ESTIMATION DANS LES ENQUÊTES REPÉTÉES : APPLICATION A L'ENQUÊTE EMPLOI EN CONTINU

N. CARON [*] *et P. RAVALET* [**]

[*] INSEE, CREST - Rennes, Laboratoire de Statistique d'Enquêtes
[**] INSEE, Unité Méthodes Statistiques

RÉSUMÉ

Après avoir décrit les principaux types d'enquêtes répétées et leurs objectifs, nous nous intéressons plus particulièrement à l'estimation du niveau et de l'évolution d'une caractéristique de la population à partir d'une enquête à échantillon rotatif à un niveau. Nous présentons les estimateurs les plus fréquemment utilisés et nous comparons leurs performances exprimées en termes de variance selon le taux de renouvellement de l'échantillon et les caractéristiques de la variable d'intérêt. Nous nous limitons dans un premier temps au cas de deux périodes et d'un plan de sondage aléatoire simple. Les résultats sont ensuite étendus à plusieurs périodes et nous examinons le passage à un plan de sondage complexe.

Nous étudions l'application de ces techniques dans le cadre de la future enquête Emploi française en continu, en particulier pour l'estimation du niveau et de l'évolution - en glissements trimestriel et annuel - de l'emploi et du chômage. Les performances des estimateurs ont été simulées à partir des données du dispositif léger mis en place en juillet 1998 et servant de test à l'enquête en vraie grandeur. Les estimateurs composites, du type estimateur K, offrent des gains de précision appréciables pour l'estimation des évolutions et plus modestes pour les niveaux. Ces gains sont d'autant plus importants que la variable d'intérêt présente une corrélation temporelle élevée. Enfin, la forme optimale d'un estimateur composite étant fonction de la variable étudiée, le respect d'une cohérence entre les différentes variables impose une forme unique pour l'estimateur et tend à diminuer les gains associés.

MOTS CLÉS : Enquêtes répétées ; échantillons rotatifs ; estimations d'évolution ; estimateurs composites ; calcul de précision ; enquête Emploi en continu.

1. Introduction

Pour suivre les évolutions de certaines grandeurs économiques (nombre de chômeurs, revenu moyen,...), la méthode habituelle consiste à réaliser une série d'enquêtes répétées à des intervalles de temps en général réguliers. La conception et l'exploitation de ce type d'enquête se heurtent aux mêmes difficultés que pour une enquête ponctuelle, mais le caractère répétitif de l'opération tend à en accroître l'ampleur. En particulier, la composition temporelle des échantillons est un élément fondamental. Ainsi, pour améliorer la précision des évolutions entre deux périodes, on a intérêt à conserver le même échantillon, c'est-à-dire à ré-enquêter les mêmes unités statistiques. En revanche, si on s'intéresse au niveau de la variable sur une période, il est préférable de renouveler partiellement l'échantillon. Les objectifs du concepteur d'enquête étant en général multiples, un compromis est alors nécessaire. La technique des échantillons rotatifs qui consiste en un renouvellement partiel de l'échantillon à chaque période offre une solution satisfaisante pour l'estimation simultanée de niveaux (instantané ou moyen) et d'évolutions. En effet, elle permet de tenir compte du lien éventuel existant entre les valeurs successives mesurées sur un même individu pour construire des estimateurs de meilleure précision. Cette technique a été mise en oeuvre dans de nombreux pays, notamment dans le domaine des enquêtes sur les forces de travail.

Dans la première partie de ce document, nous revenons en détail sur les différents types d'enquêtes répétées - panels, enquêtes à échantillon rotatif à un ou plusieurs niveaux, à échantillons partagés, à échantillons distincts - en insistant plus particulièrement sur les avantages et les inconvénients de ces dispositifs selon les objectifs recherchés. Nous nous intéressons ensuite à l'estimation du niveau et de l'évolution d'une caractéristique de la population dans le cas d'une enquête à échantillon rotatif à un niveau. Dans un premier temps, on se place dans l'hypothèse d'un tirage aléatoire simple et on se limite au cas de deux périodes. Plusieurs estimateurs sont proposés : l'estimateur naturel, l'estimateur de l'évolution calculé sur la partie commune des échantillons correspondant aux deux périodes, un estimateur composite direct de l'évolution et la différence des estimateurs composites des niveaux. Les performances relatives de ces estimateurs mesurées en gain de variance par rapport à l'estimateur naturel ont été comparées pour une configuration proche de celle de la future enquête Emploi française en continu dans le cas des variables emploi et chômage. Le dispositif léger, mis en place en juillet 1998, a permis de caractériser le comportement temporel de ces deux variables d'intérêt considérées et d'obtenir ainsi une première idée des performances attendues des différents estimateurs.

Pour le calcul d'évolutions - en glissements trimestriel et annuel -, se restreindre à la partie commune de l'échantillon dans le cas de deux périodes peut conduire à des

résultats dramatiques (c'est-à-dire moins précis que l'estimateur naturel) si le taux de recouvrement et la corrélation sont faibles. C'est le cas de l'évolution du chômage en glissement annuel avec un taux de recouvrement de 33% et une corrélation de 0.45. En revanche, les estimateurs composites apportent un gain systématique et appréciable. Ainsi, un estimateur composite du niveau du type estimateur par régression offre des gains de précision de l'ordre de 15% sur la variable emploi et 6% sur la variable chômage. Pour la mesure des évolutions, l'estimateur composite de l'évolution est uniformément meilleur que la différence des estimations composites des niveaux. On montre d'ailleurs que c'est le meilleur estimateur linéaire sans biais. Ainsi, pour l'évolution trimestrielle de l'emploi, cet estimateur est 2.5 fois plus précis que l'estimateur naturel ; dans le cas du chômage, le gain en précision est plus faible (environ 16%).

Nous constatons aussi que la forme optimale de l'estimateur composite dépend fortement de la corrélation temporelle de la variable étudiée. La solution idéale consiste donc à choisir la forme de l'estimateur en fonction de la variable étudiée. Cependant, le respect d'une cohérence entre les différentes variables impose une forme unique pour l'estimateur ce qui par conséquent tend à diminuer les gains associés.

Les résultats sont ensuite généralisés au cas de plusieurs périodes. L'écriture des estimateurs composites ne se complique pas lorsqu'itérée sur plusieurs périodes. Cependant, ils ne sont en général plus optimaux ; ils constituent néanmoins une approximation tout à fait acceptable dans la pluart des cas. Deux estimateurs sont présentés : l'estimateur de Patterson (optimal sous une forme particulière de corrélation) et l'estimateur composite K (utilisé par exemple dans l'enquête américaine *Current Population Survey*). Le premier correspond à une moyenne pondérée entre l'estimateur naturel et un estimateur par régression sur la partie commune et le second à une moyenne pondérée entre l'estimateur naturel et un estimateur qui chaîne l'évolution mesurée sur la partie commune au niveau de la période précédente. La performance de l'estimateur K a été calculée sur les niveaux et évolutions des variables Emploi et Chômage. Pour la variable Emploi, l'efficacité (par rapport à l'estimateur naturel) de l'estimateur K optimal est de 164% pour le niveau et de 248% pour l'évolution trimestrielle. Dans le cas du chômage, ces efficacités sont respectivement de 106% et 114%.

Finalement, ces premiers résultats obtenus d'après le dispositif léger suggèrent pour la future enquête emploi en continu l'utilisation d'un estimateur K, de forme unique pour respecter la cohérence entre les variables. Dans la recherche de cet estimateur commun, on pourrait, comme cela est proposé dans cet article, favoriser l'estimation du chômage pour laquelle les gains sont les plus faibles. Sous les hypothèses retenues, le gain par rapport à l'estimateur naturel serait de l'ordre de 6% pour le niveau du chômage et de 15% pour le glissement annuel et l'évolution trimestrielle. Ceux-ci seraient plus importants pour l'emploi puisqu'ils atteindraient 31% pour le niveau et plus de 60% pour les évolutions trimestrielles et annuelles. Ces résultats

montrent aussi qu'une précision égale à celle de l'estimateur naturel peut être obtenue en appliquant l'estimateur K sur un échantillon de taille plus petite, ce qui permet de réduire le budget d'enquête.

Cependant, ces résultats ont été établis sous des hypothèses restrictives et leur transposition directe au cadre de l'enquête emploi en continu s'avère délicate. En particulier, nous avons considéré que les individus ont été directement sélectionnés dans une base de sondage d'individus par un plan de sondage aléatoire simple. Or, le plan de sondage de la future enquête emploi est relativement différent puisqu'il consiste en un sondage en grappe de logements dans lesquels l'ensemble des individus ont été retenus pour l'enquête. De plus, nous avons supposé qu'à un même individu n'était associé qu'un seul poids permettant d'extrapoler toutes les données le concernant (recueillies au sein des différentes enquêtes). Par conséquent, nous n'avons pas tenu compte des méthodes basées sur des informations auxiliaires qui corrigent la non-réponse et les fluctuations d'échantillonnage et conduisent à des pondérations différentes selon la date d'enquête considérée. Ainsi, s'il apparaît a priori possible de transposer directement les résultats qualitatifs obtenus dans cette étude au cas de la future enquête emploi en continu, il nous semble au contraire plus délicat d'extrapoler les résultats chiffrés.

Dans une dernière partie, on donne quelques éléments sur la généralisation des résultats au cas d'un plan de sondage complexe. La forme des principaux estimateurs ainsi que le calcul de leur précision s'adaptent facilement au cas d'un sondage complexe. Cependant, l'estimation précise des variances nécessite de disposer d'un logiciel évaluant la précision de totaux estimés à partir d'enquêtes par sondage complexes. Dans le cas de deux périodes, la comparaison relative des performances des estimateurs revient à celle des estimateurs définis dans le cadre d'un sondage aléatoire simple. Il suffit d'introduire la notion de « design effect » correspondant par définition au rapport de la variance issue du plan de sondage complexe et de celle que l'on aurait obtenue en supposant que les données sont issues d'un plan de sondage simple.

2. Objectifs et typologie des enquêtes répétées

La conception et l'exploitation d'une enquête répétée dans le temps se heurtent aux mêmes difficultés que pour une enquête ponctuelle, mais le caractère répétitif de l'opération tend à en accroître l'ampleur. D'un point de vue pratique, la gestion temporelle rigoureuse de l'échantillon, du travail de terrain, de la collecte, du contrôle et du traitement des données est indispensable et nécessite souvent une intendance conséquente. Le respect constant des délais de production devient une difficulté intrinsèque qui ne peut être surmontée que par l'allocation de moyens adéquats. De plus, l'interrogation répétée d'un même individu peut entraîner des complications - phénomène de lassitude des répondants conduisant à une attrition accélérée de l'échantillon, apparition de biais spécifiques - ou au contraire présenter des avantages - coopération accrue, possibilité de contrôle longitudinal dès la deuxième vague -. D'un point de vue théorique, les problèmes sont également nombreux. Quelle fréquence d'interrogation retenir ? Comment tirer l'échantillon et le renouveler au fil du temps ? Quel estimateur adopter ? Le choix du type d'enquête et de la méthodologie est souvent difficile et doit faire l'objet d'un compromis en fonction des objectifs que l'on se fixe.

Duncan et Kalton (1987) distinguent sept objectifs pour les enquêtes répétées :

(a) estimer une caractéristique de la population à différentes périodes (proportion de chômeurs dans la population par exemple). Le résultat est la construction d'une série temporelle qui peut faire ensuite l'objet d'analyses spécifiques ou être utilisée pour des travaux de modélisation.

(b) estimer une caractéristique de la population en moyenne sur plusieurs périodes. On peut, par exemple, s'intéresser au nombre moyen de chômeurs sur l'année. Ce calcul peut aussi être une façon d'éliminer les phénomènes saisonniers dans le cas d'une grandeur évoluant peu sur moyenne période mais qui présenterait des fluctuations infra-annuelles marquées comme, par exemple, le nombre de décès par semaine.

(c) mesurer des évolutions à un niveau agrégé.

(d) mesurer différentes composantes d'évolution au niveau individuel :

- *des flux* comme, par exemple, les passages entre différents états d'activité économique (emploi, chômage ou inactivité) entre deux périodes.

- *des évolutions moyennes individuelles* (évolution moyenne du revenu, du temps de travail) calculées sur plusieurs périodes afin de lisser les irrégularités ponctuelles.

- *une instabilité individuelle* comme les variations du revenu mensuel d'un individu ou d'un ménage par rapport à son revenu moyen annuel.

(e) calculer des grandeurs individuelles moyennes comme, par exemple, le revenu annuel d'un ménage obtenu à partir de ses déclarations mensuelles.

(f) mesurer l'occurrence, la fréquence et la durée de certains phénomènes (nombre de personnes qui ont connu une période de chômage au cours de l'année).

(g) accumuler des informations sur des populations rares en regroupant les échantillons sur plusieurs périodes. On pourra s'intéresser, par exemple, au devenir des personnes qui ont suivi un contrat d'apprentissage ou à la situation des personnes d'origine étrangère.

Duncan et Kalton définissent également quatre types d'enquêtes dont ils évaluent les mérites et les inconvénients au regard de ces sept objectifs : (1) les enquêtes répétées sur échantillons distincts, (2) les enquêtes par panel, (3) les enquêtes à échantillon partagé et (4) les enquêtes à échantillon rotatif.

2.1 Les enquêtes répétées sur échantillons distincts

Cette technique consiste simplement à réaliser une série d'enquêtes en coupe instantanée où aucun individu n'est volontairement interrogé plus d'une fois.

Tout d'abord, il faut insister sur le fait que ce principe ne garantit nullement l'indépendance des échantillons ni, par conséquent, celle des estimations. En effet, dans un plan de sondage à plusieurs degrés, on peut par exemple choisir de tirer des unités secondaires différentes à chaque date mais appartenant aux mêmes unités primaires. Ainsi, bien que tous les individus ne soient interrogés qu'une seule fois, les estimateurs calculés sur deux périodes distinctes présenteront une covariance d'échantillonnage induite par la contiguïté des unités secondaires. C'est le cas de la plupart des enquêtes « Ménages » de l'INSEE dont l'échantillon est tiré dans l'Echantillon Maître qui peut s'interpréter comme une première phase de tirage. En général, ces covariances seront faibles et seront négligées en pratique.

Par nature, ce protocole d'enquête permet d'éviter tous les problèmes liés à la ré-interrogation volontaire des mêmes individus : attrition au fil des vagues diminuant la précision, introduction de biais spécifiques (voir section sur les panels ci-dessous). Il permet aussi d'estimer correctement le niveau sur une période (a) ou en moyenne sur plusieurs périodes (b) et l'évolution (c) d'une caractéristique de la population ainsi que d'accumuler des informations sur des populations rares (g). Naturellement, le suivi des comportements individuels est impossible et les objectifs (d) et (e) dont, en particulier, l'estimation de flux ne peuvent être satisfaits. Bien qu'en principe

réalisable, le calcul de fréquences ou de la durée de certains phénomènes (f) qui doit faire appel à la mémoire des enquêtés est délicate en l'absence de tout contrôle longitudinal. Des interrogations multiples peuvent en effet aider à déceler d'éventuelles incohérences dans les déclarations successives des interrogés et améliorer ainsi la qualité des estimations. Le calcul d'une durée moyenne de chômage se heurterait sans aucun doute à cette difficulté. L'enquête « Transition sur le marché du travail » (Detour et alii (1998)) réalisée en septembre 96 avait en effet montré que l'instabilité des situations vis à vis du marché du travail chez certaines populations engendrait des réponses peu précises et que les calendriers reconstitués à cette occasion présentaient de nombreuses incohérences avec les déclarations faites à l'Enquête Trimestrielle Emploi un mois auparavant.

2.2 Les enquêtes par panel

Il s'agit ici d'interroger régulièrement les mêmes individus selon une fréquence et sur une durée qui peuvent varier considérablement d'un dispositif à l'autre.

Par exemple, le panel européen des ménages est une enquête annuelle réalisée sous l'égide d'Eurostat dans tous les pays de l'Union Européenne depuis 1994 et qui devrait se poursuivre jusqu'en 2002. En France, entre 7 000 et 8 000 individus répondent à cette enquête dont les thèmes principaux sont la situation des individus vis à vis du marché du travail et les revenus perçus.

L'échantillon d'un panel peut être mis à jour par intégration régulière de nouvelles unités afin de suivre fidèlement l'évolution de la population. Dans le cas contraire, les évolutions mesurées à partir des données de panel (c) sont des **évolutions nettes**, c'est à dire calculées à structure de la population constante.

Une **cohorte** est un cas particulier de panel où l'on s'intéresse à une population spécifique qui comprend l'ensemble des individus ayant vécu simultanément un même événement : personnes d'une même génération, promotion d'étudiants de l'Ensae, ensemble des couples mariés la même année etc. Le principal inconvénient est bien entendu que les résultats ne s'appliquent qu'à un type de population déterminé et ne peuvent être inférés à l'ensemble de la population. De plus, l'âge étant une caractéristique commune des individus, les effets de génération liés au cycle de vie ne peuvent être isolés des évolutions temporelles propres.

L'attrait des enquêtes par panel réside essentiellement dans les possibilités de suivi des trajectoires individuelles (d) ou d'agrégation temporelle de données individuelles (e). A ceci, ajoutons que ce procédé permet de recueillir un plus grand nombre d'informations sans augmenter obligatoirement le coût financier ou la charge pour l'enquêté. En effet, certaines variables structurelles (état civil, diplômes, formation) n'ont pas besoin d'être mesurées à chaque fois. Par ailleurs, la possibilité d'effectuer un contrôle longitudinal améliore en principe la réalisation de l'objectif (f). Enfin,

l'estimation d'évolutions (c) est a priori meilleure (en terme de précision) que dans les enquêtes à échantillons distincts puisqu'on bénéficie d'une corrélation, en général positive, entre les grandeurs observées pour un même individu (voir section 5).

En revanche, un panel ne peut servir à cumuler des échantillons relatifs à une population définie par une caractéristique statique (personnes d'origine étrangère par exemple). L'attrition et l'incorporation des nouveaux entrants peuvent aussi poser des problèmes pour les estimations en coupe instantanée (a) ou en moyenne sur longue période (b). Ansieau (1995) qualifie l'attrition de « fléau » pour les enquêtes par panel et donne quelques pistes pour en limiter les effets au moment de la collecte. Il insiste notamment sur le maintien du même enquêteur pour un individu donné, sur une relance personnalisée des ménages qui refusent de coopérer, sur la nécessité d'impliquer les ménages et d'instaurer un climat de confiance (par la présentation des résultats ou la distribution de cadeaux) ainsi que sur le suivi rigoureux des mouvements de personnes à la suite d'un déménagement ou d'un « éclatement » du ménage. Ce dernier point, qui doit naturellement être géré « en continu » pendant les opérations de collecte, nécessite une organisation adaptée pouvant se révéler très coûteuse.

