COMPARAISON D'ESTIMATEURS ALTERNATIFS DANS L'ENQUÊTE EMPLOI EN CONTINU

Dominique PLACE (*)

(*) Insee, Unité Méthodes Statistiques

Introduction

Dans une enquête répétée avec renouvellement partiel de l'échantillon, des corrélations structurelles existent entre les estimateurs transversaux, puisqu'une partie des individus est réinterrogée d'une période sur l'autre. Une piste naturelle d'amélioration des estimateurs réside dans la prise en compte de ces corrélations en mobilisant les données des périodes passées. Cela conduit à un ensemble d'estimateurs que l'on regroupe sous le terme d'estimateurs composites.

Certains de ces estimateurs ont pu être évalués dans le cadre de l'Enquête Emploi en Continu française (EEC). Cette enquête trimestrielle repose sur un échantillon de secteurs géographiques répartis en six groupes de rotation, dans lesquels les logements sont finalement sélectionnés. Chaque trimestre, dans cinq de ces groupes, les logements précédemment sélectionnés sont réinterrogés tandis que dans le dernier groupe de rotation, tous les logements sont remplacés. Le recouvrement entre deux échantillons trimestriels consécutifs est donc de cinq sixièmes. Toutes les personnes de plus de 15 ans sont ensuite enquêtées dans les logements sélectionnés, *a priori* six trimestres de suite.

Ce schéma de rotation représente un compromis entre un panel, plus adapté pour la mesure d'évolutions, et un échantillon avec renouvellement complet chaque trimestre, qui permet d'obtenir de meilleures estimations des moyennes annuelles. Les techniques d'estimation composite prolongent ce compromis en mettant à profit le caractère répétitif de l'enquête pour améliorer à la fois les estimateurs en niveau, trimestriels ou moyens, et ceux des évolutions, qu'ils soient entre deux trimestres consécutifs ou non, ou entre deux moyennes.

Bien que le principe en ait été défini depuis longtemps, ces estimateurs ont été relativement peu utilisés jusqu'à récemment. Aux États-Unis, la *Current Population Survey* utilise pour les principales grandeurs sur l'emploi et le chômage des estimateurs récursifs, notamment l'estimateur composite AK. Définis par Gurney et Daly (1965), ils combinent l'estimation transversale courante et une composante qui corrige l'estimation de la période précédente par l'évolution sur la partie stable de l'échantillon. Une extension, appelée pondération composite AK, proposée par Fuller (1990), est mise en œuvre depuis une dizaine d'années. Ces estimateurs peuvent être considérés comme des approximations des estimateurs linéaires optimaux à fenêtre fixe, définis dès 1942 par Jessen dans le cas de deux périodes. L'extension à des fenêtres de temps plus grandes est apparue pendant longtemps difficile à implémenter mais avec l'accroissement des moyens de calcul, ils ont pu être étudiés avec les données de l'enquête emploi australienne par Bell (1998, 2001). Ces estimateurs sont une combinaison linéaire d'estimateurs calculés sur les différents groupes de rotation pour un nombre fixe de périodes. Une autre démarche a été adoptée dans l'Enquête sur la Population Active (EPA) au Canada : les estimateurs par régression modifiée (Singh et Merkouris, 1995, Fuller et Rao, 2001). Leur principe est d'ajouter aux variables de calage usuelles portant sur la période courante des variables relatives à la précédente, avec imputation pour les individus entrants.

Dans la première partie, on présente l'Enquête Emploi en Continu, en insistant sur la méthode actuelle d'estimation de la variance. Les définitions des trois types d'estimateurs sont ensuite précisément données. Dans la troisième partie, dans une démarche semblable à celle de Bell (2001), des évaluations sont présentées avec les données de l'EEC et en discutant de l'influence du biais de rotation. Comme la méthode d'estimation de la variance actuellement utilisée ne permet pas aisément de mesurer la variance des estimateurs par régression modifiée, seuls les estimateurs AK et linéaires optimaux à fenêtre fixe seront comparés à l'estimateur actuel.

1. L'Enquête Emploi en Continu

C'est une enquête conduite chaque trimestre par l'Insee sur un échantillon d'environ 54 000 logements ordinaires de France métropolitaine. Ses objectifs majeurs sont de fournir des informations quantitatives et structurelles sur le marché du travail français, notamment en évaluant chaque trimestre la population active, le nombre de chômeurs et le taux de chômage au sens du Bureau International du Travail ainsi que les évolutions trimestrielles du chômage. Le règlement européen n° 577/98 définit deux critères de précision que doit satisfaire l'enquête sur les estimations de chômage :

- pour un groupe de chômeurs représentant 5% de la population d'âge actif, le coefficient de variation pour l'estimation des moyennes annuelles ne doit pas excéder 8% au niveau des régions de plus de 300.000 habitants ;
- l'écart type des estimations de variations au niveau national entre deux trimestres consécutifs ne doit pas dépasser 2% du nombre de chômeurs.

Ces deux critères sont opposés : pour satisfaire le premier, il y aurait intérêt à renouveler complètement l'échantillon entre deux trimestres pour interroger le plus de personnes différentes en une année, tandis que pour le second, c'est le panel qui est le plus adapté. Le schéma de rotation choisi constitue un compromis entre ces deux contraintes, tout en sachant que la première est délicate à respecter pour les régions de taille réduite. Il y a aussi une contrainte de collecte particulièrement forte : les personnes sélectionnées doivent être interrogées sur leur situation sur le marché du travail lors d'une semaine de référence pendant les deux semaines qui suivent celles-ci. Ces différentes contraintes ont conduit l'Insee à opter pour un échantillon rotatif et aréolaire dont le plan a été décrit en détail par Marc Christine (2002). Il ne s'agit ici que d'en rappeler les principes, qui ont été globalement repris pour la construction du nouvel échantillon, décrite lors de ces journées de méthodologie statistique par Vincent Loonis.

1.1. Plan de sondage

Par rapport à des enquêtes emploi d'autres pays, une particularité de l'échantillon français construit à partir du recensement de 1999 est qu'il a été conçu une fois pour toutes pour une longue durée, de 2002 à 2011. Il n'y a pas de redécoupage et de sélection de nouvelles aires de manière régulière comme par exemple au Canada et aux Pays-Bas. Le plan de l'EEC est à plusieurs degrés. Les unités primaires (UP) sont des communes ou des districts. Leur tirage a été stratifié par région et tranche de taille d'unités urbaines. Dans chaque strate, elles ont été sélectionnées avec des probabilités proportionnelles au nombre de logements. Dans chaque UP sélectionnée, des secteurs contenant entre 120 et 240 logements contigus ont été découpés, sur la base des limites géographiques repérables comme les rues, les cours d'eau, etc. Dans l'urbain très dense, ils peuvent n'être constitués que de quelques immeubles voire d'un seul. Dans chaque UP, un seul a été sélectionné par un tirage à probabilités proportionnelles au nombre de logements. Chaque secteur sélectionné a été finalement découpé en aires d'environ 20 logements contigus au nombre de 6 à 13 par secteurs. Dans chaque strate, l'échantillon de secteurs a été réparti aléatoirement en six sous-échantillons de taille égale[1]. Les sous-échantillons ont été intégrés successivement dans le dispositif lors d'une phase d'initialisation en 2001 et 2002. Dans un secteur, une seule aire est enquêtée à un trimestre donné. Elle est enquêtée pendant six trimestres puis est remplacée par une autre aire du secteur. L'ordre d'interrogation des aires dans un secteur est déterminé aléatoirement. Le taux de recouvrement de l'échantillon entre deux trimestres consécutifs est donc de 5/6 (cf. figure 1). Plus généralement, entre deux trimestres t et t', avec $|t-t'|<6$, le taux de recouvrement des échantillons trimestriels est : $o_{tt'}=1-\frac{|t-t'|}{6}$.

[1] A cause de certaines difficultés liées au découpage et à l'organisation de la collecte, l'équirépartition par région n'a pas pu être complètement respectée. Toutefois l'équirépartition est respectée au niveau national et on considère qu'un sous-échantillon de secteurs résulte de l'échantillon total de secteurs par un tirage aléatoire simple au taux de 1/6.

Un groupe de rotation est l'ensemble des aires entrant au même trimestre. Il résulte donc d'un tirage simple d'une aire dans chaque secteur d'un sous-échantillon de secteurs.

Figure 1 : Schéma de rotation de l'enquête Emploi en continu

Trimestre	Groupes de rotation											
	1A	2A	3A	4A	5A	6A	1B	2B	3B	4B	5B	6B
1	1											
2	2	1										
3	3	2	1									
4	4	3	2	1								
5	5	4	3	2	1							
6	6	5	4	3	2	1						
7		6	5	4	3	2	1					
8			6	5	4	3	2	1				
9				6	5	4	3	2	1			
10					6	5	4	3	2	1		
11						6	5	4	3	2	1	
12							6	5	4	3	2	1
13								6	5	4	3	2
14									6	5	4	3
15										*6*	5	4

Note de lecture : le chiffre dans la case indique le rang d'interrogation pour le groupe de rotation.

L'échantillon complet comprend 2554 secteurs et un des six sous-échantillons 425 ou 426. Un échantillon trimestriel inclut environ 54 000 logements, un groupe de rotation environ 9 000, ce qui représente 72 000 répondants âgés de plus de 15 ans chaque trimestre, 12 000 par groupe de rotation. Les première et dernière interrogations d'un ménage se font en face en face tandis que les interrogations se font par téléphone. En cas de déménagement, le ménage n'est pas suivi et ce sont les éventuels nouveaux occupants du logement qui sont contactés.

1.2. Estimateurs

Le calcul des probabilités d'inclusion est immédiat : pour un logement, c'est celle de l'aire à laquelle il appartient. Soit une aire k dans le secteur i, celui-ci étant découpé en N_i aires, sa probabilité d'inclusion est :

$$\pi_k = \frac{M_{str}}{N_i} \frac{\log_i}{\log_{str}}$$

où $\log_i$ est le nombre de logements repérés au recensement de 1999 dans le secteur i, $\log_{str}$ celui de la strate, et M_{str} le nombre d'UP et donc de secteurs sélectionnés dans la strate. La probabilité de sélection de l'aire dans un groupe de sélection est $\pi_k / 6$. Un calage est effectué par groupe de rotation sur les données sociodémographiques trimestrielles à un niveau national et aussi en partie à un niveau régional. Des données sur les logements au niveau national fournies par le compte satellite du logement sont aussi utilisées : le nombre total actualisé de logements, celui de résidences principales et de logements construits depuis 1999[2].

[2] Les logements construits depuis le recensement de 1999 sont inclus dans l'échantillon au moment de la phase de ratissage des aires préalable à leur interrogation : une aire est explicitement définie comme une zone

En notant x_{tl} le vecteur des variables auxiliaires en t d'un logement l, alors le poids final au trimestre t, w_{tl}, est donné par : $w_{tl} = \frac{1}{\pi_l} \mathrm{F}(x'_{tl}\lambda_t^{Ri})$.

