

No 11-522-XIF au catalogue

**La série des symposiums internationaux
de Statistique Canada - Recueil**

**Symposium 2005 : Défis
méthodologiques reliés aux
besoins futurs d'information**



2005



Statistique
Canada

Statistics
Canada

Canada

LA COMMUNICATION DE VARIANCES SERVANT À INTERPRÉTER LES CHANGEMENTS ET LES POINTS D'INFLEXION LORS D'ENQUÊTES À PASSAGES RÉPÉTÉS

David Steel et Paul Smith¹

RÉSUMÉ

Les estimations obtenues lors d'enquêtes à passages répétés servent à suivre et à analyser les changements affectant les variables dignes d'intérêt au fil du temps. Les enquêtes majeures sont souvent menées sur une base mensuelle ou trimestrielle et permettent de cerner les changements importants dans le niveau ou le taux de variation des variables, y compris les points d'inflexion. Les échantillons utilisés par ces enquêtes sont souvent vastes et donnent lieu à des chevauchements au fil du temps à cause du recours à une procédure de renouvellement. La taille d'échantillon et la structure de corrélation dans les estimations induites par le schéma de renouvellement ont une incidence sur les erreurs d'échantillonnage des estimations et sur l'analyse des changements qui les touchent. Les utilisateurs analysent et interprètent les séries chronologiques d'estimations de différentes manières, en faisant souvent appel à des estimations portant sur plusieurs périodes. Malgré la taille importante des échantillons et le degré de chevauchement entre eux pour certaines périodes, les erreurs d'échantillonnage peuvent encore affecter substantiellement les estimations des mouvements et les fonctions connexes qui servent à interpréter la série d'estimations. Nous considérons comment tenir compte des erreurs d'échantillonnage dans l'interprétation des estimations provenant d'enquêtes à passages répétés et comment informer les utilisateurs et les analystes de leur incidence éventuelle.

MOTS CLÉS: Estimations de mouvements; Schémas de renouvellement; Tendances; Points d'inflexion.

1. INTRODUCTION

1.1 Enquêtes à passages répétés

Les estimations obtenues à partir d'enquêtes à passages répétés servent à suivre et à analyser les changements affectant les variables dignes d'intérêt au fil du temps. Les enquêtes sociales et économiques majeures sont souvent menées sur une base mensuelle ou trimestrielle et visent à cerner les changements importants dans le niveau ou le taux de variation des variables, y compris les points d'inflexion. Ces enquêtes utilisent souvent des échantillons vastes et des plans qui donnent lieu à des chevauchements au fil du temps à cause du recours à une procédure de renouvellement de l'échantillon. Le chevauchement de l'échantillon amène une structure de corrélation dans la série chronologique d'estimations, ce qui a une incidence sur les erreurs d'échantillonnage des estimations et sur l'analyse des changements qui les touchent.

Les utilisateurs analysent et interprètent les séries chronologiques d'estimations de différentes manières en faisant appel à des estimations portant sur plusieurs points chronologiques. Généralement, la corrélation de la variable au niveau unitaire pour toute paire de points chronologiques varie en fonction du décalage entre les points. Elle se dégrade habituellement lorsque nous considérons de plus gros décalages, bien que la saisonnalité puisse entrer en ligne de compte. La corrélation entre les estimations pour deux points dans le temps dépend également du chevauchement de l'échantillon, de même que de la corrélation au niveau unitaire. Aucun chevauchement entre deux périodes donne lieu à une corrélation nulle entre les estimations. Même s'il y a chevauchement de l'échantillon, la corrélation au niveau unitaire peut être faible et peut mener à une corrélation faible entre les estimations. Si on se retrouve avec un chevauchement important et une corrélation unitaire élevée, alors les

¹David Steel, Centre for Statistical and Survey Methodology, University of Wollongong, Australia, NSW 2522, (courriel : dsteel@uow.edu.au); Paul Smith, Methodology Directorate, Office for National Statistics, Newport, UK, NP10 8XG (courriel : Paul.Smith@ons.gsi.gov.uk);

estimations sont fortement corrélées. Ces différences de corrélation entre les estimations compliquent l'analyse et l'interprétation de l'estimation découlant d'une enquête à passages répétés.

Malgré la taille importante des échantillons et le degré de chevauchement entre eux pour certains points, les erreurs d'échantillonnage peuvent affecter substantiellement les estimations des mouvements et les fonctions connexes qui servent à interpréter la série d'estimations. Nous devons considérer comment tenir compte des erreurs d'échantillonnage dans l'interprétation des estimations provenant d'enquêtes à passages répétés et comment informer les utilisateurs et les analystes de leur incidence éventuelle.

Parmi les exemples d'enquêtes majeures à passages répétés, on retrouve les enquêtes sur la population active, les enquêtes sur le commerce de détail et diverses enquêtes portant sur d'autres indicateurs économiques.