Rappelons enfin que les panels sont sujets à des biais spécifiques. Un **biais de sélection** est lié au fait que les individus qui acceptent de participer durant une longue période au panel ne sont pas nécessairement « représentatifs » de la population. Une longue participation à l'enquête peut aussi pousser les individus à modifier leur comportement introduisant un **biais de conditionnement**. Dans une enquête de type Budget de Famille, l'obligation de tenir un compte détaillé des achats peut inciter les enquêtés à contrôler ou maîtriser plus leurs dépenses voire à modifier leur comportement de consommation. Enfin, le comportement de réponse des enquêtés peut évoluer au fil des vagues et donner des estimations d'espérances différentes selon le rang de l'enquête. Ce phénomène, appelé biais de renouvellement, est aussi présent dans les enquêtes à échantillon rotatif.

2.3 Les enquêtes à échantillon partagé

Cette technique, proposée par Kish, consiste à combiner un échantillon permanent de type panel et un échantillon répété, c'est à dire renouvelé entièrement à chaque période.

Kish (1998) suggère de calibrer la taille du panel à environ 1/3 de celle des échantillons répétés et décrit les avantages de ce système. La composante panel permet de satisfaire correctement les objectifs (d) et (e). Elle implique en outre des corrélations entre toutes les périodes et pas seulement entre celles spécifiées arbitrairement dans le plan de rotation ; ces corrélations peuvent alors être mises à profit pour améliorer l'estimation d'évolutions (c) entre deux périodes quelconques. L'échantillon répété permet, quant à lui, de pallier les insuffisances d'un panel pour

l'estimation des niveaux instantané (a) et moyen (b). Il convient aussi pour le cumul d'informations sur des sous-populations particulières (g).

2.4 Les enquêtes à échantillon rotatif

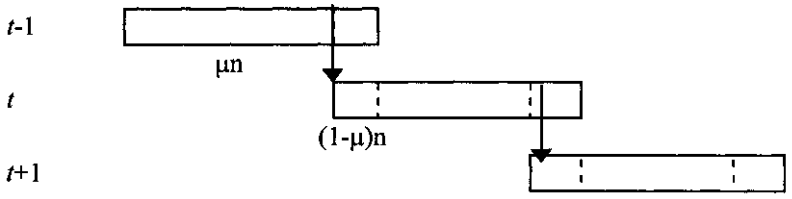
Dans ce type d'enquête, les unités statistiques ont une durée de vie limitée dans l'échantillon. A chaque période, des unités sortent de l'échantillon et sont remplacées par d'autres selon un **plan de rotation** prédéfini. On distingue habituellement trois schémas de renouvellement de l'échantillon : rotation à un niveau, demi-rotation à un niveau et rotation multi-niveaux.

Rotation à un niveau

La taille de l'échantillon n étant fixée indépendante du temps, une fraction $(1-\mu)$ des unités interrogées à la période t est conservée dans l'échantillon $t+1$ tandis que les μn autres unités sont remplacées par de nouvelles unités tirées parmi la population qui n'a pas encore été interrogée.

Dans le cas d'un fort renouvellement de l'échantillon ($\mu \geq 0,5$), seule une fraction des μn unités remplacées en t seront ré-interrogées en $t+1$, situation correspondant à la figure ci-dessous. Il faut alors remarquer que l'échantillon à la période t est constitué de trois catégories d'individus : les personnes interrogées en $t-1$ et t, les personnes interrogées en t uniquement et celles interrogées en t et $t+1$.

Plan de rotation dans le cas d'un renouvellement élevé ($\mu \geq 0,5$)



Pour un faible taux de renouvellement ($\mu \leq 0,5$) en général choisi tel que $m=1/\mu$ soit entier, les unités nouvellement introduites en t vont être interrogées m fois (i.e. pendant m mois ou trimestres selon la périodicité de l'enquête) avant de sortir de l'échantillon à la période $t+m-1$. L'échantillon est donc en permanence constitué de m **groupes** ou **sous-échantillons** de taille identique n/m, un nouveau groupe

remplaçant un sortant à chaque période. Habituellement, on repère un groupe par sa date d'entrée dans l'échantillon. De façon équivalente, on peut considérer l'échantillon comme la réunion de m **vagues** où la vague ℓ est l'ensemble des unités interrogées pour la $\ell^{\text{ième}}$ fois, c'est à dire correspond au groupe introduit à la période $t-\ell+1$.

Gurney et Daly (1965) ont introduit le concept d'**estimateur élémentaire** pour une caractéristique $\overline{Y}$ de la population. Un tel estimateur, noté $\overline{y}_{t,\ell}$, n'utilise que les données relatives à la période courante t et mesurées sur des unités appartenant à la même vague ℓ ou au groupe introduit en $t-\ell+1$. Par conséquent, les estimateurs $\overline{y}_{t,\ell}$ et $\overline{y}_{t-j,\ell-j}$ qui sont calculés sur les mêmes unités interrogées aux dates t et t-j vont être corrélés. Ces estimateurs élémentaires sont parfois affectés d'un biais appelé **biais de renouvellement** (voir Bailar (1975)) propre à chaque vague ℓ. L'existence de tels biais s'explique par un comportement de réponse différent selon le nombre de fois où les répondants ont été confrontés à l'enquête. Ce phénomène peut s'assimiler formellement à une erreur de mesure. On a par exemple remarqué que la première vague de l'enquête Emploi annuelle donnait un nombre de chômeurs en général plus élevé que les vagues deux et trois.

L'estimateur naturel $\overline{y}_t$, moyenne des estimateurs élémentaires, ne sera alors sans biais que si la somme des biais élémentaires est nulle. En revanche, $\overline{y}_t - \overline{y}_{t-1}$ estimera toujours sans biais l'évolution $\Delta\theta_t$ puisque les biais élémentaires sont supposés indépendants de t. Pour les estimateurs complexes, l'absence de biais sera liée au respect d'une condition qui dépend de la forme de l'estimateur (voir Hidiroglou et Binder (1988)). Dans le cas général, ces estimateurs sont biaisés et les comparaisons doivent s'effectuer d'après l'écart quadratique moyen.

L'enquête Emploi Canadienne, réalisée par Statistics Canada (voir Kumar et Lee (1983)), suit un plan de rotation à un niveau où les ménages sont interrogés pendant six mois consécutifs. Chaque échantillon mensuel est donc constitué de six vagues comme le montre le schéma ci-dessous où les 'X' identifient les sous-échantillons interrogés un mois donné. Le **taux de recouvrement** entre deux mois consécutif est donc de 5/6 et favorise l'estimation d'évolutions mensuelles. En revanche, les échantillons espacés d'un an sont totalement disjoints.

Plan de rotation dans l'enquête Emploi Canadienne

	Jan	Fév	Mar	Avr	Mai	Juin	Juil	Août	Sept	Oct	Nov	Déc
Sous échantillon												
1	x	x	x	x	x	x						
2		x	x	x	x	x	x					
3			x	x	x	x	x	x				
4				x	x	x	x	x	x			
5					x	x	x	x	x	x		
6						x	x	x	x	x	x	
7							x	x	x	x	x	x

L'échantillon de l'enquête Emploi annuelle française suit aussi un plan de rotation à un niveau puisqu'il est renouvelé par tiers tous les ans. L'estimateur retenu est simplement la moyenne des trois estimateurs élémentaires calculés sur chaque vague.

Les enquêtes Emploi en continu britannique et française, que l'on peut considérer comme des enquêtes trimestrielles, ont aussi adopté ce plan de rotation où les logements sont interrogés respectivement cinq et six fois consécutives. Dans le cas français, on aura des taux de recouvrement trimestriel et annuel de respectivement 83% et 33%.

Comme pour les enquêtes sur échantillons disjoints, il faut insister sur le fait que deux sous-échantillons ne sont pas nécessairement indépendants. En effet, les unités finales peuvent être issues des mêmes unités primaires dans le cas d'un plan de sondage à plusieurs degrés ou bien d'aires géographiquement proches comme dans l'enquête Emploi. De façon plus générale, Lee (1990) identifie trois types de corrélation pour les estimateurs élémentaires, constituant trois sources de **covariance d'échantillonnage**.

1) Les estimations d'une même caractéristique $\overline{Y}$, espacées de *j* périodes et calculées sur le même sous-échantillon, sont naturellement corrélées. Le coefficient de corrélation ρ_j est en général supposé indépendant du temps et du sous-échantillon considéré mais, en revanche, fonction de *j* qui ne peut dépasser *m*-1 : $\mathrm{Cov}(\overline{y}_{t,\ell}, \overline{y}_{t-j,\ell-j}) = \rho_j \sigma_y^2$ avec $1 \le j \le m-1$ et $1 \le \ell \le m$

2) Selon la méthode retenue pour tirer les unités de sondage, les estimations relatives aux périodes *t-j* et *t* et calculées respectivement sur un sous-

échantillon et son remplaçant peuvent aussi être corrélées. Formellement, on définit le coefficient γ_j par $\mathrm{Cov}(\bar{y}_{t,\ell}, \bar{y}_{t-j,\ell-j+m}) = \gamma_j \sigma_y^2$ avec $1 \le \ell \le j$ si $1 \le j \le m$ et $j-m \le \ell \le m$ si $m+1 \le j \le 2m-1$

Bien entendu, cette définition peut être étendue au cas d'un échantillon et de son $p^{\text{ième}}$ remplaçant.

3) Lee définit aussi la corrélation entre les estimations de deux caractéristiques différentes $\overline{X}$ et $\overline{Y}$ à partir du même sous-échantillon interrogé à *j* périodes d'intervalle : $\mathrm{Cov}(\bar{y}_{t,\ell}, \bar{x}_{t-j,\ell-j}) = \tau_j \sigma_y \sigma_x$ avec $1 \le j \le m-1$ et $1 \le \ell \le m$

En principe, les coefficients ρ_j et γ_j vont diminuer avec *j* et il est raisonnable de supposer ρ_j plus grand que γ_j qui pourra être négligé la plupart du temps.

Demi-rotation à un niveau

Ce schéma de rotation qui généralise le précédent a été présenté par Rao et Graham (1964) dans un cadre assez général. Chaque groupe d'unités est interrogé *p* fois consécutives, quitte l'échantillon pendant *q* périodes, puis est de nouveau interrogé *p* fois et ainsi de suite. Un groupe sortira définitivement de l'échantillon après un nombre prédéterminé de cycles (*p*, *q*).

Le *Bureau of the Census* a adopté un tel schéma de rotation pour la *Current Population Survey*. L'échantillon mensuel se compose de huit groupes qui suivent deux cycles de type (4,8). Ce plan de rotation, plus connu sous la forme 4*in*-8*out*-4*in*, consiste donc à interroger les mêmes unités pendant quatre mois consécutifs, les « désactiver » pendant huit mois et les ré-interroger pendant quatre mois. Ainsi, chaque mois, deux groupes rentrent dans l'échantillon : l'un est entièrement nouveau et l'autre correspond à celui abandonné huit mois plus tôt. L'intérêt de ce système est d'obtenir un recouvrement élevé entre deux mois consécutifs (75%) ainsi qu'entre deux mois espacés d'une année (50%), favorable à l'estimation de glissements mensuel et annuel.

Rotation multi-niveaux

Dans les deux schémas présentés plus haut, seule l'information relative à la période courante est collectée. On parlera d'échantillons multi-niveaux lorsque le questionnaire prévoit également d'enregistrer des données sur les périodes précédentes. Le cas le plus fréquent est celui d'un échantillon à deux niveaux : à chaque période *t*, on tire un nouvel échantillon complet et on recueille, pour chaque

individu, les données relatives aux périodes t et t-1. La généralisation à des niveaux supérieurs est alors immédiate. Il faut quand même noter que la variable y_{t-1} collectée en t ne peut être directement assimilée à y_{t-1} collectée en t-1. En effet, solliciter la mémoire de l'enquêté risque d'introduire une erreur de mesure supplémentaire sur y_{t-1} collectée en t qui doit être prise en compte et dont les conséquences en termes de biais et de variance doivent être étudiées.

La présence limitée d'un même individu dans l'échantillon atténue les inconvénients inhérents aux panels (biais de conditionnement, biais de sélection, attrition) et le renouvellement permanent de l'échantillon permet de conserver une bonne représentativité. La technique des échantillons rotatifs apparaît alors un bon compromis pour l'estimation de niveaux instantanés (a) ou moyens (b) et d'évolutions (c). Par rapport aux enquêtes sur échantillons distincts, l'estimation d'évolutions sera plus précise grâce au recouvrement partiel des échantillons et à l'existence d'une corrélation en général positive des grandeurs mesurées sur un même individu. De plus, certaines techniques d'estimation, en particulier les estimateurs composites, permettent de profiter de ce recouvrement pour améliorer aussi l'estimation des niveaux. Le choix du plan de rotation et de la technique d'estimation peut alors être optimisé en fonction de l'objectif que l'on privilégie.

Les possibilités de suivi des trajectoires (d) ou d'estimation de grandeurs moyennes individuelles (e) sont assez limitées et évidemment moindres que pour les panels. En général, ces objectifs ne sont pas considérés comme primordiaux dans les enquêtes à échantillon rotatif et, bien souvent, on se dispense de suivre les individus qui déménagent entre deux vagues. De ce fait, le taux de recouvrement effectif est en général inférieur au taux théorique inscrit dans le plan de rotation. Le calcul de fréquence d'événements (f) peut s'effectuer, comme pour les panels, dans de bonnes conditions. Le renouvellement régulier de l'échantillon permet, par ailleurs, le cumul d'observations (g) sur population rare définie par des caractéristiques statiques.

Enfin, des considérations de coûts peuvent aussi plaider en faveur des échantillons rotatifs. La première interview se déroulant en général en face à face, les ré-interrogations seront réalisées par téléphone à un coût a priori moindre. Par ailleurs, une prise de contact plus rapide lors des ré-interrogations, aidée notamment par la possibilité de prendre rendez-vous, facilite le travail de l'enquêteur.

3. Principes de l'estimation dans les enquêtes à échantillon rotatif

Un objectif naturel des enquêtes répétées est d'estimer la séquence $\{\theta_t\}$ d'une certaine caractéristique de la population à partir d'observations individuelles qui sont en général corrélées positivement d'une période à l'autre. La probabilité d'être au chômage à une période donnée est, selon toute vraisemblance, liée à la ou aux situations précédentes. Connaissant les observations passées, la prise en compte de ce lien doit alors permettre d'améliorer l'estimation de la période courante. Le caractère répétitif de l'enquête associé au recouvrement partiel des échantillons ouvre ainsi la voie à un ensemble de techniques qui, par nature, ne peuvent être appliquées aux enquêtes ponctuelles.

Le principe étant établi, il s'agit maintenant de préciser ces techniques d'estimation et les conséquences sur la politique de renouvellement à suivre. Comme nous l'avons signalé plus haut, les enquêtes à échantillon rotatif se concentrent principalement sur trois objectifs :

a) estimer le paramètre θ pour chaque période

b) estimer θ en moyenne sur plusieurs périodes

c) estimer l'évolution du paramètre θ entre deux périodes (successives ou non)

En termes de renouvellement de l'échantillon, ces objectifs sont malheureusement contradictoires et **le choix du plan de rotation doit faire l'objet d'un compromis.** A ce stade, on peut énoncer les principes suivants. Pour effectuer un cumul (ou calculer une moyenne) sur plusieurs périodes, on a intérêt à renouveler entièrement l'échantillon à chaque fois afin de maximiser le nombre d'observations indépendantes. En revanche, pour estimer une évolution, il est souhaitable de conserver intégralement l'échantillon alors qu'un taux de recouvrement modeste sera préférable pour l'estimation de niveaux.

Par ailleurs, il faut noter que, pour un plan de rotation donné, les gains en précision et la forme de l'estimateur optimal vont dépendre de la corrélation temporelle de la variable d'intérêt. Lorsque le coefficient de corrélation entre deux périodes ρ est faible ($\rho \leq 0{,}6$), les gains en précision sont en général limités et ne justifient pas toujours l'application de procédures complexes. Pour des valeurs élevées ($\rho \geq 0{,}8$), on obtiendra des gains substantiels sur l'estimation des évolutions et un peu plus modestes sur les niveaux.

Sauf exception rare, les enquêtes portent sur plusieurs variables d'intérêt qui n'ont naturellement pas les mêmes structures de corrélation. Ainsi, la rémanence du chômage est vraisemblablement plus faible que celle de l'emploi. Idéalement, il

faudrait donc utiliser un estimateur adapté pour chaque variable ; cette stratégie est difficilement envisageable car, au-delà d'une complexité accrue et d'un long travail de mise au point, elle pose de redoutables problèmes de cohérence (sur le bouclage de la population active par exemple). Là encore, un compromis est nécessaire. Cette remarque vaut aussi pour le cas de résultats désagrégés (par âge et par sexe) : la durée moyenne du chômage est reconnue plus faible chez les jeunes que chez les personnes de plus de 40 ans.

La question de l'estimation dans les enquêtes répétées à échantillons rotatifs peut être abordée selon deux approches : une approche dite « classique » et une approche par les séries temporelles (ou modélisée). Binder et Hidiroglou (1988) font une présentation claire et complète de ces deux approches et des techniques d'estimation associées. Nous n'en rappelons ici que les principes généraux.

Dans **l'approche classique**, la séquence $\{\theta_t\}$ des paramètres de la population est supposée inconnue et fixe, c'est à dire non aléatoire et sans aucune relation entre les valeurs successives. Seules les variables individuelles y_t^i sont supposées reliées dans le temps selon une structure de corrélation que l'on peut déterminer d'après les observations. La connaissance de cette structure permet alors de construire un estimateur associant de façon optimale l'ensemble de l'information recueillie aux périodes courante et passées.

Au contraire, **l'approche « série temporelle »** considère le paramètre θ aléatoire dont l'évolution peut être décrite par un modèle stochastique. Ayant supposé l'existence et la forme du modèle, il est alors possible d'en estimer les paramètres et d'exhiber un estimateur optimal de la valeur courante θ_t. Prenons un exemple extrême : si le paramètre θ est constant, alors la moyenne des observations successives $\sum_{t=1}^{T} \bar{y}_t / T$ sera évidemment meilleure que n'importe quelle estimation courante $\bar{y}_t$.

En définitive, cette approche reprend et systématise l'habitude qu'ont les utilisateurs de modéliser les résultats des enquêtes répétées et de les analyser par des techniques de séries temporelles comme, par exemple, pour la correction des fluctuations saisonnières.

Dans ce document, on se limite aux estimateurs « classiques » du niveau et de l'évolution d'une caractéristique de la population, en vue d'une possible application à la future enquête Emploi en continu. Contrairement aux modèles de séries temporelles dont la construction nécessite la disponibilité de données réelles suffisantes, les propriétés et performances de ces estimateurs peuvent être facilement simulées et appréciées à partir d'un ensemble minimal d'information.