Le vecteur λ_t^{Ri}, spécifique au rang d'interrogation du logement, est calculé pour satisfaire les équations de calage par rang d'interrogation : $\sum_{l \in s_t^{Ri}} w_{tl} x_{tl} = X_t$ où X_t est le vecteur des totaux de référence. En pratique, les poids finaux sont obtenus avec le progiciel Calmar sous SAS, avec la fonction exponentielle, ce qui correspond à la méthode du raking ratio (voir par exemple Tillé, 2001, pour la description de la méthode générale du calage, de l'algorithme et des propriétés des estimateurs calés).

Le calage étant effectué par groupe de rotation, on obtient des estimateurs trimestriels $\hat{y}_t^{Ri}$ d'un total y_t pour chaque rang d'interrogation Ri, appelés par la suite estimateurs élémentaires :

$$\hat{y}_t^{Ri} = 6 \sum_{l \in s_t^{Ri}} w_{tl} y_{tl} .$$

L'estimateur trimestriel actuellement utilisé est l'estimateur naturel, c'est-à-dire la moyenne arithmétique simple des six estimateurs élémentaires :

$$\hat{y}_t = \frac{1}{6} \sum_{i=1}^{6} \hat{y}_t^{Ri} .$$

Les grandeurs longitudinales comme les évolutions et les moyennes annuelles sont simplement estimées par substitution. Par exemple, l'évolution entre deux trimestres t et t' est estimée par : $\hat{\Delta}_{tt'} = \hat{y}_{t'} - \hat{y}_t$.

1.3. Biais de rotation

Selon la théorie des sondages, si les méthodes de redressement sont adaptées, les estimateurs par rangs d'interrogation doivent avoir la même espérance. Cela signifie qu'en moyenne, il n'y a pas d'écarts entre les estimations élémentaires d'un même trimestre. Mais depuis longtemps, on a remarqué que dans les panels et les enquêtes à échantillon rotatif, les estimations basées sur les rangs d'interrogation présentent généralement en moyenne des écarts systématiques. Cela est désigné par le terme de biais de rotation, introduit par Bailar (1975). C'est le signe que, conditionnellement aux variables auxiliaires utilisées dans le calage, la non-réponse reste en partie dépendante de la variable d'intérêt. Ces biais sont sans doute dus en partie à des phénomènes d'attrition et de lassitude de la part des répondants. En plus de l'attrition usuelle, il y a de l' « attrition négative » dans l'EEC puisque, dans les aires en réinterrogation, on cherche à interroger des personnes qui n'ont pas répondu à l'enquête lors des interrogations précédentes. Une conséquence est que le taux de réponse en deuxième interrogation est légèrement plus élevé qu'en première. Malgré le calage séparé par groupe de rotation, les biais de rotation demeurent et pour l'instant, aucune procédure ne permet réellement de réduire ces biais spécifiques, qui se manifestent par des écarts systématiques entre les estimations par groupe de rotation. Ces biais existent dans la plupart des enquêtes avec échantillon rotatif. Il existe plusieurs manières de les mesurer (Goux, 2005) dont les indices de Bailar. Ils sont définis pour chaque période et chaque rang d'interrogation i :

$$B_{it} = 100 \frac{\hat{y}_t^{Ri}}{\hat{y}_t} .$$

Les indices moyens sur la période T1 2003 - T4 2008 attestent d'un biais de rotation plus important sur le chômage que sur l'emploi (cf. tableau 1). On constate aussi un effet du mode d'interrogation :

---

géographique. Toutefois, on a constaté un défaut de couverture de ces logements en région parisienne, où les aires ne peuvent n'être réduites qu'à quelques étages d'un immeuble.

lors de la sixième interrogation qui se fait en face à face, il y a sensiblement plus de personnes qui sont classées comme chômeurs selon les critères BIT que lors de la cinquième interrogation effectuée par téléphone, bien qu'au fil des réinterrogations par téléphone, la proportion de chômeurs baisse en moyenne.

Tableau 1 : Indices de Bailar moyens pour les nombres de chômeurs et d'actifs occupés

	Rang d'interrogation					
	1	2	3	4	5	6
Nombre de chômeurs	105,6	102,8	99,9	97,7	96,0	98,0
Nombre d'actifs occupés	99,9	99,8	100,0	100,1	100,2	100,1

Source : EEC T1 2003 au T4 2008.
Champ : population des ménages de France métropolitaine.

1.4. Estimation des variances et des covariances

1.4.1. Estimation de la variance transversale

L'estimation de la variance dans les enquêtes de l'Insee est effectuée à l'aide de formules analytiques de variance avec décomposition par degrés et par phases de sondage. Les statistiques complexes ainsi que le calage sont pris en compte par la technique de linéarisation. Les principes ont été exposés par Ardilly et Osier (2007) qui ont utilisé pour le calcul de la variance transversale de l'enquête Emploi le progiciel de calcul Poulpe développé à l'Insee. Une difficulté de ce calcul tient à la sélection d'échantillons de taille un à l'intérieur des UP[3], ce qui ne permet pas de calcul empirique direct de la variance à ce degré de sondage. Cela nécessite de fusionner les UP et de se ramener à un plan de sondage approché. Il en résulte une légère surestimation de la variance. Ces calculs sont effectués chaque trimestre mais on observe une bonne stabilité des coefficients de variation au cours du temps.

Tableau 2 : éléments sur la précision transversale

	Au T4 2008			CV moyen
	Total estimé (milliers)	Écart type (milliers)	CV (%)	2003-2008 (%)
Actifs occupés	25 862	85	0,3	0,3
Chômeurs	2 231	46	2,1	2,1
âgés de 15 à 24 ans	619	21	3,4	3,7
âgés de 25 à 49 ans	1 272	33	2,6	2,5
âgés de plus de 50 ans	340	16	4,7	4,7

Champ : population des ménages de France métropolitaine.

1.4.2. Estimation des variances longitudinales

L'estimation de la variance d'estimateurs longitudinaux revient à celle de la covariance d'estimateurs construits sur des échantillons différents. Par exemple, la variance d'une évolution entre deux trimestres, $\hat{\Delta}_{tt'}$, s'écrit ainsi :

[3] Pour un échantillon trimestriel, la sélection d'un secteur dans une UP donnée suivie de la sélection d'une aire dans ce secteur est équivalente à la sélection directe d'une aire dans l'UP. Le plan de sondage dans l'UP est en effet complètement déterminé par les probabilités simples de sélection, qui sont aussi les probabilités des échantillons possibles, chacun réduit à une seule aire. La procédure de sélection importe peu du moment qu'elle respecte les probabilités simples de sélection. Le secteur ne constitue à ce niveau qu'un moyen pour limiter le travail de découpage des aires. Pour les estimations longitudinales, les secteurs sont à prendre en compte puisqu'ils sont le support de la rotation des aires.

$$Var(\hat{\Delta}_{tt'}) = Var(\hat{y}_t) + Var(\hat{y}_{t'}) - 2Cov(\hat{y}_t, \hat{y}_{t'}).$$

Généralement la covariance est calculée en faisant l'hypothèse d'indépendance intertemporelle des estimateurs élémentaires portant sur la partie rotative de l'échantillon. Cela est acceptable lorsque le renouvellement de l'échantillon est tiré de manière quasiment indépendante de l'échantillon initial avec la seule contrainte de disjonction[4]. Dans l'EEC, cette hypothèse apparaît particulièrement forte puisqu'une aire est toujours remplacée par une autre aire appartenant au même secteur. Un secteur est une unité assez petite et une corrélation positive non nulle peut apparaître entre les parties rotatives de l'échantillon. En décomposant selon les degrés aire et secteur, on voit qu'il y a deux effets sur la covariance : un effet positif de grappe, qui tient au fait qu'on reste dans les mêmes secteurs, et un effet propre de sondage, qui est négatif. Cet effet provient du tirage disjonctif de deux échantillons, réduits chacun à une aire, dans les secteurs soumis à rotation.

Même s'il y a plusieurs échantillons différents en jeu, il est possible pour le calcul de variance et de covariance de se ramener par linéarisation à des estimateurs de Horvitz-Thompson (voir par exemple Goga *et al.*, 2006). Ensuite on peut établir la formule théorique suivante (Place, 2008) :

$$Var(\hat{\Delta}_{tt'}) = Var(\hat{\Delta}_{tt'}^{panel}) + 2(1 - o_{tt'}) \sum_{i \in U_I} \frac{N_i^2}{\pi_i} S_{i,tt'} \quad (1)$$

où U_I est l'échantillon de secteurs, N_i est le nombre d'aires dans le secteur i, $Var(\hat{\Delta}_{tt'}^{panel})$ est la variance de l'estimateur panélisé, c'est-à-dire celui que l'on aurait si les mêmes aires étaient enquêtées aux deux périodes, et $S_{i,tt'}$ est la covariance intra-secteur des totaux par aires entre les deux trimestres :

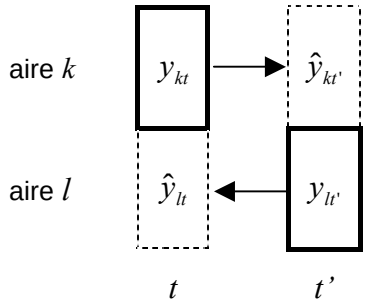
$$S_{i,tt'} = \frac{1}{N_i - 1} \sum_{k \in i} (y_{tk} - \bar{y}_{ti})(y_{t'k} - \bar{y}_{t'i}).$$

Le second terme de (1) apparaît comme une correction pour rotation, d'autant plus forte que le recouvrement des échantillons, mesuré par $o_{tt'}$, est faible. Comme les covariances intra-secteurs sont en général positives, on retrouve le fait que c'est le panel qui permet d'avoir la variance la plus faible pour l'évolution.

