underlying change in unemployment in the UK. *Journal of the Royal Statistical Society, Series A* **163**, 303-339.

Harvey, A.C. and Todd, P.H.J. (1983) Forecasting economic time series with structural and Box-Jenkins Models : a case study. *Journal of Business and Economic Statistics* **1**, 299-306.

Hillmer, S.C. and Tiao, G.C. (1982) An ARIMA-model-based approach to seasonal adjustment. *Journal of the American Statistical Association* 77, 63-70.

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

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

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

Kumar, S. and Lee, H. (1983) Evaluation of composite estimation for the Canadian labour force survey. *Survey Methodology* **9**, 1-24.

Patterson, H.D. (1950) Sampling on successive occasions with partial replacement of units. *Journal of the Royal Statistical Society, Series 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* **9**, 163-175.

Scott, A.J. and 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. and Jones, R.G. (1977) The application of time series methods to the analysis of repeated surveys. *International Statistical Review* **45**, 13-28

Tiller, R.B. (1992) Time series modeling of sample survey data from the U.S. Current Population Survey. *Journal of Official Statistics* **8**, 149-166.

Annexe : le filtre de Kalman

Une présentation plus complète peut être trouvée dans Gourieroux et Monfort (1990).

Le modèle état-mesure

Examinons tout d'abord ce qu'est un modèle état-mesure. Une observation est modélisée comme une combinaison entre un signal et un bruit :

$$Y_t = C_t Z_t + \eta_t$$

où Y_t est l'observation à la date t de taille n, Z_t les facteurs non observées (i.e l'état du système) qui expliquent Y_t , C_t une matrice déterministe de taille $n \times K$ qui contient les poids des différents facteur dans le modèle et η_t un bruit blanc normal de variance R.

Bien que l'état du système ne soit pas observé, on suppose que l'on connaît son évolution :

$$Z_{t+1} = A_t Z_t + \varepsilon_t$$ [3]

où ε_t est un bruit blanc normal de variance Q, A_t est une matrice déterministe de taille $K \times K$ qui décrit la transition du système entre les dates t et t+1. On suppose par ailleurs que Z_0 est un vecteur aléatoire de loi $N(m,P)$, indépendant des $\begin{pmatrix} \varepsilon_t \\ \eta_t \end{pmatrix}$ [4].

[3] Les résultats s'étendent facilement au cas d'un modèle plus général de la forme $Z_{t+1} = A_t Z_t + B_t \varepsilon_t$

[4] Il est possible de s'affranchir de l'hypothèse de normalité du bruit $\begin{pmatrix} \varepsilon_t \\ \eta_t \end{pmatrix}$ et de l'état initial Z_0. Il est également possible d'introduire des corrélations entre ε_t et η_t et de supposer que les matrices de variance covariance dépendent du temps.

Le filtre de Kalman

Le filtre de Kalman permet de calculer de manière récursive les estimations optimales des Z_t conditionnellement aux observations. Ce filtre permet donc de connaître la valeur réelle de l'état du système.

On notera

$$Z_{t|t'} = E(Z_t | Y_0, ..., Y_{t'}) \text{ et } \Sigma_{t|t'} = V(Z_t - Z_{t|t'})$$

$$Y_{t|t'} = E(Y_t | Y_0, ..., Y_{t'}) \text{ et } M_{t|t'} = V(Y_t - Y_{t|t'})$$

Supposons tout d'abord qu'on veuille prédire la valeur de Z_t et Y_t à la date t-1. Puisque les erreurs ε_{t-1} et η_t ne sont pas connues, les meilleures prévisions sont :

$$Z_{t|t-1} = A_{t-1} Z_{t-1|t-1}$$
$$Y_{t|t-1} = C_t Z_{t|t-1}$$

On en déduit alors :

$$\Sigma_{t|t-1} = A_{t-1} \Sigma_{t-1|t-1} A_{t-1}^{'} + Q$$
$$M_{t|t-1} = C_t \Sigma_{t|t-1} C_t^{'} + R$$