4. Cadre d'étude et notations

Dans la suite de ce document, on s'intéresse à l'estimation d'une caractéristique de la population et à la mesure de son évolution dans le cas d'une enquête à échantillon rotatif à un niveau. Dans un premier temps, on fait l'hypothèse d'un tirage aléatoire simple et on se limite au cas de deux périodes. Dans la section 5, on présente en détail les estimateurs dits « classiques » ainsi que leurs propriétés, puis, dans la section 6, on compare leurs performances pour une configuration supposée proche de celle de la future enquête Emploi en continu. La section 7 généralise ces techniques à un nombre quelconque de périodes. Enfin, on étend dans la section 8 ces résultats au cas d'un plan de sondage complexe.

On suppose la population stationnaire, c'est à dire invariante d'une période à l'autre et donc composée des N mêmes unités. Dans ce cadre, les évolutions estimées représenteront toujours des **évolutions nettes**. Le cas d'une population non stationnaire a été examiné par Holt et Skinner (1989) qui décomposent une évolution en deux éléments utiles à l'interprétation : une évolution nette et une évolution liée à une déformation de la structure de la population, aux naissances et aux disparitions d'unités. Ils traitent aussi l'estimation de flux entre sous-groupes de population définis par des caractéristiques individuelles évoluant dans le temps.

On note y_t^i la variable d'intérêt pour l'individu i à la période t dont on cherche à estimer la moyenne θ_t dans la population et l'évolution $\Delta_s\theta_t$ de sa moyenne entre les périodes t-s et t:

$$\theta_t = \frac{1}{N}\sum_{i=1}^{N} y_t^i \quad \text{et } \Delta_s\theta_t = \theta_t - \theta_{t-s}$$

Dans l'enquête Emploi, la variable Y décrira par exemple l'appartenance des individus à l'une des catégories d'activité économique - emploi, chômage ou inactivité - et θ_t représentera la proportion de chaque catégorie dans la population totale (et non dans la population active qui est a priori inconnue). On note S_t la variance empirique de la variable Y dans l'ensemble de la population que l'on supposera, pour plus de simplicité, constante dans le temps :

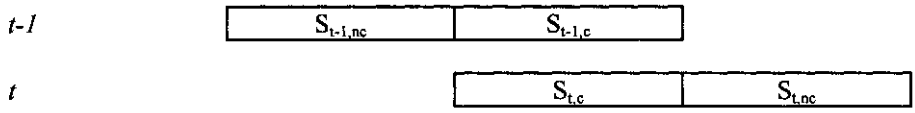
$$S_t^2 = S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_t^i - \theta_t)^2$$

La covariance entre les variables Y_t et Y_{t-s} est aussi supposée stationnaire et ne dépendre que de s :

$$Cov(Y_t, Y_{t-s}) = \rho_s S^2$$

Sauf à la section 7.1 consacrée à l'estimateur proposé par Patterson (1950) où l'on suppose une forme exponentielle pour ρ_s, on ne fera aucune hypothèse particulière sur la structure de ces corrélations.

L'échantillon, renouvelé par rotation à un niveau, est de taille *n* invariante. Un tirage aléatoire simple parmi les *N* unités de la population donne l'échantillon de première période. Aux périodes suivantes, on conserve n_c unités (par tirage aléatoire simple, *c* pour commun) de l'échantillon et les n_{nc} autres unités (*nc* pour non commun) sont remplacées par de nouvelles unités issues de la population non encore interrogée. Entre les périodes *t*-1 et *t*, les parties commune et non-commune des échantillons peuvent être représentées comme suit



Le taux de renouvellement *k* vaut par conséquent

$$k = \frac{n_{nc}}{n} = \frac{n - n_c}{n}$$

Le taux de recouvrement entre deux échantillons distants de *s* périodes sera noté 1-*k(s)*.

Un estimateur élémentaire de θ n'utilise que les données relatives à la période courante *t* et un sous-échantillon particulier. Pour un plan de sondage aléatoire simple, l'estimateur élémentaire relatif à la vague sera simplement :

$$\bar{y}_{t,\ell} = \frac{m}{n} \sum_{\ell} y_t^i \qquad \text{où n/m (entier) est la taille de chaque sous-échantillon}$$

Enfin, on peut définir $\bar{y}_{t,c}$ et $\bar{y}_{t,nc}$, estimateurs élémentaires de θ sur les parties commune et non commune de l'échantillon (par rapport à l'échantillon précédent).

Pour un schéma de rotation à un niveau,

$$\bar{y}_{t,c} = \frac{1}{m-1}\sum_{\ell=2}^{m}\bar{y}_{t,\ell} \quad \text{et} \quad \bar{y}_{t,nc} = \bar{y}_{t,1} \quad \text{où } m \text{ est le nombre de vagues}$$

Ces estimateurs élémentaires sont de plus supposés indépendants :

$$\mathrm{Cov}(\bar{y}_{t,\ell}, \bar{y}_{t,\ell'}) = 0, \text{ pour } \ell \neq \ell' \Rightarrow \mathrm{Cov}(\bar{y}_{t,c}, \bar{y}_{t,nc}) = 0$$

Avec ces notations, la moyenne empirique sur l'ensemble de l'échantillon s'écrit

$$\bar{y}_t = \frac{n_c}{n}\bar{y}_{t,c} + \frac{n_{nc}}{n}\bar{y}_{t,nc} \quad \text{soit } \bar{y}_t = (1-k)\bar{y}_{t,c} + k\bar{y}_{t,nc}$$

On suppose enfin que les estimateurs élémentaires sont sans biais. Tous les estimateurs considérés ici le seront donc aussi et leur variance pourra servir d'échelle de comparaison.

5. Enquête réalisée sur deux périodes

5.1 L'estimateur naturel

La technique la plus immédiate pour estimer la moyenne d'une variable sur deux périodes est de considérer chaque période séparément et d'y appliquer un estimateur reconnu efficace sachant le plan de sondage. Dans le cas d'un plan de sondage aléatoire simple sans remise, l'estimateur naturel du paramètre θ_t est la moyenne empirique $\bar{y}_t$ (t=1, 2). En négligeant le terme de correction pour population finie, sa variance vaut S^2 / n, indépendamment du schéma de rotation.

L'estimateur naturel de l'évolution $\Delta\theta$ est $\bar{y}_2 - \bar{y}_1$ dont la variance se calcule facilement :

$$V(\bar{y}_2 - \bar{y}_1) = V(\bar{y}_2) + V(\bar{y}_1) - 2\mathrm{Cov}(\bar{y}_2, \bar{y}_1)$$

avec

$$\mathrm{Cov}(\bar{y}_2, \bar{y}_1) = \mathrm{Cov}(k\bar{y}_{1,nc} + (1-k)\bar{y}_{1,c}, k\bar{y}_{2,nc} + (1-k)\bar{y}_{2,c}) = (1-k)^2 \mathrm{Cov}(\bar{y}_{1,c}, \bar{y}_{2,c})$$

Comme $\mathrm{Cov}(\bar{y}_{1,c}, \bar{y}_{2,c}) = \frac{S_{12}}{n_c} \approx \frac{\rho}{(1-k)} \frac{S^2}{n}$ on obtient finalement

$$\boxed{V(\bar{y}_2 - \bar{y}_1) = 2(1 - \rho(1-k)) \frac{S^2}{n}}$$

Remarques

- Dans le cas d'un renouvellement complet (k=1), les deux échantillons sont indépendants et on retrouve la formule classique $V(\bar{y}_2 - \bar{y}_1) = 2S^2 / n$.

- Le rapport des variances de l'estimateur naturel et de l'estimateur sur échantillons indépendants vaut $(1 - \rho(1-k))$. Le gain maximal est obtenu en conservant l'intégralité de l'échantillon, soit k=0, et celui-ci est d'autant plus élevé que le coefficient de corrélation ρ est grand.

5.2 Evolution estimée sur la partie commune des échantillons

Pour estimer l'évolution du paramètre θ, on peut aussi songer à n'utiliser que les unités interrogées aux deux périodes pour lesquelles on mesure directement l'évolution individuelle et de calculer, par conséquent, $\bar{y}_{2,c} - \bar{y}_{1,c}$. Les estimateurs élémentaires $\bar{y}_{1,c}$ et $\bar{y}_{2,c}$ étant supposés sans biais, leur différence l'est aussi et la variance du nouvel estimateur vaut

$$\boxed{V(\bar{y}_{2,c} - \bar{y}_{1,c}) = 2\frac{1-\rho}{1-k}\frac{S^2}{n}}$$

Remarque

- Le calcul pratique de cette variance ne nécessite pas d'estimer ρ. Il suffit en effet de construire une nouvelle variable *z* pour chaque individu, $z^i = y_2^i - y_1^i$, et d'estimer directement la variance de $\bar{z}$ par

$$\hat{V}(\bar{z}) = \frac{1}{n_c(n_c - 1)}\sum_c (z^i - \bar{z})^2$$

Comparaison de l'estimateur sur la partie commune et de l'estimateur naturel

En fonction du plan de rotation retenu, est-il préférable d'estimer l'évolution sur la partie commune des deux échantillons seulement ou bien utiliser l'ensemble des deux échantillons ?

D'après ce qui précède, le rapport des variances de ces deux estimateurs vaut

$$\frac{V(\bar{y}_{2,c} - \bar{y}_{1,c})}{V(\bar{y}_2 - \bar{y}_1)} = \frac{1-\rho}{(1-k)(1-\rho(1-k))} \text{ qui est inférieur à 1 dès que } \rho \geq \frac{1}{2-k}.$$

Ainsi, pour un taux de renouvellement *k* donné, l'estimateur sur la partie commune sera plus efficace que l'estimateur naturel si le coefficient de corrélation est suffisamment élevé. Dans un tel cas, la forte corrélation de la variable d'intérêt suffit à compenser la perte due à un échantillon de taille plus petite. En revanche, si ρ est inférieur à 1/2, il est préférable d'utiliser l'estimateur naturel, et ce, quel que soit le plan de rotation choisi.

5.3 L'estimateur optimal

Gurney et Daly (1965) ont développé un **estimateur linéaire sans biais de variance minimale** pour les enquêtes répétées en utilisant le concept d'estimateurs élémentaires. Leur approche utilisant la théorie des espaces de Hilbert peut aussi être présentée comme une application du théorème de Gauss-Markov.

L'estimateur élémentaire $\overline{y}_{t,\ell}$ estime θ_t sans biais :

$$\overline{y}_{t,\ell} = \theta_t + e_t \quad \text{où} \quad e_t \text{ est l'erreur d'échantillonnage d'espérance nulle.}$$

En empilant ces relations par vague ℓ et par période t, on obtient la forme matricielle suivante

$$\overline{Y} = X\Theta + e \text{ où } \Theta = (\theta_1,..,\theta_t)'$$

$$\overline{Y} = (\overline{y}_{1,1},..,\overline{y}_{1,m},..,\overline{y}_{t,m})'$$

X est une matrice adéquate composée de 0 et de 1.

$$E(e) = 0\,,\ E(ee') = \Omega$$

La matrice de variance-covariance Ω, de forme bloc diagonale, reflète une structure de corrélation a priori quelconque qui dépend du plan de rotation et de la variable d'intérêt. D'après le théorème de Gauss Markov, le meilleur estimateur linéaire sans biais de Θ est l'estimateur des moindres carrés généralisés $\widetilde{\Theta} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}Y$ dont la variance est $V(\widetilde{\Theta}) = (X'\Omega^{-1}X)^{-1}$. De plus, le meilleur estimateur d'une combinaison linéaire $\lambda'\Theta$ est $\lambda'\widetilde{\Theta}$ avec une variance égale à $\lambda'V(\widetilde{\Theta})\lambda$. En particulier, l'estimateur optimal de $\Delta\theta_t$ est $\widetilde{\theta}_t - \widetilde{\theta}_{t-1}$, naturellement cohérent avec l'estimateur optimal des niveaux. Cette cohérence s'obtient au prix d'une révision de l'ensemble des estimations à chaque enquête. Ceci peut toutefois poser des problèmes aux Instituts Nationaux de Statistique dont les politiques de diffusion n'autorisent, en général, qu'un nombre limité de révisions.

Gouriéroux et Roy (1978) donnent une présentation détaillée de cet estimateur et des stratégies optimales de renouvellement, **dans le cas de deux périodes et d'un sondage aléatoire simple**. En reprenant le formalisme introduit précédemment, le problème général de l'estimation de θ_1 et θ_2 s'écrit selon le modèle linéaire suivant :

$$\begin{pmatrix} \bar{y}_{1,nc} \\ \bar{y}_{1,c} \\ \bar{y}_{2,nc} \\ \bar{y}_{2,c} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} + \begin{pmatrix} e_{1,nc} \\ e_{1,c} \\ e_{2,nc} \\ e_{2,c} \end{pmatrix}$$

soit $\bar{Y} = X\Theta + e$, où $\bar{Y}$ est une statistique exhaustive sous l'hypothèse d'un modèle gaussien.

La matrice de variance-covariance des erreurs d'échantillonnage, dans le cas spécifié plus haut, s'écrit

$$E(ee') = \Omega = S^2 \begin{pmatrix} 1/n_{nc} & 0 & 0 & 0 \\ 0 & 1/n_c & 0 & \rho/n_c \\ 0 & 0 & 1/n_{nc} & 0 \\ 0 & \rho/n_c & 0 & 1/n_c \end{pmatrix}$$

L'application du théorème de Gauss-Markov donne le meilleur estimateur des niveaux θ_1 et θ_2 ainsi que de l'évolution $\theta_2 - \theta_1$. Soit,

pour les niveaux

$$\bar{y}_1^{opt} = \bar{y}_{1,nc} + \frac{1-k}{1-k^2\rho^2}(\bar{y}_{1,c} - \bar{y}_{1,nc}) + \frac{\rho k(1-k)}{1-k^2\rho^2}(\bar{y}_{2,nc} - \bar{y}_{2,c})$$

$$\bar{y}_2^{opt} = \frac{\rho k(1-k)}{1-k^2\rho^2}(\bar{y}_{1,c} - \bar{y}_{1,nc}) + \frac{1-k}{1-k^2\rho^2}(\bar{y}_{2,c} - \bar{y}_{2,nc}) + \bar{y}_{2,nc}$$

$$V(\bar{y}_1^{opt}) = V(\bar{y}_2^{opt}) = \frac{S^2}{n}\frac{1-k\rho^2}{1-k^2\rho^2}$$

pour l'évolution

$$\bar{y}_2^{opt} - \bar{y}_1^{opt} = \bar{y}_{2,nc} - \bar{y}_{1,nc} + \frac{1-k}{1-k\rho}(\bar{y}_{1,nc} - \bar{y}_{1,c} + \bar{y}_{2,c} - \bar{y}_{2,nc})$$

$$V(\bar{y}_2^{opt} - \bar{y}_1^{opt}) = 2\frac{S^2}{n}\frac{1-\rho}{1-k\rho}$$

Remarques

- Les deux périodes jouent des rôles symétriques **et l'estimation optimale de θ_1 utilise l'information recueillie à la période 2**. Il y a « rétro-estimation » de θ à la première période.

- Les estimations de niveaux et d'évolutions sont cohérentes, mais cette cohérence s'obtient au prix de révisions sur les estimations passées.

- Pour estimer les moyennes, le taux de renouvellement optimal est $k = \frac{1-\sqrt{1-\rho^2}}{\rho^2}$ pour ρ non nul et le plan de rotation est indifférent si ρ est nul.

- Pour k=0 ou 1, on retrouve l'estimateur naturel : qu'il y ait renouvellement total ou nul, l'échantillon de chaque période contient toute l'information disponible et les périodes peuvent être traitées séparément.

- Pour estimer l'évolution, il faut conserver l'échantillon si ρ est positif et le renouveler entièrement dans le cas contraire.

- Cette technique des moindres carrés peut être aussi appliquée au problème de l'estimation de flux entre différents sous-groupes de population. L'annexe 1 donne l'exemple des flux entre emploi et chômage développé par Fuller (1990).

5.4 Estimateurs composites

Afin d'éviter de trop grandes complications, notamment dans le cas de plusieurs périodes et de plans de sondage complexes, des estimateurs dits « composites » ont été développés comme alternative au meilleur estimateur linéaire sans biais. Prenant des formes plus ou moins simples et intuitives, ils suivent néanmoins un principe général qui consiste à améliorer l'estimateur instantané (i.e. qui n'utilise que l'information de la période courante) en utilisant la corrélation des observations sur la partie commune de l'échantillon.

Pour l'estimation d'évolutions, deux stratégies sont envisageables. On peut soit calculer la différence des estimations composites des niveaux, soit construire directement un estimateur composite de l'évolution. En général, la première approche n'est pas optimale - au sens de la précision - mais a l'avantage d'assurer la cohérence entre évolutions et niveaux.

Application de l'estimateur par régression sur la partie commune de l'échantillon

Jensen (1942) considéra le premier la question de l'estimation et du renouvellement de l'échantillon dans le cas de deux enquêtes successives sur une population de taille infinie. Pour la seconde période, il propose de combiner un estimateur du type estimateur par régression sur la partie commune de l'échantillon où l'information auxiliaire est la variable collectée à la première période et l'estimateur élémentaire sur la partie non commune :

$$\bar{y}_{2,c}^{reg} = \bar{y}_{2,c} + b(\bar{y}_1 - \bar{y}_{1,c})$$

où le paramètre b est le coefficient de la régression linéaire de y_2 sur y_1 :

$$b = \frac{\sum_N (y_2^i - \bar{y}_2)(y_1^i - \bar{y}_1)}{\sum_N (y_1^i - \bar{y}_1)^2}$$

On notera que cette expression correspond à l'estimation du coefficient de corrélation ρ lorsque la variance empirique de Y est constante entre les deux périodes. Ce coefficient peut être estimé directement sur la partie commune de l'échantillon.

Cette application diffère un peu du cadre strict de l'estimation par régression puisque l'on opère en deux phases. En effet, le total (resp la moyenne) de la variable auxiliaire n'est pas connu mais est estimé à l'occasion d'un premier sondage de taille n où l'on mesure cette variable auxiliaire uniquement ; on procède ensuite à un deuxième échantillonnage de taille n_c (tirage aléatoire simple) où l'on recueille la variable d'intérêt et la variable auxiliaire. Ce double échantillonnage doit être pris en compte dans les formules de variance.