La formule (1) sert de base pour les estimations de variance. L'estimation du premier terme ne pose pas de difficulté particulière : il suffit de se ramener à l'échantillon cylindré formé par les aires enquêtées aux deux périodes (voir annexe 1). En revanche, l'estimation de la correction pour rotation est plus problématique puisque, dans les secteurs soumis à rotation, une aire différente est enquêtée chaque trimestre et aucune covariance intertemporelle n'est directement observable. Pour cela, il faudrait connaître les valeurs aux deux dates pour les deux aires du secteur. On peut contourner ce défaut d'information par l'utilisation de l'imputation (cf. figure 2). L'imputation pour le statut d'activité se fait au niveau individuel sur la base des transitions observées. En pratique, pour imputer les valeurs en t', un hot-deck stratifié est effectué avec des strates formées par croisement du statut d'activité en t, du sexe, des tranches d'âge et de la région. Dans cette méthode apparaissent plusieurs difficultés liées à différents phénomènes relatifs à l'occupation des logements ordinaires. L'imputation s'effectue au niveau de l'individu, mais pour remonter au niveau du ménage, il faut tenir compte des changements dans la composition des ménages. Parmi les causes de ces modifications, on peut citer les départs des jeunes de leurs domiciles familiaux qui se produisent plutôt entre le troisième et quatrième trimestre. Il y a également des variations saisonnières sur la vacance des logements : à cause de déménagements plus fréquents à cette période, c'est en été que les logements vacants sont les plus nombreux. Cela implique d'imputer des hors champs au niveau des individus et des ménages. Par ailleurs, il faut aussi tenir compte de la non-réponse et du calage en refaisant un calage avec les valeurs imputées, ce qui contribue encore à alourdir les algorithmes et à allonger les temps de calcul.

[4] La corrélation entre deux échantillons disjoints est légèrement négative. Mais lorsque le taux de sondage est très petit, cette corrélation est négligeable.

Figure 2 : Imputation dans un secteur pour l'estimation de la covariance intra-secteur



Légende : en t, l'aire k est enquêtée et elle est remplacée en t' par l'aire l dans le même secteur. En se basant sur les aires présentes aux deux trimestres, les totaux par aire manquants pour calculer la covariance intra-secteur sont imputés par $\hat{y}_{kt'}$ *et* $\hat{y}_{lt}$ *.*

2. Les différents estimateurs composites

2.1. L'estimateur linéaire optimal à fenêtre fixe

2.1.1. Définition

C'est une combinaison linéaire bien choisie d'estimateurs élémentaires calculés sur une fenêtre fixe de f périodes :

$$\hat{y}_{tf} = \sum_{u=0}^{f-1}\sum_{i=1}^{6} a_i^u \hat{y}_{t-u}^{Ri} = A'\hat{Y}_{tf}$$

où A est le vecteur des coefficients : $A = (a_1^0,\ldots,a_6^0,a_1^1,\ldots,a_1^{f-1},\ldots a_6^{f-1})'$ et $\hat{Y}_{tf}$ celui des estimateurs élémentaires sur la fenêtre : $\hat{Y}_{tf} = (\hat{y}_t^{R1},\ldots,\hat{y}_t^{R6},\hat{y}_{t-1}^{R1},\ldots,\hat{y}_{t-f+1}^{R1},\ldots,\hat{y}_{t-f+1}^{R6})'$.

Pour avoir des estimateurs sans biais du total y_t, les coefficients a_i^u doivent respecter des conditions reposant sur des hypothèses sur les biais de rotation. Si on suppose que ces biais sont nuls, comme cela est fait dans l'enquête emploi australienne, ces conditions sont :

$$\sum_{i=1}^{6} a_i^0 = 1\,;\ \sum_{i=1}^{6} a_i^u = 0 \text{ pour } u \geq 1\,.$$

Cela s'écrit matriciellement : $C'A = c$ avec $C = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & \cdots & 1 \end{pmatrix} \otimes (1,1,1,1,1,1)'$, matrice de taille $6f \times f$.

Comme cela a été montré ci-dessus, si l'absence de biais de rotation est une hypothèse réaliste pour le nombre d'actifs occupés, cela n'est pas le cas pour le nombre de chômeurs. On suppose alors :

$$\mathrm{E}(\hat{y}_t^{Ri}) = (1+b_i)\mathrm{E}(\hat{y}_t) = (1+b_i)y_t \qquad (2)$$

avec $\sum_{i=1}^{6} b_i = 0$.

Les b_i peuvent être calculés à partir des indices moyens de Bailar ou en utilisant un algorithme de dessaisonalisation comme X11, qui donne les mêmes résultats. Sous ces hypothèses, les conditions pour avoir un estimateur de biais nul deviennent :

$$\sum_{i=1}^{6}(1+b_i)a_i^0 = 1\,;\ \sum_{i=1}^{6}(1+b_i)a_i^u = 0 \text{ pour } u \geq 1\,.$$

Matriciellement, cela s'écrit $C'A = c$ avec $C = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & \cdots & 1 \end{pmatrix} \otimes (1+b_1,\ldots,1+b_6)'$.

La minimisation de la variance de $\hat{y}_{tf}$ requiert en théorie la connaissance de la matrice de variance-covariance du vecteur des estimateurs élémentaires $\hat{Y}_{tf}$, $V = Var(\hat{Y}_{tf})$, matrice carrée d'ordre 6*f*. On a alors : $Var(\hat{y}_{tf}) = A'VA$, que l'on doit minimiser sous les contraintes précédentes. Le programme d'optimisation d'une forme quadratique sous des contraintes linéaires est classique et il conduit au résultat suivant.

L'estimateur $\hat{y}_{tf} = A'\hat{Y}_{tf}$ *est de variance minimale pour* $A = V^{-1}C(C'V^{-1}C)^{-1}c$ *et sa variance est alors :* $V_f^{opt} = c'(C'V^{-1}C)^{-1}c$.

La matrice exacte de variance-covariance V n'est jamais connue et c'est une estimation de cette matrice qui est utilisée pour obtenir un estimateur qui n'est pas exactement l'estimateur optimal défini ci-dessous, mais qui en est très proche. *A priori*, pour une fenêtre fixe, à chaque trimestre, il faut recalculer les coefficients pour obtenir cet estimateur quasi optimal. Cependant, lorsque le plan de sondage est inchangé d'un trimestre à l'autre, on peut admettre que la matrice de variance-covariance reste constante et conserver les mêmes coefficients. Afin d'avoir une estimation plus robuste des variances et des covariances, il vaut mieux calculer une matrice moyenne sur une période assez importante. Les deux procédures de moyennisation et de calcul des coefficients peuvent être interverties : sur une période de référence assez longue, on calcule des coefficients optimaux pour chaque trimestre et ensuite la moyenne en est faite pour fixer les coefficients d'estimateurs quasi optimaux sur la période de référence et pour les trimestres à venir. L'écart entre les deux méthodes est très faible même s'il semble que la seconde méthode conduise à une variance estimée légèrement inférieure. Sous les hypothèses de stabilité et d'invariance entre les groupes de rotation, à la place de la matrice des variances et des covariances, la matrice des coefficients de corrélation entre estimateurs élémentaires peut être utilisée. Dans son étude sur l'enquête emploi australienne, Bell (2001) a ainsi procédé à partir d'un modèle à quatre paramètres sur les corrélations (Bell et Carolan, 1998). Un point important à noter est que la présence ou l'absence de biais pour l'estimateur linéaire est seulement conditionnée aux contraintes posées, c'est-à-dire aux hypothèses sur les biais de rotation, et en aucune manière sur celles que l'on peut poser sur les variances ou les coefficients de corrélation.

2.1.2. Les corrélations entre estimateurs élémentaires

Le calcul des estimateurs optimaux repose entièrement sur l'estimation des variances et des coefficients de corrélation entre les estimateurs élémentaires. A partir de la méthode d'estimations des variances longitudinales, je me suis efforcée dans cette étude d'estimer le plus précisément possible les corrélations intertemporelles entre estimateurs élémentaires lorsqu'il y a renouvellement des aires. On suppose souvent qu'il n'y a aucune corrélation entre les estimations sur les rangs d'interrogation lorsqu'il y a rotation et renouvellement des logements enquêtés. Mais comme, dans l'EEC, la rotation s'effectue dans les secteurs, cette hypothèse est plus délicate à poser. Comme on l'a vu ci-dessous, deux effets se manifestent sur les covariances intertemporelles : un effet positif de grappe et un effet négatif lié à la rotation. Il y a lieu de penser que l'effet de grappe l'emporte sur l'effet de rotation et donc qu'il y ait une légère corrélation positive lorsqu'il y a renouvellement des aires. Par ailleurs, en suivant une approche « séries temporelles », les autocorrélations des estimations élémentaires ont pu

être estimé pour les données australiennes et américaines (Bell et Carolan, 1998, Pfeffermann *et al.*, 1998). Notamment, lorsqu'il y a rotation, Bell et Carolan ont obtenu pour le taux de chômage des autocorrélations de l'ordre de 0,1 jusqu'à huit mois de retard, avec une légère décroissance en fil du temps.

Dans cette étude, les coefficients de corrélation ont été estimés selon une approche purement « sondage » à partir de formules analytiques (cf. annexe 2) sous l'hypothèse d'invariance entre groupes de rotation. Les coefficients de corrélation ont été aussi estimés entre les groupes de rotation différents, c'est-à-dire correspondant à des sous-échantillons de secteurs différents. Comme attendu, ces coefficients sont tout à fait négligeables et seront posés égaux à zéro. A cause de la petitesse des échantillons et des problèmes évoqués, les estimations sont particulièrement délicates à effectuer lorsqu'il y remplacement des aires et que le retard est faible : les estimations aberrantes de covariances ont été éliminées et remplacées par les valeurs moyennes. A partir d'un retard de trois trimestres, les estimations apparaissent plus robustes. Pour un retard de quatre trimestres, les coefficients de corrélations semblent être plus élevés, ce qui correspond à un effet saisonnier assez net (cf. tableau 3). Les autres estimations sont assez voisines des résultats de Bell sur l'enquête emploi australienne, en tout cas pour les coefficients de corrélation avec remplacement d'aires. Les coefficients de corrélation sans remplacement d'aires présente une décroissance moins rapide dans l'enquête française que dans l'enquête australienne. Cela peut être dû à une attrition différente et à des modes d'interrogation différents. Il faut noter que les coefficients de corrélation obtenus sans remplacement d'aires sont plus faibles que ceux calculés directement sur les variables individuelles. Cela s'explique par un effet de grappe puisque les unités de base pour l'échantillonnage sont les aires et non les individus. Le calage contribue aussi à abaisser ces coefficients en éliminant une partie des corrélations liées aux variables de calage.

Tableau 3 : coefficients de corrélation moyens entre estimateurs élémentaires

Retard	1	2	3	4	5
Nombre de chômeurs					
Sans remplacement d'aires	0,57	0,49	0,45	0,43	0,36
Avec remplacement d'aires	0,08	0,08	0,07	0,09	0,07
Nombre d'actifs occupés					
Sans remplacement d'aires	0,76	0,71	0,68	0,66	0,61
Avec remplacement d'aires	0,08	0,08	0,08	0,10	0,07

Source : enquêtes emploi du T1 2003 au T4 2008.

L'analyse des coefficients de corrélation permet aussi d'obtenir des premiers éléments sur la détermination de la fenêtre. Comme la variance de l'estimateur optimal ne peut que décroître avec le nombre de périodes prises en compte, il est intéressant de prendre une fenêtre assez large. Comme les corrélations entre estimations par groupe de rotation diminuent assez lentement et qu'il y a un effet saisonnier non négligeable, il semble intéressant de se placer sur une fenêtre d'au moins cinq trimestres. Au-delà de six trimestres, les échantillons ne se recouvrent plus et l'avantage d'une fenêtre de plus de six trimestres paraît limité par rapport à une fenêtre moins large.