Lorsqu'une nouvelle observation devient exploitable, les estimations peuvent être modifiées pour prendre en compte cette information. On introduit tout d'abord l'erreur de prédiction

$$e_t = Y_t - Y_{t|t-1}$$

L'estimation optimale de Z_t à la date t est alors une combinaison entre l'estimation optimale à la date t-1 et l'erreur de prédiction : on réajuste l'estimation de Z_t en fonction de l'erreur qu'on avait faite pour Y_t :

$$Z_{t|t} = Z_{t|t-1} + K_t e_t$$
$$\text{avec } K_t = \Sigma_{t|t-1} C_t^{'} \left[C_t \Sigma_{t|t-1} C_t^{'} + R \right]^{-1} = \Sigma_{t|t-1} C_t^{'} V(e_t)^{-1}$$

La matrice K_t est appelée gain du filtre à la date t.

On en déduit alors :

$$\Sigma_{t|t} = [Id - K_t C_t]\Sigma_{t|t-1}$$

Cette méthode permet donc bien de calculer récursivement les estimations optimales de Z_t conditionnellement aux observations. Elle nécessite l'inversion de matrices de petite taille ($K \times K$) et est donc facilement utilisable.

Initialisation du filtre

Lorsqu'aucune information n'est disponible, il paraît raisonnable d'utiliser les valeurs $Z_{0|-1} = EZ_0 = m$ et $\Sigma_{0|-1} = V(Z_0) = P$. On peut en fait montrer que ce choix est optimal puisque le filtre donne à la date t=0 l'estimation optimale (calculable directement).

Prévision

Les meilleures prévisions de l'état du système à des dates ultérieures sont données très simplement par les formules suivantes :

$$Z_{t+h|t} = A_{t+h-1} Z_{t+h-1|t}$$
$$\Sigma_{t+h|t} = A_{t+h-1} \Sigma_{t+h-1|t} A'_{t+h-1} + Q$$

Lissage

Dans le filtre de Kalman tel que nous l'avons présenté, seule l'estimation de Z_T à la date T utilise toute l'information disponible à la date T. Il est pourtant possible de réestimer l'état du système à des dates antérieures ($t < T$) en raisonnant cette fois-ci par récurrence arrière.

Supposons par exemple qu'on connaisse $Z_{t+1|T}$ et qu'on cherche à calculer $Z_{t|T}$. On ajuste l'estimation de Z_t à la date t ($Z_{t|t}$) en utilisant la différence de prévision entre les dates t et T ($Z_{t+1|T} - Z_{t+1|t}$) pour Z_{t+1} :

$$Z_{t|T} = Z_{t|t} + F_t \left[Z_{t+1|T} - Z_{t+1|t} \right]$$
$$\text{avec } F_t = \Sigma_{t|t} A_t' \Sigma_{t+1|t}^{-1}$$
$$\Sigma_{t|T} = \Sigma_{t|t} + F_t \left[Z_{t+1|T} - Z_{t+1|t} \right] F_t'$$

Estimation

En général, les matrices précédentes (A_t, C_t, R, Q) ne sont pas connues et il faut les estimer à partir des observations. Cette estimation est réalisée par la méthode du maximum de vraisemblance. Si on note θ le vecteur de paramètres du modèle, la densité des observations s'écrit :

$$l(y;\theta) = f(y_0;\theta) f(y_1/y_0;\theta) ... f(y_T/y_0,...,y_{T-1};\theta)$$

Le terme général $f(y_t/y_0,...,y_{t-1};\theta)$ est d'après les hypothèses précédentes la densité de la loi normale de moyenne $y_{t|t-1}(\theta)$, de matrice de variance covariance $M_{t|t-1}(\theta)$.

Pour toute valeur de θ, le filtre de Kalman nous permet de calculer $y_{t|t-1}(\theta)$ et $M_{t|t-1}(\theta)$. On en déduit la log-vraisemblance du modèle donnée par :

$$L_T(\theta) = -\frac{(T+1)n}{2} Log(2\pi) - \frac{1}{2} \sum_{t=0}^{T} Log \det(M_{t|t-1}(\theta))$$
$$-\frac{1}{2} \sum_{t=0}^{T} (y_t - y_{t|t-1}(\theta))' \left[M_{t|t-1}(\theta) \right]^{-1} (y_t - y_{t|t-1}(\theta))$$

La maximisation de cette log-vraisemblance nous donne une estimation de θ.