Pour le calcul de variance, il n'est pas indispensable de faire l'hypothèse d'un modèle linéaire entre les variables d'intérêt y_2 et auxiliaire y_1, hypothèse d'ailleurs rarement vérifiée, et on ne dispose dans ce cas que de résultats asymptotiques (i.e. pour n et n_c grands). Lorsque b est estimé par les moindres carrés, l'estimateur régression est biaisé (biais en $1/n_c$) et sa variance vaut asymptotiquement (Cochran (1977))

$$V(\bar{y}_{2,c}^{reg}) = \frac{(1-\rho^2)S^2}{n_c} + \rho^2 \frac{S^2}{n}$$

On peut maintenant construire un estimateur composite de θ_2 en combinant l'estimateur par régression sur la partie commune de l'échantillon et l'estimateur élémentaire sur la partie non commune :

$$\bar{y}_2^* = (1-\phi)\bar{y}_{2,nc} + \phi\bar{y}_{2,c}^{reg}$$

soit $$\boxed{\bar{y}_2^* = (1-\phi)\bar{y}_{2,nc} + \phi(\bar{y}_{2,c} + b(\bar{y}_1 - \bar{y}_{1,c}))}$$

La variance de cet estimateur se calcule facilement puisque les deux termes sont indépendants :

$$V(\bar{y}_2^*) = (1-\phi)^2 V(\bar{y}_{2,nc}) + \phi^2 V(\bar{y}_{2,c}^{reg})$$

soit $$V(\bar{y}_2^*) = (1-\phi)^2 \frac{S^2}{kn} + \phi^2 \frac{S^2}{n}\left[\frac{1-\rho^2}{1-k} + \rho^2\right]$$

Le rapport de cette variance à la variance de l'estimateur naturel peut s'écrire sous la forme $a\phi^2 + b\phi + c$ où a est positif : il existe donc une combinaison optimale. On remarque toutefois que ce rapport vaut 1/k lorsque ϕ est nul et n'est donc pas toujours inférieur à 1 : il est impératif de bien choisir ϕ. La combinaison optimale en terme de précision sera obtenue en pondérant ces deux estimateurs indépendants proportionnellement à l'inverse de leur variance

$$\phi_{opt} = \frac{V(\bar{y}_{2,nc})}{V(\bar{y}_{2,c}^{reg}) + V(\bar{y}_{2,nc})} \quad \text{soit } \phi_{opt} = \frac{1-k}{1-k^2\rho^2}$$

Finalement $$\boxed{\bar{y}_2^* = (1 - \frac{1-k}{1-k^2\rho^2})\bar{y}_{2,nc} + \frac{1-k}{1-k^2\rho^2}(\bar{y}_{2,c} + b(\bar{y}_1 - \bar{y}_{1,c}))}$$

Si l'on se souvient que $b=\rho$, on peut alors vérifier sans difficulté que l'estimateur ainsi défini correspond exactement à l'estimateur optimal présenté dans la section précédente. Sa variance vaut donc

$$\boxed{V(\bar{y}_2^*) = \frac{S^2}{n}\frac{1-k\rho^2}{1-k^2\rho^2}}$$

Remarques

- Puisque k est inférieur à 1, cet estimateur composite est toujours meilleur que l'estimateur naturel : $\frac{S^2}{n}\frac{1-k\rho^2}{1-k^2\rho^2} \leq \frac{S^2}{n}$, pour $k \leq 1$

- Sa variance est minimale pour $k = (1-\sqrt{1-\rho^2})/\rho^2$. On note d'ailleurs que le taux de recouvrement optimal de l'échantillon diminue avec ρ. Lorsque ρ est proche de 1, il faut, conformément à l'intuition, renouveler entièrement l'échantillon et la variance est la moitié de celle de l'estimateur naturel. Pour ρ voisin de 1/2, le taux de rotation optimal est de l'ordre de 50%.

Estimation de l'évolution à partir de l'estimation composite du niveau

Pour la période 1, il n'est pas possible de construire un estimateur composite (sauf à considérer l'information de la période 2) et on conserve l'estimateur naturel. L'évolution $\Delta\theta$ peut s'estimer directement par la différence de ces deux estimateurs :

$$\bar{y}_2^* - \bar{y}_1 = (1-\phi_{opt})\bar{y}_{2,nc} + \phi_{opt}(\bar{y}_{2,c} + b(\bar{y}_1 - \bar{y}_{1,c})) - \bar{y}_1$$

La variance se calcule aisément

$$V(\bar{y}_2^* - \bar{y}_1) = V(\bar{y}_2^*) + V(\bar{y}_1) - 2Cov(\bar{y}_2^*, \bar{y}_1)$$

Or $Cov(\bar{y}_2^*, \bar{y}_1) = Cov\left((1-\phi_{opt})\bar{y}_{2,nc} + \phi_{opt}(\bar{y}_{2,c} + b(\bar{y}_1 - \bar{y}_{1,c})), (1-k)\bar{y}_{1,c} + k\bar{y}_{1,nc}\right)$

S supposé constant et b=ρ impliquent que

$$Cov(\bar{y}_2^*, \bar{y}_1) = \phi_{opt}(1-k)Cov(\bar{y}_{2,c}, \bar{y}_{1,c}) + \phi_{opt}\rho V(\bar{y}_1) - \phi_{opt}\rho(1-k)V(\bar{y}_{1,c})$$

Soit $Cov(\bar{y}_2^*, \bar{y}_1) = \phi_{opt}\rho\frac{S^2}{n}$

Finalement $$\boxed{V(\bar{y}_2^* - \bar{y}_1) = \frac{S^2}{n}\left[1 + \frac{1}{1-k^2\rho^2}(1-k\rho^2 - 2(1-k)\rho)\right]}$$

Estimateur composite direct de l'évolution

On a vu précédemment qu'il était possible, dans certains cas, d'obtenir une meilleure estimation de l'évolution en se restreignant à la partie commune de l'échantillon. Une idée logique consiste à combiner cet estimateur avec l'estimateur naturel calculé sur la partie non-commune de l'échantillon :

$$\Delta\bar{y}^* = \phi(\bar{y}_{2,c} - \bar{y}_{1,c}) + (1-\phi)(\bar{y}_{2,nc} - \bar{y}_{1,nc})$$

Comme pour le niveau, le meilleur estimateur (au sens de la précision) est obtenu en pondérant chaque élément proportionnellement à l'inverse de sa variance :

$$\phi_{opt} = \frac{V_{nc}}{V_{nc} + V_c} \quad \text{soit} \quad \phi_{opt} = \frac{1-k}{1-k\rho}$$

puisque $V_{nc} = V(\bar{y}_{2,nc} - \bar{y}_{1,nc}) = 2\frac{S^2}{kn}$ et $V_c = V(\bar{y}_{2,c} - \bar{y}_{1,c}) = 2\frac{S^2}{n}\frac{1-\rho}{1-k}$.

Finalement, $$\boxed{\Delta\bar{y}^* = \frac{1-k}{1-k\rho}(\bar{y}_{2,c} - \bar{y}_{1,c}) + (1 - \frac{1-k}{1-k\rho})(\bar{y}_{2,nc} - \bar{y}_{1,nc})}$$

Les deux estimateurs de l'évolution étant indépendants, la variance de l'estimateur composite est

$$V(\Delta\bar{y}^*) = \frac{V_{nc}V_c}{V_{nc} + V_c} \quad \text{soit} \quad \boxed{V(\Delta\bar{y}^*) = 2\frac{S^2}{n}\frac{1-\rho}{1-k\rho}}$$

Comparons cette variance avec celle de l'estimateur naturel $\bar{y}_2 - \bar{y}_1$. Le rapport des variances vaut

$$\frac{V(\Delta\bar{y}^*)}{V(\bar{y}_2 - \bar{y}_1)} = \frac{1-\rho}{1-\rho(1-k\rho(1-k))} \leq 1$$

L'estimateur composite est meilleur que l'estimateur naturel ainsi que l'estimateur construit sur la seule partie commune $\bar{y}_{2,c} - \bar{y}_{1,c}$ puisque le rapport des variances qui vaut $(1-k)/(1-k\rho)$ est toujours inférieur à 1.

Cet estimateur est rigoureusement équivalent à l'estimateur optimal $\bar{y}_2^{opt} - \bar{y}_1^{opt}$ de la section 5.3.

6. Application à l'enquête Emploi en continu

6.1. Description des données

L'échantillon de l'enquête emploi annuelle est aréolaire, c'est-à-dire composé d'aires géographiques tirées dans toute la France métropolitaine. Ces aires ont été sélectionnées par un tirage stratifié à plusieurs degrés, les strates étant définies par le croisement de la région avec la catégorie de communes. L'échantillon est renouvelé par tiers tous les ans. Ainsi, celui de l'année 2000 est constitué aux deux tiers d'aires déjà présentes en 1999. L'échantillon comporte près de 100 000 logements permettant de recueillir les réponses d'environ 150 000 individus âgés de 15 ans et plus.

L'enquête Emploi annuelle réalisée traditionnellement en mars de chaque année devrait être réalisée pour la dernière fois en 2002 et être remplacée par une enquête en continu, c'est-à-dire un dispositif selon lequel les interviews seront menés tout au long de l'année. Le schéma de rotation adopté pour cette future enquête Emploi consiste à renouveler l'échantillon par sixième tous les trimestres. Chaque unité - i.e. chaque logement - sera interrogé pendant six trimestres consécutifs sans que les individus occupant ce logement soient nécessairement les mêmes sur toute la période. Il est prévu d'enquêter 50 000 logements par trimestre, ce qui correspondrait à environ 75 000 personnes de plus de 15 ans.

Un « dispositif léger » a été mis en place en juillet 1998 pour tester le dispositif de collecte en continu et accumuler un certain nombre de données et d'informations qui faciliteront le lancement et l'exploitation de l'enquête en vraie grandeur. L'échantillon du dispositif léger a été obtenu à partir des tiers sortants des enquêtes emploi annuelles de 1997 et 1998. Celui-ci consiste plus précisément en une sélection aléatoire simple d'aires parmi l'ensemble des aires constituant le tiers concerné. En toute rigueur, le plan de sondage d'un échantillon du dispositif léger est très complexe puisqu'il correspond à un plan de sondage en deux phases : la première phase est constituée d'un plan de sondage stratifié à plusieurs degrés (c'est à dire le plan de sondage de l'enquête emploi annuelle) et la seconde d'un plan de sondage aléatoire simple. Par la suite, nous assimilerons ce plan de sondage complexe à celui d'un sondage aléatoire simple. De plus, le schéma de renouvellement est différent de celui retenu pour l'enquête finale puisque l'échantillon est renouvelé par tiers tous les trimestres.

A ce jour, seules les données des quatre premiers trimestres de collecte (de 1998T3 à 1999T2 inclus) sont disponibles. Nous n'avons pas cherché à mettre en oeuvre les estimateurs étudiés dans ce docvument vu la faible longueur des données et la forme non définitive du plan de rotation. En revanche, ces données vont nous permettre

d'évaluer les caractéristiques des variables d'intérêt, notamment leur structure de corrélation temporelle, influant sur la forme et les performances de ces estimateurs.

6.2 Taux de recouvrement théorique et taux de recouvrement effectif

Mesuré au niveau individuel, le taux de recouvrement entre deux trimestres consécutifs sera inférieur au taux théorique de 5/6, notamment en raison des naissances (entrée dans la catégorie des plus de quinze ans), des décès, des déménagements et des différentiels de non-réponse entre les vagues.

Le tableau ci-dessous compare les taux de recouvrement théorique (*1-k(s)*) et effectif *c(s)* sur les quatre premiers trimestres de collecte du dispositif léger. On précise que le taux effectif est calculé d'après le nombre d'individus techniquement appariables seulement, c'est à dire le nombre d'individus qui occupaient le même logement aux deux vagues, et que l'on a omis les personnes qui, bien qu'ayant déménagé, auraient été interrogées les deux fois (ce phénomène n'introduit pas de covariance d'échantillonnage entre les estimations). Par ailleurs, l'effet des variations - légères - du nombre d'aires interrogées selon les vagues a été neutralisé dans le calcul des taux de recouvrement effectif. Pour ce faire, on a calculé le pourcentage d'individus appariés dans les seules aires interrogées aux deux dates, puis on a multiplié le résultat par le taux de recouvrement théorique.

Taux de recouvrement effectif et théorique dans le dispositif léger (en %)

	1-*k*(1)	*c*(1)	1-*k*(2)	*c*(2)
1998T4	66,6	62,4	-	-
1999T1	66,6	62,7	33,3	30,2
1999T2	66,6	62,7	33,3	30,5

On peut, de la même façon, calculer le taux de recouvrement entre les vagues de l'enquête annuelle. On a écarté les enquêtes de 1990 et 1999 qui n'ont pas été réalisées au mois de mars, du fait de leur concomitance avec le recensement de la population.

Taux de recouvrement effectif et théorique dans l'enquête annuelle (en %)

	1-*k*(4)	*c*(4)	1-*k*(8)	*c*(8)
Mars 1992	66,6	57,2	-	-
Mars 1993	66,6	58,4	33,3	25,6
Mars 1994	66,6	58,9	33,3	26,5
Mars 1995	66,6	57,3	33,3	25,9
Mars 1996	66,6	56,6	33,3	24,8
Mars 1997	66,6	56,4	33,3	24,3
Mars 1998	66,6	56,7	33,3	24,6

Steel (1996) introduit un facteur d'ajustement $h(s)$ de la forme $h(s) = \alpha\beta^s$ tel que $c(s) = h(s)(1 - k(s))$. Le graphique 1 en annexe montre effectivement une relation à peu près linéaire entre $\log(c(s)/(1-k(s)))$ et s avec α=96,6 et β=97,3 (estimés par les moindres carrés ordinaires). Cette relation simple permet d'anticiper le taux de recouvrement de l'enquête en vraie grandeur compte tenu d'une rotation de 1/6 tous les trimestres (voir tableau suivant). Ainsi, le taux de recouvrement trimestriel serait en pratique légèrement inférieur à 80% au lieu des 83% induits par le plan de rotation ; entre deux trimestres espacés d'un an, le taux de recouvrement serait d'environ 29% (au lieu de 33% en théorie).

Taux de recouvrement anticipé pour l'enquête en continu (en %)

s	1	2	3	4	5	6
1-k(s)	83	67	50	33	17	0
c(s)	78	61	45	29	14	0

6.3 Estimation des coefficients de corrélation ρ_s

Le dispositif léger peut aussi conduire à une première estimation du coefficient de corrélation ρ_s sur les variables emploi et chômage. Ayant défini la variable y_t^i par $y_t^i = 1$ si l'individu i est au chômage (en emploi) à la période t et $y_t^i = 0$ sinon, ce coefficient de corrélation s'estime directement à partir des individus interrogés aux deux périodes t-s et t :

$$\hat{\rho}_s = \frac{\hat{P}_{11} - \hat{P}_{1.}\hat{P}_{.1}}{\sqrt{\hat{P}_{1.}(1-\hat{P}_{1.})\hat{P}_{.1}(1-\hat{P}_{.1})}}$$

où P_{11}, $P_{.1}$ et $P_{1.}$ désignent respectivement les proportions d'individus appartenant à la catégorie considérée aux deux périodes, à la première période seulement et à la seconde période seulement. En notant π_i les probabilités d'inclusion, les estimateurs d'Horvitz-Thompson de ces proportions sont :

$$\hat{P}_{11} = \frac{\sum_c \frac{y_t^i y_{t-s}^i}{\pi_i}}{\sum_c \frac{1}{\pi_i}}, \quad \hat{P}_{1.} = \frac{\sum_c \frac{y_{t-s}^i}{\pi_i}}{\sum_c \frac{1}{\pi_i}} \text{ et } \hat{P}_{.1} = \frac{\sum_c \frac{y_t^i}{\pi_i}}{\sum_c \frac{1}{\pi_i}}$$

Là encore, il est possible de réaliser des estimations à partir du dispositif léger et de l'enquête annuelle. Les résultats sont présentés dans les deux tableaux ci-dessous.

Coefficient de corrélation estimé d'après le dispositif léger

	s	1998T4	1999T1	1999T2
Emploi	1	0,91	0,93	0,94
	2	-	0,87	0,91
Chômage	1	0,60	0,65	0,66
	2	-	0,51	0,54

Coefficient de corrélation estimé d'après l'enquête annuelle (mars)

	s	1992	1993	1994	1995	1996	1997	1998
Emploi	4	0,85	0,85	0,85	0,84	0,84	0,84	0,85
	8	-	0,80	0,79	0,79	0,79	0,80	0,79
Chômage	4	0,47	0,46	0,47	0,45	0,44	0,41	0,43
	8	-	0,31	0,33	0,33	0,37	0,36	0,35

On constate sans réelle surprise que le coefficient de corrélation est nettement plus fort pour l'emploi que le chômage, quel que soit l'intervalle de temps considéré. En général, les estimateurs complexes ne permettent des gains de précision notables que pour des corrélations relativement élevées (au-delà de 0,8). Il faut donc s'attendre à ce que ceux-ci soient relativement modestes dans le cas du chômage. Steel (1996) obtient des résultats tout à fait similaires sur les données de l'enquête Emploi britannique.

On observe, par ailleurs, une bonne stabilité des estimations pour la variable emploi, même si celles-ci sont peu nombreuses pour les retards 1 et 2. Dans le cas du chômage, le coefficient de corrélation varie nettement plus selon la période (il n'est d'ailleurs pas exclu que son évolution soit liée au cycle). En revanche, la réduction du coefficient de corrélation avec s est relativement faible et ne semble pas compatible avec l'hypothèse d'une décroissance exponentielle du type ρ_0^s. Ceci est d'ailleurs confirmé par le graphique 2 en annexe où l'on peut voir que $\log(\rho_s)$ n'est manifestement pas fonction linéaire de s à la fois pour l'emploi et le chômage.

Compte tenu du faible nombre de points, il n'est pas véritablement possible d'envisager des structures de corrélation plus complexes, comme des processus ARMA par exemple. On s'est limité à une extrapolation linéaire pour déterminer la valeur du coefficient de corrélation pour s=3 et s compris entre 5 et 7, sachant que seuls les retards jusqu'à l'ordre 5 interviendront dans les applications. On se propose donc de retenir les valeurs suivantes :

Coefficient de corrélation estimé

s	1	2	3	4	5	6	7	8
Emploi	0,92	0,89	0,87	0,85	0,83	0,82	0,81	0,79
Chômage	0,64	0,52	0,49	0,45	0,42	0,39	0,36	0,34

6.4 Estimation sur deux périodes

On analyse maintenant les performances des estimateurs développés dans la partie précédente pour des situations analogues à celles de l'enquête Emploi. On s'intéresse à l'estimation du niveau et de l'évolution - en glissements trimestriel et annuel - de l'emploi et du chômage en retenant les corrélations estimées plus haut.

Estimation du niveau

Pour l'estimation du paramètre θ à la période 2, seuls deux estimateurs ont été envisagés :

- l'estimateur naturel $\overline{y}_2$
- l'estimateur composite (par régression) :
  $\overline{y}_2^* = (1-\phi_{opt})\overline{y}_{2,nc} + \phi_{opt}(\overline{y}_{2,c} + b(\overline{y}_1 - \overline{y}_{1,c}))$

On définit l'efficacité relative d'un estimateur comme 100 fois le rapport de la variance de l'estimateur naturel et de la variance de cet estimateur :

$$\text{efficacité} = 100\frac{\text{V(naturel)}}{\text{V(estimateur)}}$$

Efficacité de l'estimateur composite pour l'estimation de niveaux

	1-*k* (en %)	**ρ**	**Efficacité**	**ϕ_{opt}**
Emploi	83	0,92	113,9	0,85
Chômage	83	0,64	106,2	0,84

Pour le taux de recouvrement choisi, la pondération optimale des estimations sur partie commune et non commune est relativement insensible à la variable considérée. En revanche, les gains en précision sont appréciables sur l'emploi, mais relativement modestes sur le chômage.