2.1.3. Prolongements

Les estimateurs linéaires considérés jusqu'ici sont des combinaisons linéaires d'estimateurs élémentaires, eux-mêmes calés sur un certain nombre de variables. Or, la taille d'un groupe de rotation est limitée et ne permet pas de caler les estimateurs élémentaires sur un grand nombre de variables, surtout quand on veut obtenir des estimations sur une zone géographique limitée, comme un État australien, ou une région française, même s'il s'agit de l'Île-de-France. Une solution, utilisée maintenant dans l'enquête emploi australienne, est de calculer une combinaison linéaire d'estimateurs élémentaires calés sur un nombre réduit de variables et de recaler l'estimateur obtenu sur un plus grand nombre de variables. Dans ce calage, toutes les données de la fenêtre sont utilisées en multipliant leur poids par 6 fois le coefficient de l'estimateur élémentaire : un ménage l enquêté au trimestre t-u en $i^{\text{ème}}$ interrogation a un poids donné par $w^{*}_{t-u,l} = 6a_i^u w_{t-u,l}$. Les totaux de référence sont les mêmes que pour l'estimateur transversal naturel.

Au lieu de minimiser la variance de l'estimateur du niveau, on peut chercher à optimiser d'autres estimateurs : celui de l'évolution entre deux périodes, celui de la moyenne annuelle, etc. Un estimateur spécifique, linéaire à fenêtre fixe, peut être calculé de la même manière que l'estimateur linéaire optimal défini précédemment en modifiant le vecteur *c* apparaissant dans les contraintes. Par exemple, pour l'évolution trimestrielle, il faut poser : $c = (1,-1,0,\ldots)$; pour une moyenne annuelle, $c = (1,1,1,1,0,\ldots)/4$. Hormis cela, le calcul ne présente aucune différence.

Une autre démarche est de calculer des estimateurs linéaires des niveaux tels que leurs combinaisons permettent d'optimiser l'estimation longitudinale privilégiée. Par exemple, pour l'évolution trimestrielle, c'est $Var(\hat{y}_{tf} - \hat{y}_{t-1,f})$ que l'on optimise, et cette variance s'écrit :

$A'U'Var(\hat{Y}_{t,f+1})UA$ où U est la matrice $6(f+1)\times 6f$ définie par blocs par : $U = \begin{pmatrix} I_{6f} \\ 0 \end{pmatrix} - \begin{pmatrix} 0 \\ I_{6f} \end{pmatrix}$.

Cela conduit à remplacer dans le calcul de l'estimateur optimal la matrice $V = Var(\hat{Y}_{tf})$ par $U'Var(\hat{Y}_{t,f+1})U$, matrice qui reste inversible.

2.2. L'estimateur AK

2.2.1. Définition

Pour éviter les difficultés de calcul des estimateurs optimaux, Gurney et Daly ont défini un estimateur récursif, l'estimateur AK :

$$\hat{y}_t^{AK} = (1-K)\hat{y}_t + K\left(\hat{y}_{t-1}^{AK} + \hat{\Delta}_{t-1,t}^{c}\right) + A\left(\hat{y}_t^{R1} - \frac{1}{5}\sum_{i=2}^{6}\hat{y}_t^{Ri}\right).$$

$\hat{\Delta}_{t-1,t}^{c}$ est l'estimateur naturel de l'évolution entre les trimestres *t-1* et *t*, basé sur la partie commune des deux échantillons trimestriels, composée des groupes de rotation où il n'y a pas de remplacement de logements : $\hat{\Delta}_{t-1,t}^{c} = \frac{1}{5}\sum_{i=2}^{6}\hat{y}_t^{Ri} - \frac{1}{5}\sum_{i=1}^{5}\hat{y}_{t-1}^{Ri}$.

Les coefficients A et K sont choisis si possible pour minimiser la variance des estimateurs. L'estimateur K, qui fut initialement utilisé dans la CPS, correspond à un estimateur AK où A est fixé à zéro, c'est-à-dire sans la troisième composante qui est une composante d'ajustement. La deuxième composante permet de prendre en compte l'évolution entre deux trimestres consécutifs, mesurée sur les unités, ici les aires, présentes aux deux périodes. Dans la CPS, jusqu'aux années 1980, un estimateur K avec un coefficient K de 0,5 a été employé pour les estimations concernant le marché du travail. Ensuite, un estimateur AK avec K = 0,4 et A = 0,2 a été mis en œuvre, les deux coefficients ayant été choisis dans le but d'avoir de meilleures estimations du chômage. Depuis 1998, une méthode de pondération composite est utilisée (Lent *et al.*, 1999). Elle consiste à calculer des estimations composites pour des catégories de chômeurs et d'actifs occupés (par sexe et âge, par race, par État). Les coefficients fixés sont K = 0,4 et A = 0,3 pour le chômage et K = 0,7 et A = 0,4 pour l'emploi. Dans une seconde étape, un recalage est effectué sur ces estimations composites pour l'ensemble de l'échantillon. Cela permet d'avoir un jeu unique de poids permettant de retrouver les estimations composites quasi optimales pour les catégories de référence. Il faut noter que cette étape de pondération composite peut être aussi utilisée avec les estimateurs linéaires optimaux.

2.2.2. Influence des biais de rotation

Le biais de l'estimateur AK ainsi qu'il est défini ci-dessus dépend fortement des biais de rotation éventuels. Si on suppose qu'ils sont nuls et que l'estimateur naturel est sans biais, on a :

$$\mathrm{E}(\hat{y}_t^{AK}) = (1-K)\,\mathrm{E}(\hat{y}_t) + K\left[\mathrm{E}(\hat{y}_{t-1}^{AK}) + \mathrm{E}(\hat{y}_t) - \mathrm{E}(\hat{y}_{t-1})\right].$$

On en déduit par récurrence que l'estimateur AK a exactement la même espérance que l'estimateur naturel[5]. Comme celui-ci est un estimateur calé, il a un biais négligeable et donc l'estimateur AK aussi.

Lorsqu'on se place sous les hypothèses (2), l'espérance de $\hat{y}_t^{AK}$ dépend entre autres de l'espérance à la période précédente :

$$\mathrm{E}(\hat{y}_t^{AK}) = \left(1 + \frac{b_1}{5}(6A - K)\right) y_t + K\left(\frac{b_6}{5} - 1\right) y_{t-1} + K\,\mathrm{E}(\hat{y}_{t-1}^{AK}).$$

Sauf à admettre des hypothèses très spécifiques, l'estimateur AK a un léger biais. En particulier, si y_t reste constant, le biais tendra assez rapidement vers une valeur non nulle. Dans le cas général, le biais sera variable d'une période à l'autre mais sauf exception, il ne sera pas nul. Un moyen pour obtenir un estimateur AK avec un biais négligeable en présence de biais de rotation est de modifier les termes $\hat{\Delta}_{t-1,t}^{c}$ et d'ajustement en divisant chaque estimateur élémentaire par $1 + b_i$:

$$\hat{y}_{bt}^{AK} = (1 - K)\hat{y}_t + K\left(\hat{y}_{b,t-1}^{AK} + \hat{\Delta}_{t-1,t}^{bc}\right) + A\beta_{bt}$$

où $\hat{\Delta}_{t-1,t}^{bc} = \frac{1}{5}\sum_{i=2}^{6}\frac{\hat{y}_t^{Ri}}{1+b_i} - \frac{1}{5}\sum_{i=1}^{5}\frac{\hat{y}_{t-1}^{Ri}}{1+b_i}$ et $\beta_{bt} = \frac{\hat{y}_t^{R1}}{1+b_1} - \frac{1}{5}\sum_{i=2}^{6}\frac{\hat{y}_t^{Ri}}{1+b_i}$.

2.2.3. L'estimateur AK vu comme un estimateur linéaire à fenêtre variable

L'estimateur AK est une combinaison linéaire d'estimateurs élémentaires définis depuis la première période. Si on se place dans l'espace des suites et en notant γ_t le vecteur des coefficients de $\hat{y}_{bt}^{AK}$ classés dans le même ordre que ceux du vecteur des estimateurs linéaires à fenêtre fixe, on a :

$$\gamma_t = (1 - K)\gamma_{nat} + K(\mathrm{d}_6(\gamma_{t-1}) + \delta) + A\beta$$

où γ_{nat} est le vecteur des coefficients de l'estimateur naturel : $\gamma_{nat} = (1/6,\ldots,1/6,0,\ldots)$ avec les six premières composantes non nulles, δ correspond au vecteur des coefficients de l'estimation de l'évolution : $\delta = \frac{1}{5}(0,1/(1+b_2),\ldots,1/(1+b_6),-1/(1+b_1),\ldots,-1/(1+b_5),0,\ldots)$; β correspond à β_{bt} : $\beta = (1/(1+b_1),-1/5(1+b_2),\ldots,-1/5(1+b_6),0,\ldots)$ et d_6 est l'opérateur de décalage à droite de 6 composantes. En posant $\xi = (1 - K)\gamma_{nat} + K\delta + A\beta$, qui ne dépend pas de *t*, on la formule de récurrence plus simple :

$$\gamma_t = K\,\mathrm{d}_6(\gamma_{t-1}) + \xi.$$

avec la valeur initiale $\gamma_1 = \gamma_{nat}$. On a donc :

$$\gamma_t = \sum_{u=0}^{t-2} K^u\,\mathrm{d}_6^u(\xi) + K^{t-1}\,\mathrm{d}_6^{t-1}(\gamma_{nat}). \qquad (3)$$

Ainsi, après u trimestres, les coefficients correspondants aux retards inférieurs à u ne varient plus et les coefficients varient en fonction du retard comme K^u et décroissent rapidement avec les valeurs usuelles fixées pour K.

La formule (3) ou la formule de récurrence permet un calcul rapide des vecteurs γ_t. Ensuite, avec toute la série des estimations élémentaires, les estimations AK s'obtiennent comme les estimations linéaires à fenêtre fixe optimales. Connaissant les matrices de variance-covariance et avec les

[5] Lors de la première période d'observation, l'estimateur AK est égal à l'estimateur naturel.

hypothèses faites sur les coefficients de corrélation au-delà de six trimestres de retard, on obtient une estimation de la grande matrice de variance-covariance des estimateurs élémentaires sur la période complète de définition des estimateurs composites. Le calcul des variances des estimateurs AK est aisé avec la donnée de cette matrice et des vecteurs γ_t. La recherche des coefficients A et K optimaux se fait ensuite de manière classique par tâtonnement.