Estimation de l'évolution

Pour l'estimation de glissements trimestriels ou annuels, on considère les quatre estimateurs suivants :

- l'estimateur naturel : $\bar{y}_2 - \bar{y}_1$
- l'estimateur (I) sur partie commune de l'échantillon : $\bar{y}_{2,c} - \bar{y}_{1,c}$
- la différence des estimateurs composites (II) :
  $\bar{y}_2^* - \bar{y}_1 = (1-\phi_{opt})\bar{y}_{2,nc} + \phi_{opt}(\bar{y}_{2,c} + b(\bar{y}_1 - \bar{y}_{1,c})) - \bar{y}_1$
- l'estimateur composite de l'évolution (III) :
  $\Delta\bar{y}^* = \phi_{opt}(\bar{y}_{2,c} - \bar{y}_{1,c}) + (1-\phi_{opt})(\bar{y}_{2,nc} - \bar{y}_{1,nc})$

On définit l'efficacité de ces estimateurs comme précédemment.

Efficacité de différents estimateurs pour l'estimation d'évolution

	1-*k* (%)	ρ	Estimateur I Efficacité	Estimateur II Efficacitéi	ϕ_{opt}	Estimateur III Efficacité	ϕ_{opt}
Emploi	83	0,92	245,3	151,5	0,85	249,3	0,98
	33	0,85	158,3	154,2	0,49	206,5	0,77
Chômage	83	0,64	108,1	108,2	0,84	116,1	0,93
	33	0,45	51,1	104,9	0,36	108,1	0,47

Lorsqu'on se restreint à la partie commune de l'échantillon, les résultats sont très inégaux : ils peuvent être dramatiques si le taux de recouvrement et la corrélation sont faibles, comme dans le cas du glissement annuel du chômage. En revanche, les estimateurs composites apportent un gain systématique à condition que la pondération relative ϕ soit correctement choisie. Ce gain n'est toutefois appréciable que pour des corrélations élevées (variable emploi).

On constate logiquement que l'estimateur composite de l'évolution est uniformément meilleur que la différence des estimations composites. L'avantage n'est toutefois important que sur la variable emploi.

7. Enquête réalisée sur plusieurs périodes

Comme nous l'avons déjà évoqué dans la section 5.3, Gurney et Daly (1965) ont résolu le problème général de l'estimateur optimal (i.e. de variance minimale parmi les estimateurs linéaires sans biais) dans le cas de plusieurs périodes en utilisant le concept d'estimateurs élémentaires. Si la forme théorique de l'estimateur optimal est assez simple, son calcul explicite, qui nécessite en particulier l'inversion de la matrice de variance Ω, peut poser quelques difficultés pour un nombre important de périodes. Par ailleurs, l'ensemble des estimations doit être révisé à chaque nouvelle enquête ; ceci n'est pas toujours souhaitable ou envisageable pour les Instituts Nationaux de Statistique.

Auparavant, Patterson (1950) avait généralisé l'approche de Jensen (1942) au cas de plusieurs périodes pour un plan de rotation à un niveau et une variable d'intérêt qui suit un processus de Markov du type

$$y_t^i - \theta_t = \rho(y_{t-1}^i - \theta_{t-1}) + \eta_t^i \quad \text{où} \quad V(\eta_t^i) = (1-\rho^2)S^2$$

Sous ces hypothèses, Patterson montre que le meilleur estimateur linéaire sans biais prend une forme récursive assez simple, comparable à celle obtenue dans le cas de deux périodes :

$$\bar{y}_t^{Pat} = (1-\phi_t)\bar{y}_{t,nc} + \phi_t(\bar{y}_{t,c} + \rho(\bar{y}_{t-1}^{Pat} - \bar{y}_{t-1,c}))$$

Il s'agit toujours d'une moyenne pondérée d'un estimateur par régression sur la partie commune de l'échantillon et d'un estimateur simple sur la partie renouvelée. La simplicité de cette écriture tient uniquement à la forme particulière de la structure des corrélations ρ^s qui est supposée exponentielle entre des observations séparées par s périodes. Avec une structure de corrélation quelconque, il n'est plus possible de trouver une formulation de récurrence pour l'estimateur optimal. La section 6.1 présente en détail les propriétés de cet estimateur.

Les travaux de Patterson ont été étendus dans plusieurs directions. Eckler (1955) étudie le cas de plan de rotation à deux et trois niveaux. Rao et Graham (1964) retiennent l'hypothèse d'une population finie et incorporent un terme de correction de population finie dans les expressions de variance. Singh (1968) retrouve, quant à lui, des formes proches de l'estimateur de Patterson dans le cas d'un sondage en deux phases avec présence de corrélation entre unités secondaires d'une même unité primaire.

Des estimateurs dit « composites » ont été élaborés afin d'éviter la complexité des estimateurs optimaux. Avec une formulation qui s'inspire de celle établie dans le cas de deux périodes, ils constituent une approximation, en général tout à fait

satisfaisante, du meilleur estimateur linéaire sans biais. Pour l'estimation d'un niveau, un procédé classique consiste à retenir la forme suivante

$$\overline{y}_t^c = (1-\phi)\overline{y}_t + \phi(\overline{y}_{t-1}^c + d_{t,t-1})$$

où $d_{t,t-1}$ est un estimateur de l'évolution de θ entre t-1 et t (calculée par exemple sur la partie commune de l'échantillon). L'idée semble finalement assez naturelle puisqu'il s'agit d'une moyenne (optimale) entre l'estimateur naturel et un estimateur qui chaîne l'évolution mesurée sur la partie commune au niveau de la période précédente. Cet estimateur, dénommé estimateur composite K (la constante ϕ étant parfois notée K), a été retenu pour la *Current Population Survey* américaine (voir Rao et Graham (1964)) avec une constante égale à 1/2. La section 6.2 présente ses propriétés en détail.

Enfin, Gurney et Daly (1965) ont montré que cet estimateur pouvait être encore amélioré en distinguant les b sous-échantillons interrogés pour la première fois et en surpondérant les b estimateurs élémentaires associés par rapport aux (m-b) autres estimateurs élémentaires :

$$\overline{y}_t^c = \frac{1}{m}\left[(1-\phi+\psi)\sum_{\ell=1}^{b}\overline{y}_{t,\ell} + (1-\phi-\frac{b}{m-b}\psi)\sum_{\ell=b+1}^{m}\overline{y}_{t,\ell}\right] + \phi(\overline{y}_{t-1}^c + d_{t,t-1})$$

Les performances de cet estimateur, parfois appelé estimateur AK (les constantes ϕ et ψ étant notées A et K), ont été analysées par Kumar et Lee (1983) sur l'exemple de l'enquête Emploi canadienne.

7.1 L'estimateur de Patterson

Patterson (1950) se place dans le cas d'un plan de rotation à un niveau (avec tirage aléatoire simple à chaque étape) où le taux de rotation k et la taille de l'échantillon n sont constants. La variance empirique S^2 de la variable d'intérêt est aussi supposée stationnaire. Enfin, les corrélations à l'ordre s sont de la forme ρ^s indépendante du temps et proviennent du modèle individuel suivant :

$$y_t^i - \theta_t = \rho(y_{t-1}^i - \theta_{t-1}) + \eta_t^i$$

Sous ces hypothèses, Patterson montre que le meilleur estimateur linéaire sans biais de la caractéristique θ est de la forme

$$\overline{y}_t^{Pat} = (1-\phi_t)\overline{y}_{t,nc} + \phi_t(\overline{y}_{t,c} + \rho(\overline{y}_{t-1}^{Pat} - \overline{y}_{t-1,c}))$$

Naturellement, les poids ϕ_t et $1-\phi_t$ vont être choisis de façon à minimiser la variance de l'estimateur global. Cochran (1977) étudie en détail les propriétés de cet estimateur et montre en particulier que les poids convergent rapidement vers une valeur limite et que le taux de renouvellement optimal tend vers 1/2, et ce, indépendamment de la valeur de ρ. La variance totale de l'estimateur se stabilise elle aussi très rapidement (au bout de quelques périodes) tant que ρ est inférieur à 0,9.

Il est évidemment plus simple et plus pratique de fixer dès le départ la pondération ϕ et le taux de recouvrement de l'échantillon. Cochran (1977) indique d'ailleurs que la perte d'efficacité est très faible dès la deuxième période. Dans ce cas, l'estimateur s'écrit

$$\bar{y}_t^{Pat} = (1-\phi)\bar{y}_{t,nc} + \phi(\bar{y}_{t,c} + \rho(\bar{y}_{t-1}^{Pat} - \bar{y}_{t-1,c}))$$

On peut montrer que la variance de cet estimateur tend vers une valeur limite (voir annexe 2) :

$$V_\infty(\bar{y}^{Pat}) = \frac{S^2}{n}\frac{(1-k)\phi^2 + (1-k)(1-\phi)^2(1-\rho^2)}{k(1-k)(1-\rho^2(1-\phi)^2)}$$

Cette variance sera minimale (voir annexe 2) pour

$$\phi_{opt} = 1 - \frac{\sqrt{(1-\rho^2)(1-\rho^2(1-4k(1-k)))} - (1-\rho^2)}{2(1-k)\rho^2}$$

et vaudra (voir annexe 2)

$$\boxed{V_\infty(\bar{y}_{opt}^{Pat}) = \frac{S^2}{n}\frac{1-\phi_{opt}}{k}}$$

L'évolution est estimée directement par $\bar{y}_t^{Pat} - \bar{y}_{t-1}^{Pat}$ dont la variance se calcule de la façon suivante

$$V(\bar{y}_t^{Pat} - \bar{y}_{t-1}^{Pat}) = V(\bar{y}_t^{Pat}) + V(\bar{y}_{t-1}^{Pat}) - 2Cov(\bar{y}_t^{Pat}, \bar{y}_{t-1}^{Pat})$$

Or $$Cov(\bar{y}_t^{Pat}, \bar{y}_{t-1}^{Pat}) = Cov((1-\phi_{opt})\bar{y}_{t,nc} + \phi_{opt}(\bar{y}_{t,c} + b(\bar{y}_{t-1}^{Pat} - \bar{y}_{t-1,c})), \bar{y}_{t-1}^{Pat})$$

$$Cov(\bar{y}_t^{Pat}, \bar{y}_{t-1}^{Pat}) = \rho\phi_{opt}V(\bar{y}_{t-1}^{Pat}) + \phi_{opt}Cov(\bar{y}_{t,c} - b\bar{y}_{t-1,c}, \bar{y}_{t-1}^{Pat})$$

Le modèle de corrélation individuelle $y_t^i - \theta_t = \rho(y_{t-1}^i - \theta_{t-1}) + \eta_t^i$ implique que sur la partie commune de l'échantillon

$$\bar{y}_{t,c} - \theta_t = \rho(\bar{y}_{t-1,c} - \theta_{t-1}) + \bar{\eta}_{t,c}$$

d'où $\quad \mathrm{Cov}(\bar{y}_{t,c} - b\bar{y}_{t-1,c}, \bar{y}_{t-1}^{Pat}) = 0$

soit $\quad \mathrm{Cov}(\bar{y}_t^{Pat}, \bar{y}_{t-1}^{Pat}) = \rho\phi_{opt} V(\bar{y}_{t-1}^{Pat})$

Finalement,

$$\boxed{V_\infty(\bar{y}_t^{Pat} - \bar{y}_{t-1}^{Pat}) = 2\frac{S^2}{kn}(1-\rho\phi_{opt})(1-\phi_{opt})}$$

Remarques

- Pour l'évolution, il est aussi possible d'utiliser l'estimateur composite direct présenté à la section 5.4. Bien entendu, les estimations auront l'inconvénient de ne pas être cohérentes avec les niveaux $\bar{y}_t^{Pat}$.

- L'estimateur proposé par Patterson dans un cadre particulier, peut être aussi considéré comme un estimateur composite, moyenne de l'estimateur par régression et de l'estimateur naturel sur la partie non commune :

$$\bar{y}_t^{c'} = (1-\phi_t)\left[\bar{y}_{t,c} + \rho(\bar{y}_{t-1}^{c'} - \bar{y}_{t-1,c})\right] + \phi_t \bar{y}_{t,nc}$$

- Cochran (1977) montre que le coefficient de la régression ρ peut être remplacé par 1 sans trop de perte de précision dans le cas où sa vraie valeur est supérieure à 0,8. L'estimateur devient dans ce cas

$$\bar{y}_t = (1-\phi)\bar{y}_{t,nc} + \phi(\bar{y}_{t-1} + \bar{y}_{t,c} - \bar{y}_{t-1,c})$$

Cette forme est proche de celle de l'estimateur K, la seule différence étant que $\bar{y}_{t,nc}$ remplace $\bar{y}_t$.

7.2 L'estimateur K

L'estimateur K est une forme simplifiée de l'estimateur AK proposé par Gurney et Daly (1965) :

$$\bar{y}_t^K = (1-\phi)\bar{y}_t + \phi(\bar{y}_{t-1}^K + d_{t,t-1})$$

où $d_{t,t-1}$ est l'estimateur de l'évolution $\theta_t - \theta_{t-1}$ sur la partie commune de l'échantillon

$$d_{t,t-1} = \bar{y}_{t,c} - \bar{y}_{t-1,c}$$

Pour l'étude des propriétés de cet estimateur, il est plus simple d'introduire explicitement le plan de rotation et de faire apparaître les estimateurs élémentaires. Pour un schéma de rotation à un niveau comme celui de la future enquête Emploi en continu, l'estimateur K est défini par :

$$\bar{y}_t^K = (1-\phi)\bar{y}_t + \phi(\bar{y}_{t-1}^K + d_{t,t-1}) \quad \text{avec} \quad \bar{y}_t = \frac{1}{m}\sum_{\ell=1}^{m} \bar{y}_{t,\ell}$$

$$\text{et } d_{t,t-1} = \frac{1}{m-1}\sum_{\ell=2}^{m} (\bar{y}_{t,\ell} - \bar{y}_{t-1,\ell-1})$$

Pour simplifier les calculs, on se limite maintenant au cas particulier m=6, la transposition des résultats à d'autres valeurs de *m* étant immédiate. Les hypothèses retenues pour l'étude de cet estimateur sont les suivantes :

- Les sous-échantillons sont indépendants, par conséquent $\mathrm{Cov}(\bar{y}_{t,\ell}, \bar{y}_{t,\ell'}) = 0, \forall t, \ell \neq \ell'$.

- La variance des estimateurs élémentaires et celle de l'estimateur naturel sont stationnaires :
  $$V(\bar{y}_t) = \frac{S^2}{n} \quad \text{et} \quad V(\bar{y}_{t,\ell}) = 6\frac{S^2}{n}$$

- Les corrélations temporelles ρ_s sur un même sous-échantillon sont aussi stationnaires (i.e. ne dépendent que de *s, s=1,..,5*).

- Les corrélations γ entre un sous-échantillon et son prédécesseur sont nulles.

- Il n'y a pas de biais de renouvellement : $E(\bar{y}_{t,\ell}) = \theta_t, \forall \ell$. Par conséquent, l'estimateur naturel et l'estimateur K sont eux-mêmes sans biais.

Kumar et Lee (1983) excluent ces deux dernières hypothèses dans l'analyse des propriétés de l'estimateur AK appliqué à l'enquête Emploi Canadienne. La prise en compte des corrélations γ complique un peu les calculs sans toutefois en modifier la logique. En revanche, l'existence de biais de renouvellement introduit un biais pour

l'estimateur AK qui est fonction des poids ϕ et ψ. L'optimisation de ces poids peut alors se faire relativement à la variance ou à l'écart quadratique.

Les calculs de la variance de $\bar{y}_t^K$, de l'évolution $\bar{y}_t^K - \bar{y}_{t-1}^K$ et du glissement annuel $\bar{y}_t^K - \bar{y}_{t-4}^K$ (pour une enquête trimestrielle) figurent dans les annexes 3 et 4. Ces variances s'expriment en fonction de la variance de l'estimateur élémentaire et de la suite des ρ_s. On notera enfin que la variance d'échantillonnage du niveau n'est pas stationnaire si $\phi = 1$, c'est à dire lorsque l'estimateur K consiste à chaîner les évolutions pour estimer le niveau courant.

7.3 L'estimateur K pour l'enquête Emploi en continu

Les variances du niveau, de l'évolution trimestrielle et du glissement annuel ont été calculées, en faisant varier ϕ entre 0 et 1, pour les variables Emploi et Chômage en utilisant les corrélations estimées à la section 6 (les résultats exprimés en unité de S^2/n sont présentés sur les graphiques 4 et 5 en annexe). Pour chaque situation, la variance de l'estimateur est une fonction convexe de ϕ ; il est donc possible de déterminer une valeur optimale de ϕ qui minimise cette variance. La valeur optimale apparaît plus faible pour l'estimation des niveaux que pour les estimations d'évolution et le choix d'autant plus critique que les corrélations temporelles sont élevées (les optima sont relativement « plats » pour la variable chômage).

Le tableau ci-dessous présente l'efficacité de l'estimateur K optimal pour les trois grandeurs (niveau, évolution trimestrielle et glissement annuel) et les deux variables d'intérêt. Pour chaque catégorie, la valeur optimale de ϕ a été déterminée par balayage. On rappelle que l'estimateur naturel, qui sert de référence à la comparaison, est :

- pour le niveau: $\bar{y}_t$ avec $V(\bar{y}_t) = \dfrac{S^2}{n}$
- pour l'évolution trimestrielle: $\bar{y}_t - \bar{y}_{t-1}$ avec $V(\bar{y}_t - \bar{y}_{t-1}) = 2(1-\rho_1(1-k(1)))\dfrac{S^2}{n}$
- pour le glissement annuel : $\bar{y}_t - \bar{y}_{t-4}$ avec $V(\bar{y}_t - \bar{y}_{t-4}) = 2(1-\rho_4(1-k(4)))\dfrac{S^2}{n}$

Efficacité de l'estimateur K optimal

		Estimateur composite K	
		ϕ optimal	Efficacité
Emploi	Niveau	0,85	164,0
	Evolution trimestrielle	0,96	248,2
	Glissement annuel	0,96	334,2
Chômage	Niveau	0,49	106,8
	Evolution trimestrielle	0,61	114,2
	Glissement annuel	0,70	115,8

Remarques

- Les gains, très importants dans le cas de l'emploi, sont liés au niveau élevé de l'ensemble des corrélations. Ces résultats sont relativement sensibles à la valeur des ρ_s. Le tableau 1 en annexe donne l'efficacité de l'estimateur composite K pour différents profils de corrélation que l'on a choisi de la forme (0,91-*u*; 0,88-*u*; 0,86-*u*; 0,84-*u*; 0,82-*u*) où *u* varie de 0 à 0,1 en étant incrémenté de 0,01. Les gains diminuent rapidement avec le niveau des corrélations ; ceux-ci restent néanmoins appréciables pour des ρ_s de l'ordre de 0,8.

- Pour le chômage, les gains sont beaucoup plus modestes, de l'ordre de 6% pour le niveau et de 15% pour le glissement annuel ainsi que l'évolution trimestrielle.

- Kumar et Lee (1983) ont montré sur l'exemple de l'enquête Emploi Canadienne que l'efficacité obtenue sur le niveau était surestimée du fait de la non prise en compte des corrélations γ entre sous-échantillons successifs. L'ampleur de la surestimation dépend de la valeur des γ et peut être importante pour certaines sous-populations comme, par exemple, les agriculteurs. En effet, ceux-ci se trouvant majoritairement en zone rurale, le fait de remplacer une aire par une aire voisine introduit une corrélation forte entre les échantillons qui augmente la variance de l'estimateur composite (comme un recouvrement trop fort peut nuire à la précision).

- L'estimateur K du niveau est finalement très proche d'un estimateur composite par régression (cf. section 5.4) pour lequel on aurait fixé b=1. Lorsque la corrélation ρ est élevée (cas de l'emploi), l'approximation b=1 est acceptable et les valeurs optimales de ϕ pour ces deux estimateurs sont très proches (0,85 dans le cas de la variable emploi). En revanche, on ne peut plus supposer b=1 si la corrélation ρ est faible (cas du chômage). L'estimateur $\bar{y}^K_{t-1} + (\bar{y}_{t,c} - \bar{y}_{t-1,c})$ étant

alors moins bon que l'estimateur par régression $\overline{y}_{t,c} + b(\overline{y}_{t-1} - \overline{y}_{t-1,c})$, on lui accorde un poids ϕ plus faible.

- La valeur optimale de ϕ dépend de la grandeur à estimer (niveau, évolution trimestrielle ou glissement) et de la variable. L'optimum apparaît d'autant plus « plat » (donc le choix de la valeur de ϕ optimale moins critique) que les corrélations sont faibles.

- Bien entendu, les propriétés d'additivité des estimateurs ne seront conservées que si l'on retient une valeur commune pour ϕ. Dans la recherche d'un système de poids commun, il est légitime de favoriser l'estimation du chômage pour laquelle les gains sont les plus faibles. On remarquera à ce propos qu'un mauvais choix de ϕ peut rapidement conduire à des résultats dramatiques, i.e. à une variance supérieure à celle de l'estimateur naturel (lorsque ϕ tend vers 1, la variance du niveau tend vers $+\infty$). Une valeur de 0,6 (voir graphique 4 en annexe) apparaît être un bon compromis pour minimiser les variances des trois estimations. Le tableau ci-dessous donne l'efficacité de l'estimateur composite correspondant. Les gains obtenus sur le chômage ne sont pas trop détériorés par le choix de cette valeur commune de ϕ ; pour l'emploi, la perte est nettement plus importante, mais les gains associés demeurent néanmoins appréciables.

Efficacité de l'estimateur K commun

		Estimateur composite K	
		ϕ commun	**Efficacité**
Emploi	Niveau	0,6	131,5
	Evolution trimestrielle	0,6	194,5
	Glissement annuel	0,6	161,9
Chômage	Niveau	0,6	105,8
	Evolution trimestrielle	0,6	114,2
	Glissement annuel	0,6	115,1

Comparaison avec l'enquête annuelle

L'échantillon de l'enquête Emploi annuelle suit un plan de rotation à un niveau et est renouvelé par tiers tous les ans (voir section 6). Les niveaux et évolutions annuels sont calculés par les estimateurs naturels $\overline{y}^a_{t'}$ et $\overline{y}^a_{t'} - \overline{y}^a_{t'-1}$ (t' désigne ici des années) :

$$\bar{y}^a_{t'} = \frac{1}{3}\sum_{\ell=1}^{3} \bar{y}^a_{t',\ell} \qquad \text{avec} \quad V(\bar{y}^a_{t'}) = \frac{S^2}{n^a},\ n^a \text{ taille de l'échantillon Emploi annuel}$$

$$V(\bar{y}^a_{t'} - \bar{y}^a_{t'-1}) = 2(1-\rho_4(1-k^a))\frac{S^2}{n^a} \quad \text{où} \quad 1-k^a = \frac{2}{3},\ \rho_4 \text{ a été estimé à la section 6.}$$

On rappelle qu'il s'agit ici d'une estimation en moyenne sur le mois de mars de l'année

t', ce qui est différent d'une moyenne trimestrielle. Il est néanmoins intéressant de comparer les précisions des estimations issues des dispositifs annuel et en continu.

Le nombre n^a de personnes (de plus de 15 ans) interrogées à l'enquête annuelle est d'environ 150 000 alors qu'on envisage d'enquêter 75 000 personnes par trimestre pour l'enquête en continu. Le rapport des variances des estimateurs naturels entre les deux dispositifs est donc de $n/n^a \approx 1/2$ pour les niveaux et de $\frac{n}{n^a}\frac{1-\rho_4(1-k^a)}{1-\rho_4(1-k(4))}$ pour les glissements, soit environ 0,3 pour la variable emploi et 0,4 pour la variable chômage. L'efficacité de l'estimateur K commun par rapport aux estimateurs naturels de l'enquête annuelle est donnée dans le tableau ci-dessous. Ainsi, le dispositif d'enquête en continu entraînerait une perte de précision sensible sur l'estimation des niveaux et des évolutions annuelles due à un échantillon nettement plus petit et un taux de recouvrement annuel plus faible, particulièrement pénalisant lorsque la corrélation de la variable d'intérêt est forte (cas de l'emploi). Bien entendu, l'enquête Emploi en continu fournira en contrepartie des estimations d'évolutions trimestrielles.

Efficacité de l'estimateur K par rapport aux estimateurs naturels de l'enquête annuelle

		ϕ=0,6
Emploi	Niveau	66
	Glissement annuel	49
Chômage	Niveau	53
	Glissement annuel	47

8. Généralisation à des plans de sondage complexes

8.1 Généralités

Dans cette partie, on s'intéresse à l'estimation de l'évolution d'une caractéristique de la population dans le cas d'une enquête à échantillon rotatif à un niveau avec un plan de sondage complexe. Afin d'alléger les formules, on étudiera dans un premier temps le cas de 2 périodes notées 1 et 2 et les résultats seront ensuite généralisés au cas de plusieurs périodes. Conformément à la partie 2, on supposera que l'échantillon de chaque période est composé de m sous-échantillons élémentaires de taille identique n/m que l'on supposera entier. Chaque sous-échantillon est représentatif de l'ensemble de la population et la probabilité d'inclusion de l'individu i appartenant à l'un d'eux sera notée π^i, indépendante du rang d'interrogation. Autrement dit, les pondérations des individus appartenant à la vague l sont telles que $\sum_{i\in l} \frac{1}{\pi^i} = N$ où N est la taille de la population.

L'estimateur élémentaire du total de la variable Y est l'estimateur d'Horvitz-Thompson

$$\hat{Y}_{t,\ell} = \sum_{\ell} \frac{1}{\pi^i} y_t^i$$

et un estimateur élémentaire de la moyenne de Y sera

$$\hat{\bar{Y}}_{t,\ell} = \frac{1}{N} \sum_{\ell} \frac{1}{\pi^i} y_t^i$$

Par analogie au cas d'un plan de sondage aléatoire simple, on définit les estimateurs « naturels » du total et de la moyenne par

$$\hat{Y}_t = \frac{1}{m} \sum_{\ell=1}^{m} \hat{Y}_{t,\ell} \quad \text{et} \quad \hat{\bar{Y}}_t = \frac{1}{m} \sum_{\ell=1}^{m} \hat{\bar{Y}}_{t,\ell}$$

Le problème d'estimation dans les enquêtes à échantillon rotatif peut aussi être présenté sous un angle plus opérationnel lorsqu'on dispose d'un logiciel permettant de calculer la précision. Comme, pour une période donnée, m vagues indépendantes sont sollicitées simultanément, celles-ci peuvent être considérées comme provenant d'un même plan de sondage. En effet, on peut supposer que l'on a réalisé un plan de sondage stratifié au sein d'une population de taille m*N et que les vagues

disponibles ont été sélectionnées dans chacune des strates correspondant à la même population de taille N. Cette hypothèse revient à donner à chaque individu i le poids d'extrapolation $w^i = \frac{1}{m * \pi^i}$. Les formules des estimateurs et de leur précision restent identiques à celles présentées ci-dessous.

Variance des estimateurs élémentaires

Pour un sous-échantillon donné, la variance de l'estimation du total et de son évolution entre deux périodes consécutives s'écrit :

$$V(\hat{Y}_{t,l}) = V\left(\sum_{\ell} \frac{1}{\pi^i} y_t^i\right) \quad \text{et } V(\hat{Y}_{t,l} - \hat{Y}_{t-1,l-1}) = V\left(\sum_{\ell} \frac{1}{\pi^i} (y_t^i - y_{t-1}^i)\right)$$

(pour $\ell > 1$)

Pour l'estimation de ces variances, deux cas se présentent :

① **On dispose d'un logiciel** qui permet d'évaluer la précision de statistiques - notamment d'un total - issues d'enquêtes par sondage complexes.

Dans ce cas, l'estimation de ces variances ne pose aucune difficulté. Pour l'évolution, il suffit de construire une nouvelle variable z^i qui vaut $y_2^i - y_1^i$ pour chaque individu *i* et d'estimer la variance du total de la variable z.

② **On ne dispose pas d'un tel logiciel.**

Le calcul exact de la variance d'un total est d'autant plus difficile à conduire à la main que le plan de sondage est complexe. La solution consiste alors à essayer de se « ramener » au cas d'un sondage aléatoire simple où les calculs de variance sont abordables. Dans ce but, on a l'habitude de définir en théorie des sondages le design effect, c'est-à-dire le rapport entre la variance obtenue d'après le véritable plan de sondage à celle qu'on aurait obtenue avec des données issues d'un plan de sondage aléatoire simple. L'estimation de ce rapport permet, en particulier, d'apprécier un effet de grappe si le plan de sondage est à plusieurs degrés.

Plus précisément, en posant $\hat{T}$ (respectivement $\hat{T}^{sas}$) l'estimateur du total de la variable T pour un plan de sondage quelconque (respectivement un plan de sondage aléatoire simple sans remise de même taille fixe), le design effect est défini par

$$deff = \frac{V(\hat{T})}{V_{sas}(\hat{T}^{sas})}$$

Ainsi, si par des calculs antérieurs, il est possible d'avoir une idée de l'ordre de grandeur du design effect, la variance des estimateurs élémentaires s'en déduit facilement. Cependant, le design effect dépend à la fois du plan de sondage et des variables d'intérêt. Ainsi, on peut observer au sein d'une même enquête (enquêtes « Ménages » de l'Insee par exemple) des design effects qui varient de un à trois selon les variables. On supposera ici que l'on a défini un design effect « moyen » relatif à l'enquête et que celui-ci s'appliquera indifféremment à toute variable. On peut alors écrire,

$$V(\hat{Y}_{t,\ell}) = \text{deff}\, V(\hat{y}^{sas}_{t,\ell}) \qquad \text{où} \qquad \hat{y}^{sas}_{t,\ell} = N\overline{y}_{t,\ell}$$ est l'estimateur élémentaire du total de Y dans le cas d'un sondage aléatoire simple

et $$V(\hat{Y}_{t,\ell} - \hat{Y}_{t-1,\ell-1}) = \text{deff}\, V(\hat{y}^{sas}_{t,\ell} - \hat{y}^{sas}_{t-1,\ell-1})$$

Le calcul de la variance des estimateurs élémentaires du niveau et de l'évolution se ramène donc au cas d'un sondage aléatoire simple, à condition de disposer d'une estimation du design effect. Dans les sections suivantes, on montre que ce principe vaut aussi pour l'ensemble des estimateurs considérés dans les parties précédentes.

8.2 Estimateur naturel de l'évolution

L'estimateur naturel d'une évolution est

$$\hat{\overline{Y}}_2 - \hat{\overline{Y}}_1 = \frac{1}{N}\left(\hat{Y}_2 - \hat{Y}_1\right) \text{ soit } \quad \hat{\overline{Y}}_2 - \hat{\overline{Y}}_1 = \frac{1}{m}\sum_{\ell=1}^{m}\left(\hat{\overline{Y}}_{2,\ell} - \hat{\overline{Y}}_{1,\ell}\right)$$

Cette différence d'estimateurs peut s'écrire

$$\hat{\overline{Y}}_2 - \hat{\overline{Y}}_1 = \frac{1}{m}\frac{1}{N}\left[\sum_{S_c}\frac{1}{\pi^i}(y_2^i - y_1^i) + \sum_{S_{2,n_c}}\frac{1}{\pi^i}y_2^i - \sum_{S_{1,n_c}}\frac{1}{\pi^i}y_1^i\right]$$

Les trois termes du membre de droite étant supposés indépendants, la variance est par conséquent

$$V(\hat{\overline{Y}}_2 - \hat{\overline{Y}}_1) = \frac{1}{m^2}\frac{1}{N^2}\left[V\left(\sum_{S_c}\frac{1}{\pi^i}(y_2^i - y_1^i)\right) + V\left(\sum_{S_{2,n_c}}\frac{1}{\pi^i}y_2^i\right) + V\left(\sum_{S_{1,n_c}}\frac{1}{\pi^i}y_1^i\right)\right]$$

Si on dispose d'un logiciel de calcul de variance, on peut estimer directement chacun des trois termes. Pour le premier terme, il suffit de construire une nouvelle variable z^i qui vaut $y_2^i - y_1^i$ pour tous les individus interrogés lors des deux vagues successives et 0 pour les autres. Pour les deux autres, on considère la valeur de la variable d'intérêt pour les individus interrogés une seule fois et on met à 0 les autres valeurs de l'échantillon considéré.

On peut aussi estimer cette variance directement à partir de la variance des estimateurs élémentaires puisque

$$V(\hat{\bar{Y}}_2 - \hat{\bar{Y}}_1) = \frac{1}{m^2}\frac{1}{N^2}\left[\sum_{\ell=2}^{m} V(\hat{Y}_{2,\ell} - \hat{Y}_{1,\ell-1}) + V(\hat{Y}_{2,1}) + V(\hat{Y}_{1,6})\right]$$

Dans le cas où l'on ne dispose pas d'un logiciel de calcul de variance, il suffit d'introduire le design effect :

$$V(\hat{\bar{Y}}_2 - \hat{\bar{Y}}_1) = \frac{1}{m^2}\frac{1}{N^2}\left[\sum_{\ell=2}^{m} \text{deff}\, V(\hat{y}^{sas}_{2,\ell} - \hat{y}^{sas}_{1,\ell-1}) + \text{deff}\, V(\hat{y}^{sas}_{2,1}) + \text{deff}\, V(\hat{y}^{sas}_{1,6})\right]$$

soit finalement

$$V(\hat{\bar{Y}}_2 - \hat{\bar{Y}}_1) = \text{deff}\, V(\bar{y}_2 - \bar{y}_1)$$

8.3 Estimateur sur la partie commune

Par analogie au cas d'un sondage aléatoire simple, l'estimateur sur la partie commune (m-1 vagues) s'écrit :

$$\hat{\bar{Y}}_{2,c} - \hat{\bar{Y}}_{1,c} = \frac{1}{N}(\hat{Y}_{2,c} - \hat{Y}_{1,c})$$

soit

$$\hat{\bar{Y}}_{2,c} - \hat{\bar{Y}}_{1,c} = \frac{1}{N}\frac{1}{m-1}\sum_{i \in s_c}\frac{1}{\pi^i}(y_2^i - y_1^i)$$

$$\hat{\bar{Y}}_{2,c} - \hat{\bar{Y}}_{1,c} = \frac{1}{N}\frac{1}{m-1}\sum_{\ell=2}^{m}\left(\hat{Y}_{2,1} - \hat{Y}_{1,\ell-1}\right)$$

La variance de cet estimateur vaut donc

$$V(\hat{\bar{Y}}_{2,c} - \hat{\bar{Y}}_{1,c}) = \frac{1}{N^2}\frac{1}{(m-1)^2} V\left(\sum_{i \in s_c} \frac{1}{\pi^i}\left(y_2^i - y_1^i\right)\right)$$

Là encore, cette variance peut s'estimer directement avec un logiciel approprié. Sinon, on peut toujours se ramener au cas d'un sondage aléatoire simple puisque

$$\hat{\bar{Y}}_{2,c} - \hat{\bar{Y}}_{1,c} = \frac{1}{N}\frac{1}{m-1}\sum_{\ell=2}^{m}\left(\hat{Y}_{2,\ell} - \hat{Y}_{1,\ell-1}\right)$$

Par un raisonnement analogue au précédent, il devient évident que

$$V(\hat{\bar{Y}}_{2,c} - \hat{\bar{Y}}_{1,c}) = \text{deff}\, V(\bar{y}_{2,c} - \bar{y}_{1,c})$$

On constate que comparer les deux estimateurs envisagés jusqu'à présent revient simplement à comparer leurs équivalents dans le cas d'un sondage aléatoire simple, c'est-à-dire au résultat de la section 5.

8.4 Estimateur composite direct de l'évolution

Dans le cas d'un sondage complexe, l'estimateur composite direct de l'évolution s'exprime en retenant une forme similaire à celui défini par un sondage aléatoire simple :

$$\Delta\hat{\bar{Y}}^{*} = \phi(\hat{\bar{Y}}_{2,c} - \hat{\bar{Y}}_{1,c}) + (1-\phi)(\hat{\bar{Y}}_{2,nc} - \hat{\bar{Y}}_{1,nc})$$

Les deux évolutions étant indépendantes, la variance de l'estimateur composite optimal est avec les notations similaires à celles de la partie 5.4 :

$$V(\Delta\hat{\bar{Y}}^{*}) = \frac{V_{nc} V_c}{V_{nc} + V_c}$$

Là encore, un calcul immédiat montre que $V(\Delta\hat{\bar{Y}}^{*}) = \text{deff}\, V(\Delta\bar{y}^{*})$ et que, par ailleurs, les poids optimaux sont identiques. La comparaison avec l'estimateur naturel conduit, par conséquent, à des conclusions identiques à celles exposées dans le cas du sondage aléatoire simple.

8.5 L'estimateur K pour plusieurs périodes

Les développements de la partie 6.2. et des annexes 3 et 4 s'adaptent facilement au cas d'un plan de sondage complexe. En effet, comme nous l'avons déjà remarqué, les expressions de la variance du K estimateur ainsi que celles de l'évolution et du glissement annuel sont fonction de la variance de l'estimateur naturel et de la suite des corrélations temporelles ρ_s.