2.3. L'estimateur par régression modifiée ou par calage composite

C'est l'estimateur actuellement utilisé dans l'enquête sur la population active au Canada, enquête qui a le même schéma de rotation que l'EEC, avec des ménages présents uniquement six périodes de suite, la périodicité de l'enquête étant le mois et non le trimestre comme en France. La définition de l'estimateur employé pour l'EPA peut donc s'appliquer telle quelle pour l'EEC[6]. Le principe de cet estimateur est d'ajouter aux variables de calage usuelles, essentiellement des variables sociodémographiques et d'autres relatives au plan de sondage, des variables se rapportant à la période précédente. Comme le calage est souvent effectué avec la fonction linéaire, cela revient à considérer l'estimateur par régression généralisée (voir Tillé, 2001) d'où le premier nom de l'estimateur, mais il est possible d'employer d'autres fonctions de calage, qui devraient amener à des résultats semblables, puisque les différentes méthodes de calage conduisent à des estimations voisines sur de grands échantillons. La terminologie n'est pas encore tout à fait fixée et le terme plus récent d'estimateur par calage composite paraît le plus approprié. Ce type d'estimateur présente l'avantage de prendre en compte les corrélations intertemporelles des variables d'intérêt tout en s'affranchissant de l'étape de la pondération composite. Un système unique de poids est obtenu à l'issue du calage composite sans qu'il y ait besoin de calculer d'estimations composites spécifiques pour un certain nombre de variables déterminées *a priori* comme cela est le cas pour l'estimateur linéaire optimal et l'estimateur AK. On peut s'attendre aussi à un gain de variance intéressant sur les évolutions.

Le problème essentiel qui apparaît immédiatement avec l'ajout de variables relatives à la période précédente est que leurs valeurs ne sont pas connues pour le sixième entrant ainsi que pour les ménages n'ayant pas répondu précédemment, notamment ceux qui viennent d'emménager. Une première solution, qui avait été expérimentée par Février et Givord (2002), est de ne garder pour les sous-échantillons non entrants que les répondants à la date précédente tandis que pour l'échantillon entrant, le calage reste inchangé. Évidemment cette solution présente l'inconvénient d'une perte d'information qui peut assez conséquente : les nouveaux ménages ne sont quasiment pas pris en compte dans les exploitations finales alors qu'ils ont des caractéristiques assez spécifiques (plus actifs que la moyenne, avec plus d'enfants notamment). Une autre solution consiste à imputer les valeurs manquantes pour toutes les nouvelles unités.

La première méthode d'imputation est l'imputation par la moyenne, ce qui donne le vecteur modifié des variables auxiliaires composites pour un ménage k :

$$z_{\bullet k}^{(1)} = \begin{cases} z_{t-1,k} & \text{si } k \in s_{tr} - s_{tr}^b \\ np_k \hat{Z}_{t-1} / N_{t-1} & \text{si } k \in s_{tr}^b \end{cases},$$

où s_{tr} est l'ensemble des ménages répondants en t, s_{tr}^b celui des nouveaux ménages répondants en t, np_k est le nombre de personnes de plus de 15 ans du ménage, N_{t-1} est le nombre total de personnes de plus de 15 ans dans les ménages ordinaires et $\hat{Z}_{t-1}$ est le vecteur des estimations des totaux des variables auxiliaires composites pour la période précédente. Ici on suppose que celles-ci sont des variables individuelles ramenées au niveau ménage, typiquement des nombres de chômeurs et d'actifs occupés. La méthode se généralise aisément à des variables ménages : au lieu du nombre de personnes de plus de 15 ans, c'est le nombre de ménages qu'il faut introduire. Le total de référence est le total estimé à la date précédente. Cette méthode permet surtout une amélioration de l'estimation en niveau.

[6] L'EPA est stratifié par province et toutes les estimations sont calées par province.

Dans la seconde méthode basée sur une imputation rétrospective, on utilise le vecteur modifié :

$$z_{\bullet k}^{(2)} = \begin{cases} z_{t-1,k} + (\delta_k^{-1} - 1)(z_{t-1,k} - z_{tk}) & \text{si } k \in s_{tr} - s_{tr}^b \\ z_{tk} & \text{si } k \in s_{tr}^b \end{cases},$$

où δ_k est la probabilité pour k d'appartenir à $s_{tr} - s_{tr}^b$ sachant $k \in s_{tr}$. Dans l'EPA et dans l'EEC, si on suppose qu'il n'y a pas de déménagement ni de non-réponse, c'est la probabilité d'être dans les rangs d'interrogation 2 à 6 étant donné l'appartenance à l'échantillon trimestriel, soit : $\delta_k = 5/6$. En général, on estime δ_k par $\hat{\delta}_k = \sum_{k \in s_{tr} - s_{tr}^b} w_k^1 \Big/ \sum_{k \in s_{tr}} w_k^1$ où les w_k^1 sont les poids obtenus après correction de la non-réponse totale ou les poids calés sur les variables de calage de la période *t*. Pour les ménages entrants, les valeurs du vecteur *z* à la période précédente sont imputées par les valeurs en *t*, mais cela oblige à modifier les valeurs pour les ménages présents aux deux dates pour conserver un biais asymptotique négligeable. Comme les valeurs aux deux périodes sont bien corrélées, la correction est en général minime, et nulle pour nombre de ménages. Cette méthode vise plutôt à améliorer l'estimation de l'évolution.

On peut utiliser une seule de ces méthodes, ce qui conduit respectivement aux estimateurs MR1 et MR2 de Singh *et al.* (1997) mais de manière analogue à celle des estimateurs linéaires optimaux, une combinaison des deux méthodes, notée MR, peut être utilisée pour la définition des variables auxiliaires composites. Ainsi dans l'EPA, est employé le vecteur défini par :

$$z_{\bullet k} = (1 - \alpha) z_{\bullet k}^{(1)} + \alpha\, z_{\bullet k}^{(2)}$$

où α est choisi afin d'obtenir un compromis sur les améliorations des estimations en niveau et en évolution. Dans l'EPA, cette constante a été fixée à 2/3.

On peut généraliser la formule précédente en remplaçant α par un vecteur ayant autant de composantes qu'il y a de variables auxiliaires composites. La difficulté du choix des différents coefficients est alors d'autant plus importante. Avec une méthode d'estimation de la variance par jackknife, comme celle qui est pratiquée dans l'EPA, cela conduit à des temps de calculs informatiques très importants. Une solution peut être de regrouper les variables auxiliaires composites par groupe, par exemple celles relatives au chômage et celles sur les actifs occupés, ce qui limiterait la recherche à celle de deux coefficients. La linéarisation de l'estimateur par régression modifiée est complexe puisqu'elle fait intervenir des linéarisées à toutes les dates depuis la première période de mise en œuvre de l'estimateur. Avec une autre fonction de calage que la fonction linéaire, on peut conjecturer qu'il en aille de même.

Par ailleurs l'estimateur MR pourrait permettre de limiter le phénomène de dérive de l'estimateur MR2 tout en conservant une partie de ses avantages. Ce phénomène de dérive a été décrit par Fuller et Rao (2001). Il se produit pour une variable composite auxiliaire $z_{tk,m}$ lorsque la corrélation entre deux périodes est élevée. Dans ce cas, la corrélation avec la variable modifiée $z_{\bullet k,m}^{(2)}$ (la composante correspondante du vecteur modifié pour le calage) est particulièrement élevée par construction, ce qui entraîne une forte influence de l'estimation $\hat{z}_{t-1,m}^{MR2}$ sur l'estimation $\hat{z}_{t,m}^{MR2}$. Ainsi toute évolution importante à un moment donné aura un effet prolongé sur les estimations pendant une assez longue période. De légers biais sur l'évolution pourront aussi se cumuler au cours du temps et entraîner un écart sensible entre les estimations usuelles et les estimations MR2. Ce phénomène de dérive est moins prononcé sur l'estimateur MR1 puisqu'une partie des valeurs de la variable de calage modifiée $z_{\bullet k,m}^{(1)}$ est obtenue à partir d'une imputation par la moyenne en *t*-1, ce qui contribue à abaisser la corrélation de cette variable avec la variable d'intérêt $z_{tk,m}$. La combinaison des deux méthodes devrait donc limiter la dérive éventuelle de l'estimateur. Cependant, les calculs effectués par Bell (2001) sur les données de l'enquête emploi australienne montrent que l'estimateur MR avec α = 0,7 reste proche de l'estimateur MR2 et qu'une dérive importante est présente : sur le nombre de chômeurs, des écarts persistants de l'ordre de 3% avec l'estimateur naturel sont observés.

3. Comparaison des méthodes

3.1. Points de comparaison

Les estimations ont été calculées pour les deux premières méthodes sur la période allant du T1 2004 au T4 2008 pour les deux premières méthodes. Les estimateurs linéaires optimaux ont été calculés pour une fenêtre de six trimestres. Seules deux grandeurs ont été considérées : les nombres de chômeurs et d'actifs occupés en France métropolitaine. Un prolongement de ce travail consisterait à comparer des estimations d'autres totaux : nombres de femmes chômeuses, actives occupées, d'hommes chômeurs, actifs occupés, de chômeurs âgés de 15 à 24 ans, etc. Pour chaque variable, deux versions ont été considérées selon la prise en compte du biais de rotation. Pour les estimateurs AK, les coefficients permettant d'avoir des variances estimées les plus faibles en moyenne sont :

- pour le nombre de chômeurs, K = 0,6 et A = 0,08 sans la prise en compte du biais de rotation, K = 0,55 et A = 0,08 avec prise en compte du biais de rotation ;
- pour le nombre d'actifs occupés, K = 0,75 et A = 0,08 avec et sans la prise en compte du biais de rotation.

La comparaison repose avant tout sur celle des écarts types : l'intérêt des estimateurs composites et la raison principale de leur mise en œuvre résident dans l'amélioration des écarts types en niveau (valeur trimestrielle et moyenne annuelle) et en évolution (entre deux trimestres consécutifs ou sur un an, c'est-à-dire entre t-4 et t). Les estimations de variance pour les estimateurs linéaires à fenêtre fixe optimaux et les estimateurs AK reposent entièrement sur le calcul des variances et des covariances des estimateurs élémentaires selon les formules données précédemment. Pour l'estimateur linéaire optimal, ce calcul est consubstantiel à la recherche des coefficients optimaux. Comme l'estimateur AK est un estimateur linéaire défini sur la fenêtre complète des observations, l'estimation de sa variance nécessite d'avoir des covariances avec des retards supérieurs à six trimestres et donc de prolonger les données du tableau 3. On postule en tendance une lente décroissance des coefficients de corrélation au cours du temps conformément à ce qui est observé jusqu'à cinq trimestres d'écart. Pour des retards multiples de quatre, cette tendance est contrecarrée par un effet saisonnier et on suppose qu'il y a alors une augmentation des coefficients de corrélation. Cependant, comme la décroissance des coefficients élémentaires de l'estimateur AK est rapide, l'estimation de sa variance est peu sensible aux hypothèses sur les covariances. Pour l'estimateur par régression modifiée, le calcul de sa précision par des formules analytiques requiert celui des linéarisées, qui se fait par récurrence, et les estimations des covariances entre ces linéarisées. Ces calculs sont particulièrement voraces en temps et n'ont pas pu être effectués dans le cadre de cette étude.