Les formules obtenues sont donc identiques à celles obtenues dans le cas d'un sondage aléatoire simple à condition de poser $V(\hat{\bar{Y}}_t) = \sigma^2$ (à la place de S^2/n, soit $\sigma^2 = \text{deff}\ S^2/n$) et de définir ρ_s par :

$$\rho_S = \frac{\text{Cov}(\hat{Y}_{t,\ell}, \hat{Y}_{t-s,\ell-s})}{\sqrt{V(\hat{Y}_{t,\ell})}\ \sqrt{V(\hat{Y}_{t-s,\ell-s})}}$$

où les différents termes sont des **variances** et des **covariances d'échantillonnage.**

En supposant la stationnarité du processus, ρ_s s'écrit plus simplement $\rho_S = \frac{\text{Cov}(\hat{Y}_{t,\ell}, \hat{Y}_{t-s,\ell-s})}{V(\hat{Y}_{t,\ell})}$. Notons que contrairement au cas d'un sondage aléatoire simple, ρ_s ne correspond plus pour un sondage complexe au coefficient de corrélation linéaire empirique calculé sur l'échantillon.

Par conséquent, toute la difficulté consiste à estimer du mieux possible la suite des ρ_s à partir des données. Une solution est de remplacer le terme de covariance d'échantillonnage souvent difficile à estimer par différents termes de variance ce qui conduit par exemple à :

$$\rho_S = \frac{V(\hat{Y}_{t,\ell} + \hat{Y}_{t-s,\ell-s}) - 2V(\hat{Y}_{t,\ell})}{2V(\hat{Y}_{t,\ell})}$$

où $V(\hat{Y}_{t,\ell} + \hat{Y}_{t-s,\ell-s})$ s'estime à partir des individus interrogés lors des vagues d'enquêtes des dates *t* et *t-s* en construisant pour chaque individu *i* une nouvelle variable z, définie par $y_t^i + y_{t-s}^i$, et en calculant la variance de l'estimateur d'Horvitz-Thompson du total de cette variable.

Le coefficient de corrélation ρ_s, supposé indépendant de t et du numéro de sous-échantillon ℓ, ne dépend que de s ; Cependant, le choix de t et de ℓ n'est pas sans conséquence lorsque l'on souhaite estimer ρ_s. En effet, il y a autant de valeurs différentes pour l'estimation ρ_s que de choix possibles pour t et ℓ. Afin d'avoir une meilleure estimation (au sens de la précision), il est préférable d'utiliser toute l'information disponible à une date donnée t. Lee (1990) suggère, par exemple, d'estimer les différents termes de variance sur l'ensemble des sous-échantillons qui sont communs entre les dates t et t-s.

Là encore, si on considère le design effect identique pour chaque variable, les ρ_s ainsi calculés ne seront vraisemblablement pas très éloignés de ceux obtenus pour un plan de sondage aléatoire simple, c'est à dire des coefficients de corrélation linéaire empiriques.

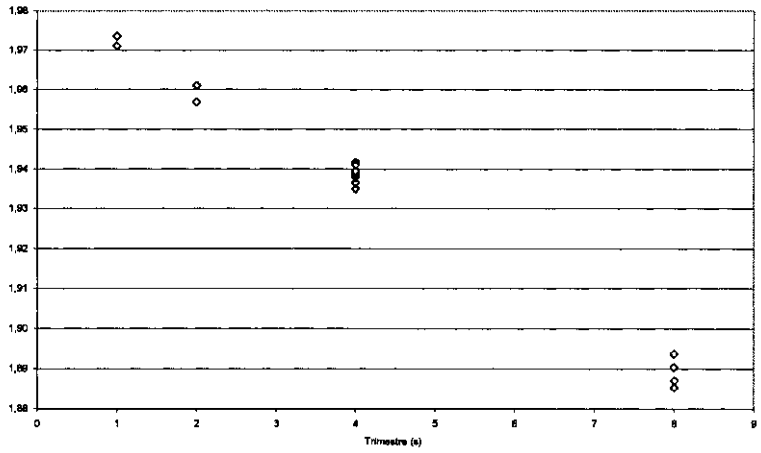
8.6 Remarques

Pour certains plans de sondage, la taille de l'échantillon est variable. C'est par exemple le cas pour un plan de sondage bernoullien, poissonnien mais aussi pour un plan de sondage à plusieurs degrés dont les deux degrés sont des sondages aléatoires simples ou un plan de sondage en grappe. Lorsque la taille de la population est connue, vaut-il mieux, pour estimer une moyenne, utiliser l'estimateur d'Horvitz-Thompson $\hat{\bar{Y}}_\pi = \frac{\hat{Y}_\pi}{N}$ ou l'estimateur par le ratio $\hat{\bar{Y}}_r = \frac{\hat{Y}_\pi}{N_\pi}$? Le second estimateur est plus « logique » que le premier au sens où si on recueille la même valeur c pour l'ensemble des individus enquêtés, l'estimateur $\hat{\bar{Y}}_r$ conduit à cette valeur alors que $\hat{\bar{Y}}_\pi$ est égal à $c\frac{\hat{N}_\pi}{N}$. Par construction, l'estimateur par le ratio, dont la variance correspond à celle de l'estimateur du total de la variable artificielle $u^i = \frac{1}{N}(y^i - \hat{\bar{Y}}_\pi)$ est en général plus précis que l'estimateur d'Horvitz-Thompson. En effet, l'estimateur par le ratio $\bar{Y}_r$, assurant une certaine stabilité à l'estimateur, se comporte mieux que $\hat{\bar{Y}}_\pi$. Ainsi, si la taille de l'échantillon obtenue est supérieure ou inférieure à son espérance, le numérateur et le dénominateur comportent le même nombre de termes.

9. Annexes

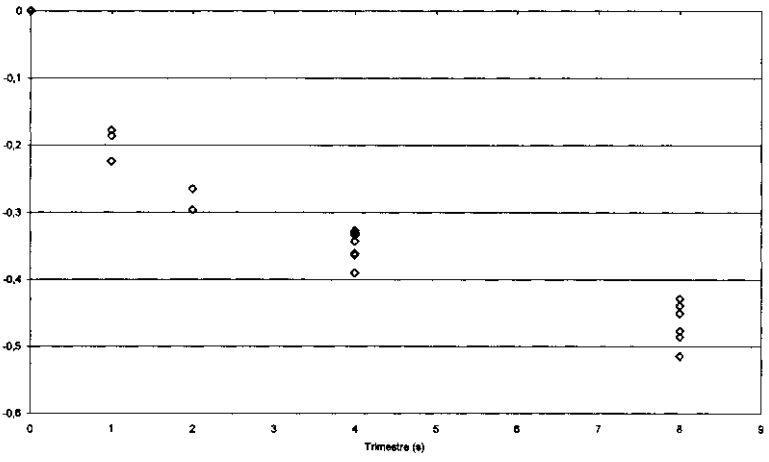
Graphe 1 : Log(*c(s)*-Log(1-*k(s)*)

Dispositif léger pour s=1,2 et enquête Emploi annuelle pour s=4,8.



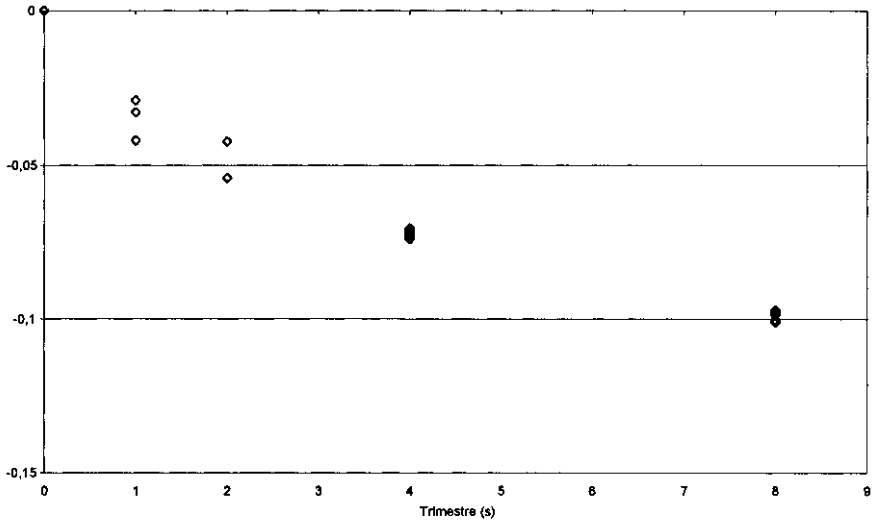
Graphe 2 : Log(ρ(s)) (Chômage)

Dispositif léger pour s=1,2 et enquête Emploi annuelle pour s=4,8.

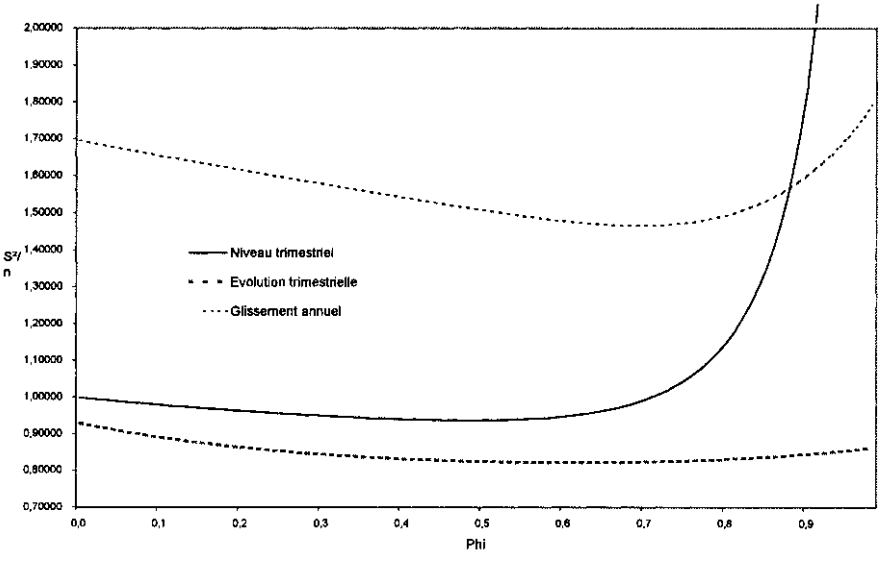


Graphe 3 : Log(ρ(s)) (Emploi)

Dispositif léger pour s=1,2 et enquête Emploi annuelle pour s=4,8.



Graphe 4 : Variance de l'estimateur K (Chômage)



Graphe 5 : Variance de l'estimateur K (Emploi)

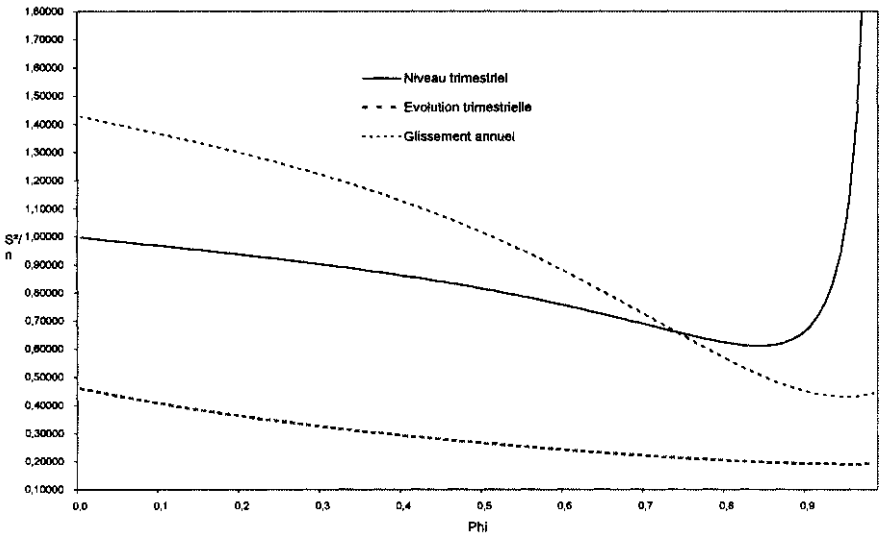


Tableau 1 : efficacité de l'estimateur K pour différents profils de corrélation

Le profil est du type (0,91-*u*; 0,88-*u*; 0,86-*u*; 0,84-*u*; 0,82-*u*)

u	*0*	*0,01*	*0,02*	*0,03*	*0,04*	*0,05*	*0,06*	*0,07*	*0,08*	*0,09*
Niveau	159,3	155,0	151,1	147,6	144,3	141,3	138,6	136,0	133,7	131,5
Evolution trimestrielle	228,5	212,7	199,9	189,1	180,1	172,3	165,6	159,8	154,6	150,1
Glissement annuel	310,3	289,8	272,0	256,6	243,0	231,0	220,3	210,7	202,1	194,2

Annexe 1 : La technique des moindres carrés appliquée à l'estimation de flux

Fuller (1990) a appliqué la technique des moindres carrés au problème de l'estimation de flux, sur l'exemple des transitions entre emploi et chômage entre deux périodes. Le problème consiste alors à estimer les proportions définies dans le tableau croisé ci-dessous :

		Période 2		
		Emploi	Chômage	Total
Période 1	Emploi	P_{EE}	P_{EC}	$P_{E.}$
	Chômage	P_{CE}	P_{CC}	$P_{C.}$
	Total	$P_{.E}$	$P_{.C}$	1

On suppose implicitement que la population active est constante dans le temps. Dans le cas contraire, il serait nécessaire de considérer un tableau de dimension 3x3 en y ajoutant la catégorie des inactifs.

Pour un tableau 2x2, il suffit d'estimer trois grandeurs. Retenons, par exemple, les proportions d'individus en emploi aux deux périodes P_{EE} ainsi que les marges $P_{.E}$ et $P_{E.}$ et posons $\Theta = (P_{E.}, P_{.E}, P_{EE})'$.

Le vecteur d'observation est composé des proportions
$\overline{Y} = (\hat{P}_{E.,nc}, \hat{P}_{E.,c}, \hat{P}_{EE,c}, \hat{P}_{.E,c}, \hat{P}_{.E,nc})$ et le modèle linéaire peut se mettre sous la forme habituelle

$$\overline{Y} = X\Theta + e \qquad \text{où } X = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$

Sous l'hypothèse d'un sondage aléatoire simple, la matrice de variance-covariance Ω de e s'estime facilement et l'estimateur des moindres carrés est
$\hat{\Theta} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}Y$

Par rapport aux estimateurs naturels des proportions, l'estimateur des moindres carrés apporte un gain appréciable sur la probabilité croisée P_{EE} et plus modeste sur les marges $P_{.E}$ et $P_{E.}$. Fuller examine aussi le cas où l'on s'interdit de réviser les estimations de la première période, c'est à dire ici $P_{.E}$. Pour ce faire, il développe une procédure de moindres carrés généralisés contraints. La perte d'efficacité sur l'estimation de P_{EE} et $P_{E.}$ serait très faible.

Annexe 2 : Variance de l'estimateur de Patterson

L'estimateur de Patterson s'écrit

$$\bar{y}_t^{Pat} = (1-\phi)\bar{y}_{t,nc} + \phi(\bar{y}_{t,c} + \rho(\bar{y}_{t-1}^{Pat} - \bar{y}_{t-1,c}))$$

Les deux termes étant indépendants, la variance vaut

$$V(\bar{y}_t^{Pat}) = (1-\phi)^2 V(\bar{y}_{t,nc}) + \phi^2 V(\bar{y}_{t,c}^{Reg})$$

soit
$$V(\bar{y}_t^{Pat}) = (1-\phi)^2 \frac{S^2}{n_{nc}} + \phi^2 \left[(1-\rho^2)\frac{S^2}{n_c} + \rho^2 V(\bar{y}_{t-1}^{Pat}) \right]$$

$$V(\bar{y}_t^{Pat}) = \frac{S^2}{n}\left[\frac{(1-\phi)^2}{k} + \frac{\phi^2}{1-k}(1-\rho^2) + \phi^2\rho^2 V(\bar{y}_{t-1}^{Pat})\frac{n}{S^2} \right]$$

Par conséquent, $V(\bar{y}_t^{Pat})/\frac{S^2}{n}$ suit une relation de récurrence du type

$$u_t = a + bu_{t-1} \text{ où } \quad a = \frac{(1-\phi)^2}{k} + \frac{\phi^2}{1-k}(1-\rho^2) \text{ et } b = \phi^2\rho^2$$

Soit
$$u_t = \frac{a(1-b^{t-1})}{1-b} + b^{t-1}u_1 \quad (\text{avec } u_1 = 1)$$

Puisque $b = \phi^2\rho^2$ est inférieur à 1 (le cas $\phi = 1$ n'est de toute façon pas très intéressant), u_t converge vers $a/(1-b)$. Par conséquent,

$$\boxed{V_\infty(\bar{y}^{Pat}) = \frac{S^2}{n}\frac{(1-k)(1-\phi)^2 + k\phi^2(1-\rho^2)}{k(1-k)(1-\rho^2\phi^2)}}$$

Cette variance peut être minimisée relativement à ϕ. En dérivant cette expression par rapport à ϕ, on trouve la condition du premier ordre :

$$\rho^2(1-k)\phi^2 - ((1-k)(1+\rho^2) + k(1-\rho^2))\phi + (1-k) = 0$$

La seule racine acceptable (inférieure à 1) est :

$$\phi_{opt} = \frac{(1-\rho^2)+2\rho^2(1-k)-\sqrt{(1-\rho^2)(1-\rho^2+4k(1-k)\rho^2)}}{2\rho^2(1-k)}$$

soit

$$\boxed{\phi_{opt} = 1-\frac{\sqrt{(1-\rho^2)}\left(\sqrt{1-\rho^2+4k(1-k)\rho^2}-\sqrt{(1-\rho^2)}\right)}{2\rho^2(1-k)}}$$

Pour cette valeur de ϕ, il est facile de montrer que

$$\boxed{V_\infty(\bar{y}_{opt}^{Pat}) = \frac{S^2}{n}\frac{1-\phi_{opt}}{k}}$$

En effet,

$$V_\infty(\bar{y}_{opt}^{Pat}) = \frac{S^2}{n}\frac{(1-k)(1-\phi_{opt})^2+k\phi_{opt}^2(1-\rho^2)}{k(1-k)(1-\rho^2\phi_{opt}^2)}$$

soit

$$V_\infty(\bar{y}_{opt}^{Pat}) = \frac{S^2}{kn}\left[1+\phi_{opt}\frac{(1-k)(2+\phi_{opt})+k\phi_{opt}(1-\rho^2)+\rho^2\phi_{opt}(1-k)}{(1-k)(1-\rho^2\phi_{opt}^2)}\right]$$

La condition du premier ordre définissant ϕ_{opt}

$$\rho^2(1-k)\phi_{opt}^2-((1-k)(1+\rho^2)+k(1-\rho^2))\phi_{opt}+(1-k)=0$$

équivaut à

$$(1-k)(1-\rho^2\phi_{opt}^2) = 2(1-k)-\phi_{opt}((1-k)(1+\rho^2)+k(1-\rho^2))$$

On obtient finalement

$$V_\infty(\bar{y}_{opt}^{Pat}) = \frac{S^2}{n}\frac{1-\phi_{opt}}{k}$$

Annexe 3 : Variance de l'estimateur K

L'estimateur K s'écrit :

$$\bar{y}_t^K = (1-\phi)\bar{y}_t + \phi(\bar{y}_{t-1}^K + d_{t,t-1}) \quad \text{avec} \quad d_{t,t-1} = \frac{1}{5}\sum_{\ell=2}^{6}\bar{y}_{t,\ell} - \bar{y}_{t-1,\ell-1}$$

$$\text{et} \quad \bar{y}_t = \frac{1}{6}\sum_{\ell=1}^{6}\bar{y}_{t,\ell}$$

Sa variance peut alors s'écrire

$$V(\bar{y}_t^K) = \phi^2 V(\bar{y}_{t-1}^K) + (1-\phi)^2 V(\bar{y}_t) + \phi^2 V(d_{t,t-1}) + 2\phi(1-\phi)\mathrm{Cov}(\bar{y}_t, \bar{y}_{t-1}^K) + 2\phi(1-\phi)\mathrm{Cov}(\bar{y}_t, d_{t,t-1}) + 2\phi^2\mathrm{Cov}(\bar{y}_{t-1}^K, d_{t,t-1})$$

On se place directement après un nombre suffisant de périodes pour considérer la forme et la variance de l'estimateur composite stationnaires, c'est à dire

$$V(\bar{y}_t^K) = V(\bar{y}_{t-1}^K) \quad \text{pour} \quad t \geq t_0$$

Notons dès à présent que le cas ϕ=1 est incompatible avec l'hypothèse de stationnarité de la variance. L'intérêt de ce cas étant très académique, on suppose donc ϕ<1 et la variance de l'estimateur K s'écrit

$$(1-\phi^2)V(\bar{y}_t^K) = (1-\phi)^2 V(\bar{y}_t) + \phi^2 V(d_{t,t-1}) + 2\phi(1-\phi)\mathrm{Cov}(\bar{y}_t, \bar{y}_{t-1}^K) + 2\phi(1-\phi)\mathrm{Cov}(\bar{y}_t, d_{t,t-1}) + 2\phi^2\mathrm{Cov}(\bar{y}_{t-1}^K, d_{t,t-1}) \quad (2.1)$$

Il faut maintenant calculer les cinq termes du membre de droite. Pour ceux faisant intervenir l'estimateur K retardé d'une période $\bar{y}_{t-1}^K$, une astuce consiste à ré-exprimer ce terme en fonction de l'estimateur naturel $\bar{y}_t$ et des évolutions $d_{t,t-1}$ ainsi que leur retards en itérant la définition de l'estimateur K :

$$\bar{y}_{t-1}^K = \sum_{j=1}^{6}\phi^{j-1}(1-\phi)\bar{y}_{t-j} + \phi^j d_{t-j,t-j-1} + \phi^6\bar{y}_{t-7}^K \quad (2.2)$$

Bien entendu, l'idée sous-jacente est que la covariance entre $d_{t,t-1}$ ou $\bar{y}_t$ et $\bar{y}^K_{t-7}$ sera nulle et qu'il ne reste plus qu'à calculer les covariances avec les $d_{t-j,t-j-1}$ et les $\bar{y}_{t-j}$ pour j=1 à 6.

2.A $V(\bar{y}_t)$

Par hypothèse,

$$\boxed{V(\bar{y}_t) = \frac{S^2}{n}} \tag{2.3}$$

2.B $V(d_{t,t-1})$

$$V(d_{t,t-1}) = \frac{1}{25}\sum_{\ell=2}^{6} V(\bar{y}_{t,\ell}) + V(\bar{y}_{t-1,\ell-1}) - 2\mathrm{Cov}(\bar{y}_{t,\ell}, \bar{y}_{t-1,\ell-1})$$

or $$\mathrm{Cov}(\bar{y}_{t,\ell}, \bar{y}_{t-1,\ell-1}) = \rho_1 6\frac{S^2}{n}$$

soit $$\boxed{V(d_{t,t-1}) = \frac{12}{5}(1-\rho_1)\frac{S^2}{n}} \tag{2.4}$$

2.C $\mathrm{Cov}(\bar{y}_t, \bar{y}^K_{t-1})$

En utilisant l'écriture développée de $\bar{y}^K_{t-1}$ (2.2), on a

$$\mathrm{Cov}(\bar{y}_t, \bar{y}^K_{t-1}) = \sum_{j=1}^{6} \phi^{j-1}(1-\phi)\mathrm{Cov}(\bar{y}_t, \bar{y}_{t-j}) + \phi^j \mathrm{Cov}(\bar{y}_t, d_{t-j,t-j-1}) + 0$$

Pour $1 \le j \le 6$,

$$\mathrm{Cov}(\bar{y}_t, \bar{y}_{t-j}) = \mathrm{Cov}\left(\frac{1}{6}\sum_{\ell=1}^{6} y_{t,\ell}, \frac{1}{6}\sum_{\ell=1}^{6} y_{t-j,\ell}\right)$$

$$= \frac{1}{36}(6-j)\rho_j 6\frac{S^2}{n}$$

soit $$\mathrm{Cov}(\bar{y}_t, \bar{y}_{t-j}) = \frac{1}{6}\frac{S^2}{n}(6-j)\rho_j$$

De même, pour $1 \le j \le 6$,

$$\mathrm{Cov}(\bar{y}_t, d_{t-j,t-j-1}) = \frac{1}{6}\frac{1}{5}\mathrm{Cov}\left(\sum_{\ell=1}^{6} y_{t,\ell}, \sum_{\ell=2}^{6} \bar{y}_{t-j,\ell} - \bar{y}_{t-j-1,\ell-1}\right)$$
$$= \frac{1}{30}\left(6\frac{S^2}{n}(5-j)\rho_j - 6\frac{S^2}{n}(5-j)\rho_{j+1}\right) * (j \le 5)$$

soit $$\mathrm{Cov}(\bar{y}_t, d_{t-j,t-j-1}) = \frac{1}{5}\frac{S^2}{n}(5-j)(\rho_j - \rho_{j+1}) * (j \le 5)$$

Finalement,

$$\boxed{\mathrm{Cov}(\bar{y}_t, \bar{y}_{t-1}^K) = \sum_{j=1}^{5}\frac{1}{6}\frac{S^2}{n}(6-j)\rho_j\phi^{j-1}(1-\phi) + \frac{1}{5}\frac{S^2}{n}(5-j)(\rho_j - \rho_{j+1})\phi^j} \quad (2.5)$$

2.D $\mathrm{Cov}(\bar{y}_t, d_{t,t-1})$

$$\mathrm{Cov}(\bar{y}_t, d_{t,t-1}) = \frac{1}{6}\frac{1}{5}\mathrm{Cov}\left(\frac{1}{6}\sum_{\ell=1}^{6} y_{t,\ell}, \sum_{\ell=2}^{6} \bar{y}_{t,\ell} - \bar{y}_{t-1,\ell-1}\right)$$
$$= \frac{1}{30}\left(5*6\frac{S^2}{n} - 5*6\frac{S^2}{n}\rho_1\right)$$

soit $$\boxed{\mathrm{Cov}(\bar{y}_t, d_{t,t-1}) = (1-\rho_1)\frac{S^2}{n}} \quad (2.6)$$

2.E $\mathrm{Cov}(\bar{y}_{t-1}^K, d_{t,t-1})$

$$\mathrm{Cov}(\bar{y}_{t-1}^K, d_{t,t-1}) = \sum_{j=1}^{6}\phi^{j-1}(1-\phi)\mathrm{Cov}(\bar{y}_{t-j}, d_{t,t-1}) + \phi^j\mathrm{Cov}(d_{t-j,t-j-1}, d_{t,t-1})$$

Pour $1 \le j \le 6$,

$$\mathrm{Cov}(\bar{y}_{t-j}, d_{t,t-1}) = \frac{1}{6}\frac{1}{5}\mathrm{Cov}\left(\sum_{\ell=1}^{6} y_{t-j,\ell}, \sum_{\ell=2}^{6} \bar{y}_{t,\ell} - \bar{y}_{t-1,\ell-1}\right)$$

$$= \frac{1}{30}\left((6-j)\rho_j 6\frac{S^2}{n} - (6-j)\rho_{j-1} 6\frac{S^2}{n}\right)$$

soit $$\mathrm{Cov}(\bar{y}_{t-j}, d_{t,t-1}) = \frac{1}{5}(6-j)(\rho_j - \rho_{j-1})\frac{S^2}{n}$$

Pour $1 \le j \le 6$,

$$Cov(d_{t-j,t-j-1}, d_{t,t-1}) = \frac{1}{5}\frac{1}{5}Cov\left(\sum_{\ell=2}^{6} \bar{y}_{t-j,\ell} - \bar{y}_{t-j-1,\ell-1}, \sum_{\ell=2}^{6} \bar{y}_{t,\ell} - \bar{y}_{t-1,\ell-1}\right)$$

$$= \frac{1}{25} 6\frac{S^2}{n}\left((5-j)(j \le 5)\rho_j - (5-j)(j \le 5)\rho_{j-1} - (5-j)(j \le 5)\rho_{j+1} + (5-j)(j \le 5)\rho_j\right)$$

soit $$\mathrm{Cov}(d_{t-j,t-j-1}, d_{t,t-1}) = \frac{1}{25} 6\frac{S^2}{n}(5-j)(j \le 5)\left(2\rho_j - \rho_{j-1} - \rho_{j+1}\right)$$

Finalement,

$$\boxed{\mathrm{Cov}(\bar{y}^K_{t-1}, d_{t,t-1}) = \sum_{j=1}^{5} \phi^{j-1}(1-\phi)\frac{1}{5}\frac{S^2}{n}(6-j)(\rho_j - \rho_{j-1}) + \phi^j \frac{6}{25}\frac{S^2}{n}(5-j)(2\rho_j - \rho_{j-1} - \rho_{j+1})}$$

(2.7)

Pour calculer $V(\bar{y}^K_t)$, il ne reste plus qu'à remplacer les termes du membre de droite de (2.1) par leur expression donnée par les formules (2.3) à (2.7).

Annexe 4 : Variance de l'évolution de l'estimateur K.

Cas d'une évolution trimestrielle : $\bar{y}_t^K - \bar{y}_{t-1}^K$

On suppose là aussi que l'estimateur a atteint un régime stationnaire,

$$V(\bar{y}_t^K - \bar{y}_{t-1}^K) = 2V(\bar{y}_t^K) - 2Cov(\bar{y}_t^K, \bar{y}_{t-1}^K)$$

Or $$\bar{y}_t^K = (1-\phi)\bar{y}_t + \phi(\bar{y}_{t-1}^K + d_{t,t-1})$$

permet de calculer facilement le terme de covariance. En effet,

$$\bar{y}_t^K - \phi\bar{y}_{t-1}^K) = (1-\phi)\bar{y}_t + \phi d_{t,t-1}$$

soit

$$(1+\phi^2)V(\bar{y}_t^K) - 2\phi Cov(\bar{y}_t^K, \bar{y}_{t-1}^K) = (1-\phi)^2 V(\bar{y}_t) + \phi^2 V(d_{t,t-1}) + 2\phi(1-\phi)Cov(\bar{y}_t, d_{t,t-1})$$

Or d'après (2.3), (2.4) et (2.6)

$$V(\bar{y}_t) = \frac{S^2}{n}, \quad V(d_{t,t-1}) = \frac{12}{5}(1-\rho_1)\frac{S^2}{n} \quad \text{et}$$

$$Cov(\bar{y}_t, d_{t,t-1}) = (1-\rho_1)\frac{S^2}{n}$$

donc, pour $\phi \neq 0$ (le cas $\phi = 0$, qui correspond à l'estimateur naturel, a déjà été étudié)

$$2Cov(\bar{y}_t^K, \bar{y}_{t-1}^K) = \frac{1+\phi^2}{\phi}V(\bar{y}_t^K) - \frac{S^2}{n}\left[\frac{(1-\phi)^2}{\phi} + \frac{12}{5}\phi(1-\rho_1) + 2(1-\phi)(1-\rho_1)\right]$$

Soit finalement,

$$\boxed{V(\bar{y}_t^K - \bar{y}_{t-1}^K) = \frac{-(1-\phi)^2}{\phi}V(\bar{y}_t^K) - \frac{S^2}{n}\frac{1}{\phi}\left[(1-\phi)^2 + \frac{2}{5}\phi(\phi+5)(1-\rho_1)\right]}$$

Cas d'un glissement annuel : $\bar{y}_t^K - \bar{y}_{t-4}^K$

De la même façon, on peut écrire en régime stationnaire

$$V(\bar{y}_t^K - \bar{y}_{t-4}^K) = 2V(\bar{y}_t^K) - 2Cov(\bar{y}_t^K, \bar{y}_{t-4}^K)$$

Pour le calcul du terme de covariance, il suffit d'écrire $\bar{y}_t^K$ et $\bar{y}_{t-4}^K$ sous forme développée :

$$\bar{y}_t^K = \sum_{i=0}^{3} \phi^i (1-\phi)\bar{y}_{t-i} + \phi^i d_{t-i,t-i-1} + \phi^4 \bar{y}_{t-4}^K$$

et $$\bar{y}_{t-4}^K = \sum_{j=4}^{9} \phi^{j-4}(1-\phi)\bar{y}_{t-j} + \phi^{j-3} d_{t-j,t-j-1} + \phi^6 \bar{y}_{t-10}^K$$

Soit

$$Cov(\bar{y}_t^K, \bar{y}_{t-4}^K) = \sum_{i=0}^{3} \phi^i (1-\phi) Cov(\bar{y}_{t-i}, \bar{y}_{t-4}^K) + \phi^i Cov(d_{t-i,t-i-1}, \bar{y}_{t-4}^K) + \phi^4 V(\bar{y}_{t-4}^K)$$

3.A $Cov(\bar{y}_{t-i}, \bar{y}_{t-4}^K)$

$$Cov(\bar{y}_{t-i}, \bar{y}_{t-4}^K) = \sum_{j=4}^{9} \phi^{j-4}(1-\phi) Cov(\bar{y}_{t-i}, \bar{y}_{t-j}) + \phi^{j-3} Cov(d_{t-j,t-j-1}, \bar{y}_{t-i})$$

Pour $4 \le j \le 9$,

$$Cov(\bar{y}_{t-i}, \bar{y}_{t-j}) = \frac{1}{6}\frac{1}{6} Cov\left(\sum_{\ell=1}^{6} y_{t-i,\ell}, \sum_{\ell=1}^{6} y_{t-j,\ell}\right)$$

$$= \frac{1}{6}\frac{S^2}{n}(6 - |i-j|)\rho_{|i-j|} * (|i-j| \le 5)$$

Pour $4 \le j \le 9$,

$$\mathrm{Cov}(d_{t-j,t-j-1}, \bar{y}_{t-i}) = \frac{1}{6}\frac{1}{5}\mathrm{Cov}\left(\sum_{\ell=1}^{6}\bar{y}_{t-i,\ell}, \sum_{\ell=2}^{6}\bar{y}_{t-j,\ell} - \bar{y}_{t-j-1,\ell-1}\right)$$
$$= \frac{1}{5}\frac{S^2}{n}(5-|i-j|)(\rho_{|i-j|} - \rho_{|i-j+1|}) * (|i-j| \le 5)$$

3.B $\mathrm{Cov}(d_{t-i,t-i-1}, \bar{y}^K_{t-4})$

$$\mathrm{Cov}(d_{t-i,t-i-1}, \bar{y}^K_{t-4}) = \sum_{j=4}^{9}\phi^{j-4}(1-\phi)\mathrm{Cov}(d_{t-i,t-i-1}, \bar{y}_{t-j}) + \phi^{j-3}\mathrm{Cov}(d_{t-i,t-i-1}, d_{t-j,t-j-1})$$

Pour $4 \le j \le 9$,

$$\mathrm{Cov}(d_{t-i,t-i-1}, \bar{y}_{t-j}) = \frac{1}{5}\frac{S^2}{n}(5-|i-j|)(\rho_{|i-j|} - \rho_{|i-1-j|}) * (|i-j| \le 5)$$

Pour $4 \le j \le 9$,

$$\mathrm{Cov}(d_{t-i,t-i-1}, d_{t-j,t-j-1}) = \frac{1}{25}\mathrm{Cov}\left(\sum_{\ell=2}^{6}\bar{y}_{t-j,\ell} - \bar{y}_{t-j-1,\ell-1}, \sum_{\ell=2}^{6}\bar{y}_{t-i,\ell} - \bar{y}_{t-i-1,\ell-1}\right)$$
$$= \frac{6}{25}\frac{S^2}{n}(5-|i-j|)(2\rho_{|i-j|} - \rho_{|i-j+1|} - \rho_{|i-j-1|}) * (|i-j| \le 5)$$

Finalement, la variance du glissement annuel se calcule en injectant les résultats de 3.A et 3.B dans l'expression suivante :

$$V(\bar{y}^K_t - \bar{y}^K_{t-4}) = 2(1-\phi^4)V(\bar{y}^K_t) - 2\sum_{i=0}^{3}\phi^i(1-\phi)\mathrm{Cov}(\bar{y}_{t-i}, \bar{y}^K_{t-4}) + \phi^i\mathrm{Cov}(d_{t-i,t-i-1}, \bar{y}^K_{t-4})$$

Bibliographie

Ansieau, D. (1998). Quelles mesures prendre pour limiter les effets de l'attrition d'un panel lors de la collecte ? *Actes des Journées de Méthodologie Statistique*, Insee Méthodes 84-86, 83-100.

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

Binder, D. A. and Dick, J. P. (1989). Modelling and Estimation for repeated Surveys. *Survey Methodology*, 1, 29-45.

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

Cochran, W. G. (1977). *Sampling Techniques*. Third Edition, Wiley,Toronto.

Detour, C., Thiesset, C. et Schuhl, P. (1995). Contrôle de qualité de l'enquête trimestrielle Emploi : résultats de l'enquête Transition sur le marché du travail. *Actes des Journées de Méthodologie Statistique*, Insee Méthodes 59-61, 415-480.

Duncan, G. J. and Kalton, G. (1987). Issues of Design and Analysis of Surveys Across Time. *International Statistical Review*, 5, 97-117.

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

Fuller, W. (1990). Analysis of Repeated Surveys. *Survey Methodology*, 2 , 167-180.

Gouriéroux, C. et Roy, G. (1978). Enquête en deux vagues : renouvellement de l'échantillon. *Annales de l'Insee*, 29, 115-135.

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 Research Methods*, 247-257.

Holt, D. and Skinner, C. J. (1989). Component of Change in Repeated Surveys. *International Statistical Review*, 57, 1-18.

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

Kish, L. (1998). Space/Time Variations and Rolling Samples. *Journal of Official Statistics*, 1, 31-46.

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

Lee, H. (1990). Estimation of Panel Correlations for the Canadian Labour Force Survey. *Survey Methodology*, 2, 283-292.

Rao, J. and Graham, J. (1964). Rotation designs for sampling on repeated occasions. *Journal of the American Statistical Association*, 50, 492-509.

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

Steel, D. (1996). Options for Producing Monthly Estimates of Unemployment According to the ILO Definition. *Mimeo.*

Yates, F. (1979). *Sampling Methods for Censuses and Surveys*. Fourth Edition. Charles Griffin, London.