Un autre critère de comparaison est la présence d'une dérive, surtout présente pour l'estimateur par régression modifiée mais qui pourrait apparaître pour les autres estimateurs composites. Elle se manifeste par des écarts avec l'estimateur naturel importants et persistant pendant plusieurs périodes. Les estimations produites ne sont pas corrigées des variations saisonnières.

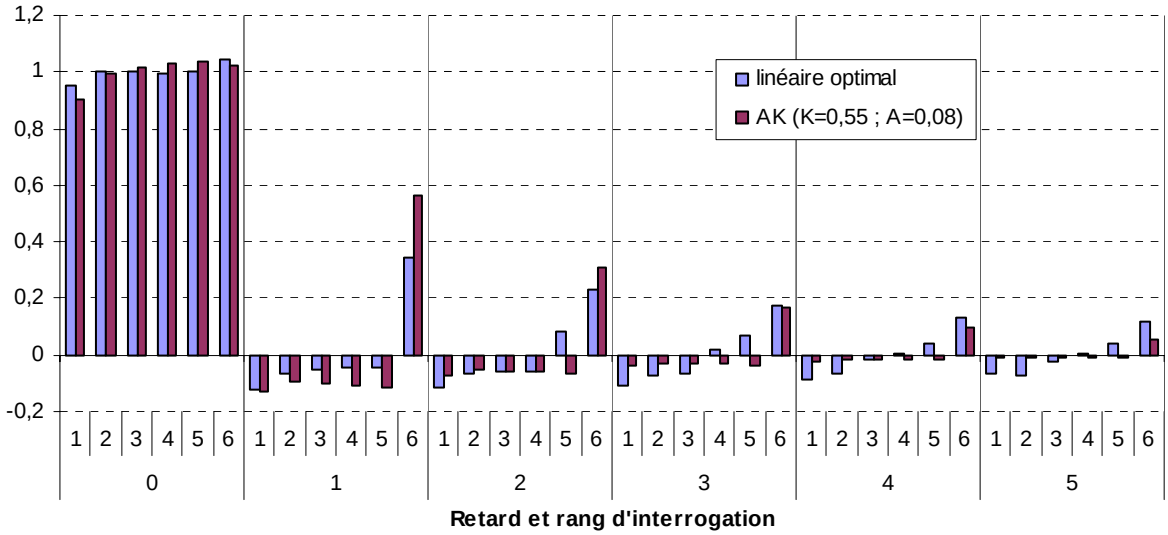
3.2. Coefficients

En les multipliant par six, les coefficients des estimateurs élémentaires s'analysent directement comme des multiplicateurs des poids actuels des ménages en faisant intervenir une contribution de ceux interrogés aux trimestres précédents. Pour les estimateurs AK, les coefficients des estimateurs élémentaires sont calculées selon la formule (3). Bien que la fenêtre de cet estimateur soit variable, les multiplicateurs sont donnés jusqu'à un retard de cinq trimestres correspondant à la fenêtre des estimateurs linéaires optimaux. On constate que pour les trimestres retardés, les sixièmes rangs d'interrogation sont pondérés de manière plus importante que les autres rangs d'interrogation et toujours positivement (cf. graphique 1). Par exemple, dans l'estimateur AK, les ménages du sixième rang d'interrogation avec un retard de un trimestre interviennent pour plus de la moitié de leur poids trimestriel. Cela permet d'incorporer une information corrélée avec le niveau de la variable au trimestre courant et qui n'est plus disponible aux trimestres suivants. Avec l'estimateur optimal, les trimestres avec un retard important, supérieur à quatre, gardent une contribution non négligeable au contraire de l'estimateur AK. Cela constitue une justification supplémentaire pour se placer sur une fenêtre d'au moins cinq trimestres.

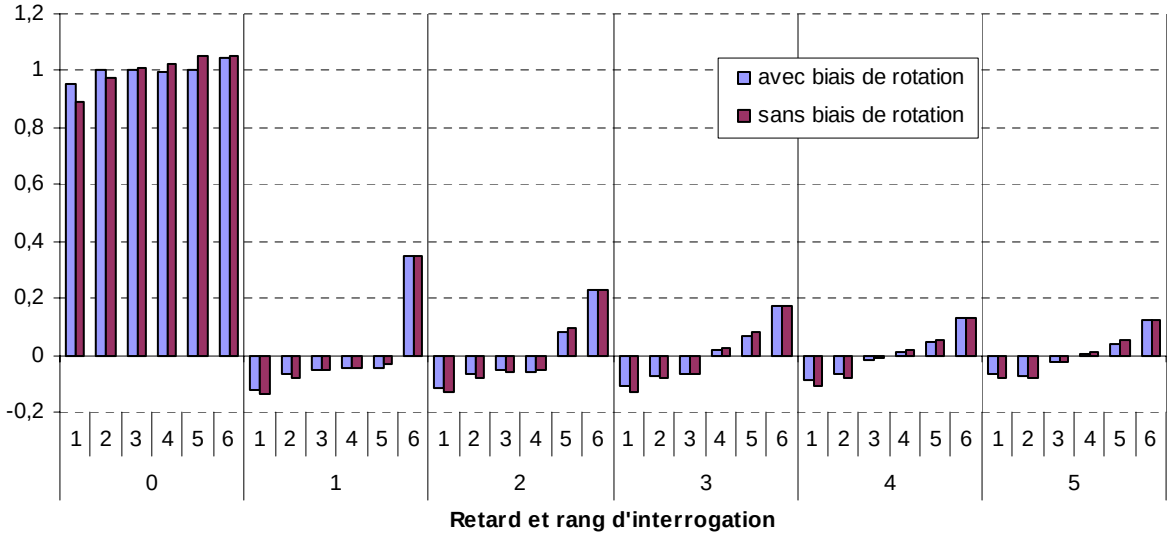
Pour l'estimateur linéaire optimal, la prise en compte du biais de rotation contribue à uniformiser les repondérations des rangs d'interrogations intermédiaires et à relever la pondération du premier rang d'interrogation.

Les repondérations sont plus importantes pour le nombre d'actifs occupés à cause des corrélations intertemporelles plus fortes entre estimateurs élémentaires (cf. graphique 3). Même au cinquième trimestre de retard, le sixième rang d'interrogation conserve un poids de plus de 25% de son poids trimestriel initial. La prise en compte du biais de rotation ne modifie ici que peu les coefficients.

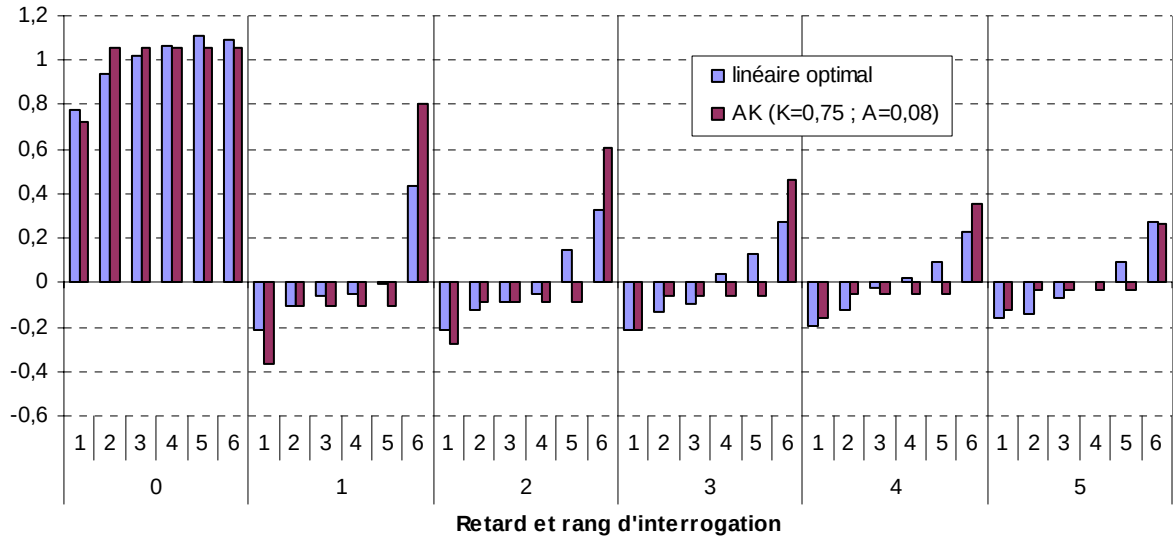
Graphique 1 : multiplicateurs des poids pour le nombre de chômeurs avec prise en compte du biais de rotation



Graphique 2 : comparaison des multiplicateurs des poids pour l'estimateur linéaire optimal du nombre de chômeurs selon la prise en compte ou non du biais de rotation



Graphique 3 : multiplicateurs des poids pour le nombre d'actifs occupés sans prise en compte du biais de rotation

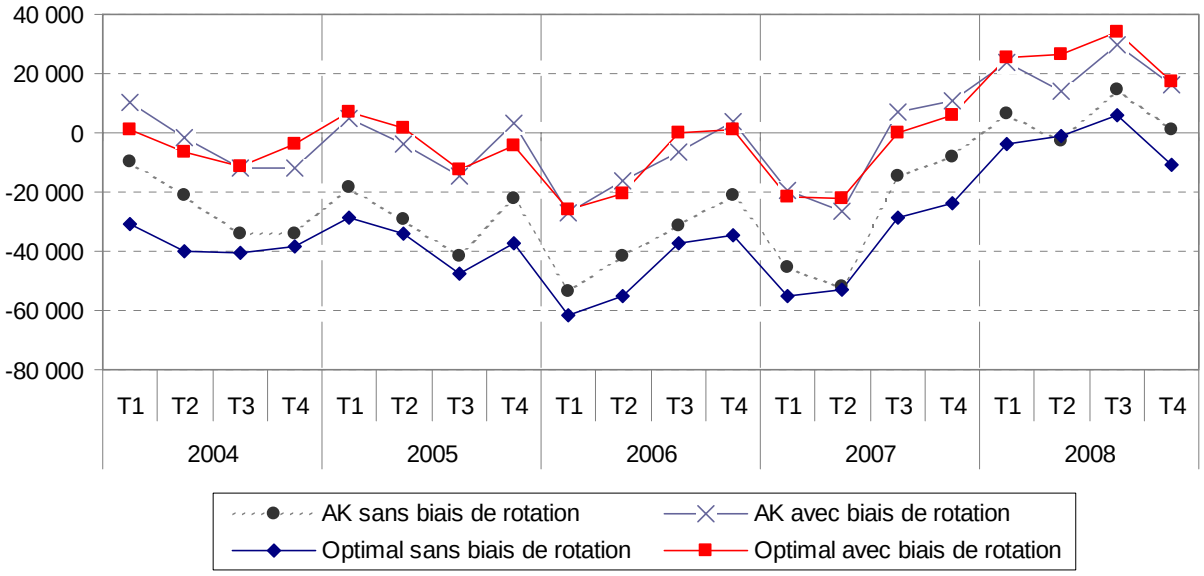


3.3. Écarts avec l'estimateur actuel

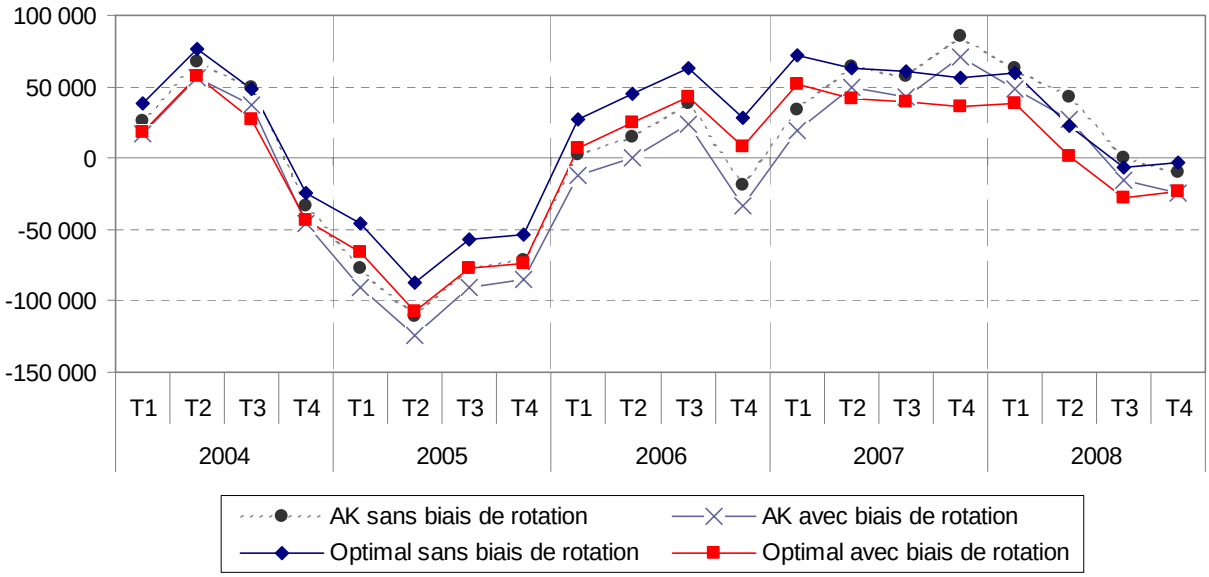
Plutôt que de montrer les graphiques des différentes estimations alternatives, il est plus intéressant de montrer les écarts avec l'estimateur actuel, qui est un estimateur purement transversal avec des calages séparés par groupe de rotation (graphiques 4 et 5). Afin de mettre ces écarts en perspective, il faut les comparer avec les niveaux des écarts types de l'estimateur actuel : au T4 2008, ils étaient de 46 000 pour le nombre de chômeurs et de 85 000 pour le nombre d'actifs occupés.

Les estimateurs linéaires et AK donnent des résultats très semblables surtout pour le nombre de chômeurs, les différences apparaissant un peu plus importantes pour le nombre d'actifs occupés. La prise en compte du biais de rotation ne change pas l'évolution des écarts mais contribue à ajouter un écart à peu près constant. Pour l'estimateur linéaire optimal du nombre de chômeurs, cet écart est d'environ 30 000, pour le nombre d'actifs occupés, il est plus réduit et négatif, de l'ordre de –20 000. Si on considère les estimateurs construits sous l'hypothèse de biais de rotation, on constate que les différences pour le nombre de chômeurs sont peu marquées de 2004 à 2007, mais qu'elles le sont un peu plus en 2008. En revanche, pour le nombre d'actifs occupés, ces écarts sont marqués en 2005, où au deuxième trimestre, ils sont supérieurs à un écart type. En 2006 et 2007, c'est plutôt le contraire qui se produit. Il y a pas de variation très brusque des écarts ni de phénomène de dérive comme cela peut se constater avec les estimateurs MR2 et MR dans l'enquête emploi australienne.

Graphique 4 : Écarts sur le nombre de chômeurs



Graphique 5 : Écarts sur le nombre d'actifs occupés



3.4. Les écarts types

Les écarts types des estimateurs sous hypothèse de biais de rotation sont présentés dans le graphique 6 et le tableau 4 pour des estimations en niveau et en évolution :

- le total trimestriel ;
- la moyenne annuelle ;
- l'évolution trimestrielle entre deux trimestres consécutifs ;
- l'évolution sur un an entre les trimestres T-4 et T.

Dans le graphique, ils sont calculés en pourcentage des écarts types des estimations actuelles et la moyenne de ces pourcentages est calculée pour la période sur laquelle les estimations ont été produites : du T1 2004 au T4 2008 pour le niveau trimestriel, du T1 2005 au T4 2008 pour l'évolution sur un an. Ces pourcentages sont stables au cours du temps.

Graphique 6 : Écarts types en pourcentage moyen des écarts types actuels

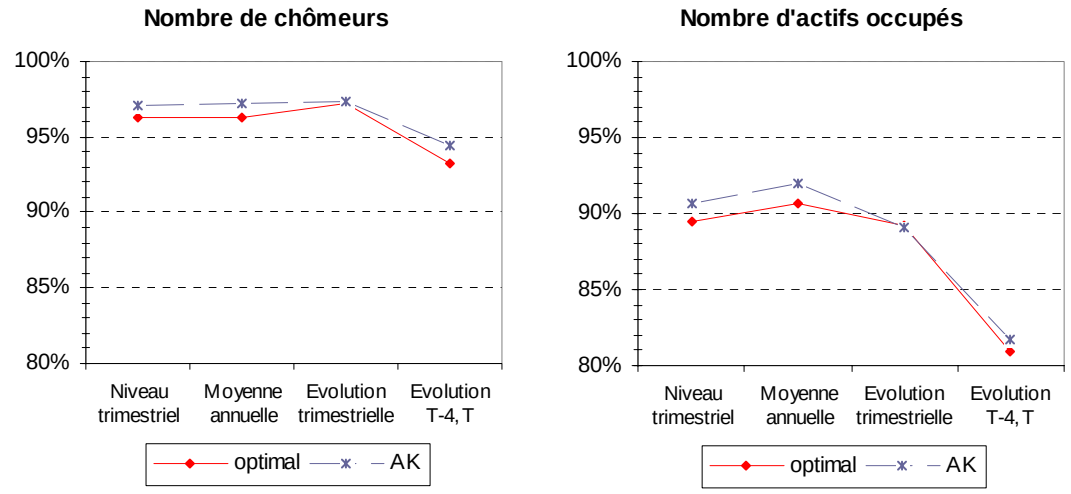


Tableau 4 : Écarts types moyens en milliers

	Nombre de chômeurs			Nombre d'actifs occupés		
Estimateurs	Lin. opt.	AK	Actuel	Lin. opt.	AK	Actuel
Niveau trimestriel	45,3	45,6	47,0	75,9	76,8	84,7
Moyenne annuelle	34,3	34,6	35,6	63,8	64,8	70,4
Évolution trimestrielle	45,1	45, 1	46,4	59,4	59,4	66,6
Évolution T-4,T	54,3	55,0	58,3	81,2	82,0	100,1

Champ : population des ménages de France métropolitaine.

Les précisions des estimateurs AK et linéaires optimaux sont très voisines. Si les estimateurs linéaires optimaux apparaissent toujours un peu plus précis, les estimateurs AK, malgré la forme particulière imposée aux coefficients des estimateurs élémentaires, en constituent de bonnes approximations. Pour le nombre de chômeurs, les gains de précision sont modestes, de 4% en niveau et de 7% pour l'évolution sur un an. Ces gains sont identiques pour le taux de chômage puisque les coefficients de variation des estimations du taux de chômage sont quasiment les mêmes que ceux des estimations du nombre de chômeurs. Pour le nombre d'actifs occupés, les gains sont plus importants à cause des corrélations plus fortes entre les estimations trimestrielles naturelles. Ils sont de l'ordre de 10% pour le niveau et comme pour le nombre de chômeurs, il est plus intéressant pour l'évolution T-4,T, à 19%.

Ces résultats sont du même ordre que ceux obtenus pour l'enquête emploi australienne, compte tenu du fait qu'elle est mensuelle et donc qu'elle présente des corrélations intertemporelles plus élevées (Bell, 2001). Les rapports des précisions pour le nombre de chômeurs y pouvaient être chiffrés à 95% en niveau pour un estimateur comparable, noté B1, et à 91% pour l'évolution de la moyenne trimestrielle. Pour le nombre d'actifs occupés, ils sont là aussi plus intéressants, de l'ordre de 90% en niveau et de 80% pour l'évolution de la moyenne trimestrielle. Bell a obtenu des gains plus importants pour les estimateurs par régression modifiée, MR et MR2, l'écart type étant à 86% de l'écart type actuel pour l'évolution mensuelle du nombre de chômeurs mais restant de l'ordre de 95% pour le niveau mensuel, et sensiblement égal à l'écart type actuel pour la moyenne trimestrielle. Cependant, comme ces estimateurs présentent un risque élevé de biais et de dérive, c'est finalement une variante d'un estimateur linéaire optimal défini sur une fenêtre de sept mois qui a été adopté dans l'enquête emploi australienne en 2007.

4. Conclusion

Les estimateurs composites évalués ici présentent des gains en précision assez limités sur le chômage mais plus intéressants pour l'emploi. Les gains sont toujours plus importants pour la mesure d'une évolution sur un an puisque ces estimateurs font intervenir un nombre important de périodes. Les estimateurs AK et les estimateurs linéaires optimaux avec une fenêtre de six trimestres présentent les mêmes propriétés et les mêmes difficultés de mise en œuvre. Même si les estimateurs linéaires optimaux considérés ici nécessitent l'agrégation des données sur six trimestres, cela ne constitue pas réellement une difficulté supplémentaire par rapport à l'estimateur AK qui, du fait de sa définition récursive, ne nécessite *a priori* que les données et l'estimation du trimestre précédent en plus de celles du trimestre courant. Cependant, l'estimation de la variance est du même ordre de complexité et se base dans les deux cas sur les données de plusieurs trimestres.

La présence de biais de rotation n'est pas en mesure de bouleverser les évolutions des différentes estimations et il ne se traduit que par un écart supplémentaire par rapport à l'estimateur naturel. Un prolongement de ce travail serait de calculer des estimations composites sur certaines catégories : les femmes, les hommes, les hommes, les régions les plus peuplées comme l'Île-de-France ou PACA, etc. Cela permettrait de fixer des estimations cibles qui, avec la procédure de pondération composite, seraient incorporées dans les poids diffusés.

Annexe 1 : Estimations de la variance panélisée

La formule théorique de la variance de l'estimateur naturel de l'évolution entre deux trimestres, $\Delta_{tt'} = Y_{t'} - Y_t$, est la suivante :

$$Var(\hat{\Delta}_{tt'}) = Var(\hat{\Delta}_{tt'}^{panel}) + 2(1-o_{tt'})\sum_{i\in U_I}\frac{N_i^2}{\pi_i}S_{i,tt'}$$

où U_I est l'échantillon complet de secteurs, N_i est le nombre d'aires dans le secteur i, $Var(\hat{\Delta}_{tt'}^{panel})$ est la variance de l'estimateur panélisé, c'est-à-dire celui que l'on aurait si les mêmes aires étaient enquêtées aux deux périodes, et $S_{i,tt'}$ est la covariance intra-secteur des totaux par aires entre les deux trimestres.

Le second terme s'estime par le biais d'imputations tandis que la variance panélisée s'estime à partir de la variance d'un estimateur cylindré. Celui-ci est défini sur le recouvrement des deux échantillons trimestriels d'aires $s_{cyl} = s_t \cap s_{t'}$:

$$\hat{\Delta}_{tt'}^{cyl} = \frac{1}{o_{tt'}}\sum_{k\in s_{cyl}}\frac{y_{kt'} - y_{kt}}{\pi_k}.$$

C'est l'estimateur de Horvitz-Thompson de la variable $\delta_k = y_{kt'} - y_{kt}$ calculée à partir des totaux des linéarisées par aires. Comme s_{cyl} regroupe plusieurs rangs d'interrogations, il est obtenu à partir de l'échantillon complet par un sondage aléatoire simple au taux $o_{tt'}$. En utilisant les résultats pour les plans à deux phases où la seconde phase est un sondage aléatoire simple (voir par exemple Tillé, 2001), on a :

$$\hat{Var}(\hat{\Delta}_{tt'}^{cyl}) = \hat{Var}(\hat{\Delta}_{tt'}^{panel}) + N\left(\frac{1}{o_{tt'}} - 1\right)S_{cyl}^2$$

où S^2_{cyl} est la variance empirique de la variable $\breve{\delta}_k = \delta_k / \pi_k$, calculée sur l'échantillon cylindré : $S^2_{cyl} = \frac{1}{o_{tt'}N-1}\sum_{k\in s_{cyl}}(\breve{\delta}_k - \breve{\bar{\delta}})^2$. $\hat{V}ar(\hat{\Delta}^{cyl}_{tt'})$ se calcule directement comme une variance transversale. Ensuite $\hat{V}ar(\hat{\Delta}^{panel}_{tt'})$ est calculée par différence avec la formule précédente. On obtient en général que le rapport $\hat{V}ar(\hat{\Delta}^{panel}_{tt'}) / \hat{V}ar(\hat{\Delta}^{cyl}_{tt'})$ est proche de $o_{tt'}$.

Annexe 2 : Estimation des variances et des covariances des estimateurs élémentaires

Pour calculer toutes les variances et les covariances, on fait une hypothèse d'invariance entre les groupes de rotation. Ainsi les variances transversales des six estimateurs élémentaires sont supposées égales ainsi que les covariances entre deux estimateurs élémentaires d'un même trimestre.

Soit $\hat{V} = \hat{V}ar(\hat{y}_t)$ la variance estimée de l'estimateur naturel du total et $S^2_t = \frac{1}{N-1}\sum_{s_t}\left(\breve{y}_{kt} - \breve{\bar{y}}_t\right)^2$

où $\breve{y}_{kt} = y_{kt} / \pi_k$ est le total (de la linéarisée) par aire divisé par la probabilité d'inclusion de l'aire. Alors on a :

- $\hat{V}ar(\hat{y}^{Ri}_t) = \hat{V} + 5NS^2_t$
- $\hat{C}ov(\hat{y}^{Ri}_t, \hat{y}^{Rj}_t) = \hat{V} - NS^2_t$ si $i \neq j$.

L'estimation des covariances transversales ne pose pas de problèmes spécifiques et donne comme attendu des résultats proches de zéro.

Soit t et t' tels que $t-6 < t' < t$. On introduit $C^{panel}_{tt'}$ la covariance panélisée qui est définie de manière similaire à la variance panélisée dans l'annexe 1. Son estimation se calcule comme celle de la variance panélisée d'une évolution :

$$\hat{C}^{panel}_{tt'} = \hat{C}^{cyl}_{tt'} - N\left(\frac{1}{o_{tt'}} - 1\right)S_{tt'}$$

où $S_{tt'} = \frac{1}{o_{tt'}N-1}\sum_{s_t \cap s_{t'}}(\breve{y}_{kt} - \breve{\bar{y}}_t)(\breve{y}_{kt'} - \breve{\bar{y}}_{t'})$ est la covariance empirique calculée sur l'échantillon cylindré.

Ensuite, comme on a un plan à deux phases, on peut obtenir des estimations de la covariance selon que l'on est sur le même groupe de rotation et qu'il y a eu rotation ou non :

- sur le même groupe de rotation sans rotation Ri et Rj avec $i = j + t - t'$:
$$\hat{C}ov(\hat{y}^{Ri}_t, \hat{y}^{Rj}_{t'}) = \hat{C}^{panel}_{tt'} + 5NS_{tt'} \ ;$$
- sur deux groupes de rotation non correspondant :
$$\hat{C}ov(\hat{y}^{Ri}_t, \hat{y}^{Rj}_{t'}) = \hat{C}^{panel}_{tt'} - NS_{tt'} \ ;$$
- avec rotation, i.e. si $i = j + t - t' - 6$ et $j \geq 7 + t' - t$:
$$\hat{C}ov(\hat{y}^{Ri}_t, \hat{y}^{Rj}_{t'}) = \hat{C}^{panel}_{tt'} + 5NS_{tt'} - \frac{6}{1-o_{tt'}}\sum_{i\in s^{rot}_I}\frac{N^2}{\pi^2_i}\hat{S}_{itt'}$$

  où $\hat{S}_{itt'}$ est l'estimation de la covariance intra-secteur du secteur i et s^{rot}_I l'échantillon des secteurs entre les deux trimestres t et t'.

Bibliographie

[1] Ardilly P. et Osier G. (2007). « Cross-sectional variance estimation for the French Labor Force Survey », *Survey Research Methods* [Online] 1:2. Available: http://w4.ub.uni-konstanz.de/srm/article/view/77

[2] Bailar B.A. (1975). « The effect of rotation group bias on estimates from panel surveys », *Journal of American Statistical Association*, 70, pp. 23-29.

[3] Bell P.A. et Carolan A. (1998). « Trend estimation for small areas from a continuing survey with controlled sample overlap », Working Papers in *Econometrics and Applied Statistics*, n° 98/1, cat. n° 1351.0, ABS, Canberra.

[4] Bell P.A. (1998). « Using state space models and composite estimation to measure the effects of telephone interviewing on labour force estimates », Working Papers in *Econometrics and Applied Statistics*, n° 98/2, cat. n° 1351.0, ABS, Canberra.

[5] Bell P.A. (2001). « Comparison of Alternative Labour Force Survey Estimators », *Survey Methodology*, 16, pp. 167-180.

[6] Caron N. et Ravalet P. (2002). « Estimation dans les enquêtes répétées : application à l'enquête emploi en continu » in Actes des journées de méthodologie statistique des 4 et 5 décembre 2000, *Insee Méthodes*, n° 100, pp. 327-391.

[7] Christine M. (2002). « La construction de la future enquête emploi en continu à partir du recensement de 1999 », in Actes des journées de méthodologie statistique des 4 et 5 décembre 2000, *Insee Méthodes*, n° 100, pp. 175-229.

[8] Février P. et Givord P. (2002), « Repondérations dans la nouvelle enquête emploi en continu », article présenté aux VIII^{es} journées de méthodologie statistique de l'Insee.

[9] Fuller W.A. (1990). « Analysis of repeated surveys », *Survey Methodology*, 16, pp. 167-180.

[10] Fuller W.A et Rao J.N.K. (2001). « A Regression Composite Estimator with Application to the Canadian Labour Force Survey », *Survey Methodology*, 27, pp. 45-51.

[11] Goga C., Deville J.-C. et Ruiz-Gazen A. (2006). « Linéarisation par la fonction d'influence pour des données issues de deux échantillons », in *Méthodes d'enquêtes et sondages*, Paris, Dunod, pp. 382-387.

[12] Goux D. (2005). « L'impact des changements intervenus dans l'enquête Emploi en 2003 sur la qualité de ses résultats », article présenté aux IX^{es} journées de méthodologie statistique de l'Insee, Paris, France.

[13] Gurney M. et Daly J.F. (1965). « A multivariate approach to estimation in periodic sample surveys », *Proceedings of the Survey Research Methods Section*, American Statistical Association, pp. 247-257.

[14] Jessen R.J. (1942). « Statistical investigation of a farm survey for obtaining farm facts », *Iowa Agricultural station research Bulletin*, 304.

[15] Lent J., Miller S., Cantwell P. et Duff M. (1999). « Effects of composite weights on some estimates from the current population survey », Journal of Official Statistics, vol. 15, n°3, pp.431-448.

[16] Pfeffermann D., Feder M. et Signorelli D. (1998). « Estimation of autocorrelations of survey errors with application to trend estimation in small areas », *Journal of Business & Economic Statistics*, vol. 16, n°3, pp. 339-348.

[17] Place D. (2008). « Calcul de la précision des estimations longitudinales dans l'Enquête Emploi en Continu », in *Méthodes de Sondage*, P. Guilbert, D. Haziza, A. Ruiz-Gazen, Y. Tillé, Paris, Dunod,

[18] Singh A.C. et Merkouris P. (1995). « Composite estimation by modified regression for repeated surveys », *Proceedings of the Survey Research Methods Section*, American Statistical Association, pp 420-425.

[19] Singh A.C., Kennedy B., Wu S. et Brisebois F. (1997) « Composite estimation for the Canadian Labour Force Survey », *Proceedings of the Survey Research Methods Section*, American Statistical Association, pp 300-305.

[20] Tillé Y. (2001). *Théorie des sondages : Échantillonnage et estimation en populations finies : cours et exercices*, Paris, Dunod.