2. PLAN ET ANALYSE DES ENQUÊTES À PASSAGES RÉPÉTÉS

2.1 Schémas de renouvellement

Lors des enquêtes à passages répétés, l'échantillon est souvent choisi de manière à ce qu'il y ait chevauchement substantiel entre les enquêtes successives. Cette approche réduit non seulement les variances d'échantillonnage des estimations entre les deux enquêtes, mais encore les coûts comparativement à la constitution d'un échantillon indépendant à chaque période. Les échantillons chevauchants sont produits en utilisant un schéma de renouvellement qui peut aussi être conçu afin de répartir le fardeau de réponse. Un schéma de renouvellement raisonnable consisterait à sélectionner une unité pour a mois consécutifs, à la retirer pour b mois, puis à la réintégrer pour une autre période de a mois. Ce processus est répété jusqu'à ce que les unités soient incluses pour un total de m mois. Ce schéma peut être exprimé par a - b - $a(m)$. En posant $b=0$, nous nous retrouvons avec un schéma de renouvellement prévoyant l'inclusion pour m mois. Voir Rao et Graham (1964).

Exemples

L'Enquête australienne mensuelle sur la population active (MLFS) utilise un schéma de renouvellement prévoyant l'inclusion pour 8 mois. Ce schéma mène à un chevauchement d'échantillon avec un décalage de s , $v_s = 1-s/8$ lorsque $s=1, \dots, 7$. Le chevauchement réel est légèrement moindre en raison du mouvement des gens et des ménages entre les enquêtes. Il n'y a pas de chevauchement d'échantillon pour un décalage de $s=12$. Ce schéma débouche sur un chevauchement important des échantillons lorsque le décalage est court, ce qui contribue à réduire les variances d'échantillonnage pour les estimations des variations à court terme (sur 1, 2 ou 3 mois, par exemple).

L'Enquête (mensuelle) sur la population active canadienne utilise un schéma de renouvellement prévoyant l'inclusion pour 6 mois. Ce schéma mène à un chevauchement d'échantillon théorique avec un décalage de s , $v_s = 1-s/6$ lorsque $s=1, \dots, 5$. Encore, il n'y a pas de chevauchement d'échantillon pour un décalage de $s=12$. Ce schéma débouche également sur un chevauchement important des échantillons lorsque le décalage est court, ce qui contribue à réduire les variances d'échantillonnage pour les estimations des variations sur 1, 2 ou 3 mois.

L'enquête sur la population actuelle des États-Unis utilise un schéma de renouvellement à 4 - 8 - $4(8)$. Le chevauchement d'échantillon qui en résulte est de $v_s = 1-s/4$ lorsque $s=1, \dots, 3$, $v_s = 0$ lorsque $s=4, \dots, 8$ et $v_s = 4/8 - |s-12|/8$ lorsque $s=9, \dots, 15$. Le chevauchement d'échantillon lorsque $s=12$ est $v_{12} = 4/8$. Le chevauchement d'échantillon diminue plutôt rapidement lorsque le décalage augmente, bien qu'il réapparaisse après 9 mois.

L'enquête sur la population active du Royaume-Uni est conçue comme une enquête trimestrielle, mais peut aussi être vue comme un échantillon mensuel avec un schéma de renouvellement 1 - 2 - $1(5)$. Le chevauchement d'échantillon est de $v_s = 1-s/(3m)$ lorsque $s=3, 6, 9, \dots, 3m$, sinon $v_s = 0$. Il y a un faible chevauchement d'échantillon de $v_{12} = 1/5$ lorsque $s=12$.

L'enquête australienne mensuelle sur le commerce de détail utilise un schéma de renouvellement pour les petites entreprises en vertu duquel le douzième des entreprises sont écartées de l'enquête aux trois mois, alors qu'aucun

renouvellement n'est effectué lors des deux autres mois du trimestre. Les petites entreprises ne se retrouvent donc pas dans l'échantillon pendant plus de 36 mois.

Ces exemples montrent qu'il existe de nombreux schémas de renouvellement en vigueur. Des schémas différents donnent lieu à des degrés de chevauchement d'échantillon différents. Ils génèrent des corrélations entre les estimations pour différentes périodes avec un chevauchement d'échantillon appréciable.

2.2 Résultats et analyses clés

Les enquêtes à passages répétés peuvent fournir des estimations pour chaque période \hat{y}_t . Étant donné les coûts engendrés par leur réalisation sur une base répétitive, la valeur réelle des enquêtes à passages répétés tient à leur capacité de fournir des estimations de changement, telles que :

$$\hat{y}_t - \hat{y}_{t-s} = \Delta^{(s)}\hat{y}_t.$$

L'accent est souvent mis sur $s=1$, mais pour une enquête répétée sur une base mensuelle, on examine souvent les changements lorsque $s=2, 3, 12$. Un chevauchement d'échantillon au décalage s mène généralement à une corrélation positive entre les estimations. Puisque

$$\text{Var}(\Delta^{(s)}\hat{y}_t) = \text{Var}(\hat{y}_t) + \text{Var}(\hat{y}_{t-s}) - 2\sqrt{\text{Var}(\hat{y}_t)}\sqrt{\text{Var}(\hat{y}_{t-s})}\text{Corr}(\hat{y}_t, \hat{y}_{t-s}) \quad (1)$$

cela réduit la variance de $\Delta^{(s)}\hat{y}_t$ par rapport à ce qu'on aurait obtenu sans chevauchement d'échantillon.

Il peut aussi y avoir un intérêt à connaître les changements dans le taux de changement, par exemple,

$$\Delta^{(s)}\hat{y}_t - \Delta^{(s)}\hat{y}_{t-k} = \hat{y}_t - \hat{y}_{t-s} - (\hat{y}_{t-k} - \hat{y}_{t-k-s}).$$

Si $s=k$, nous nous retrouvons avec

$$\Delta^{(s)}\hat{y}_t - \Delta^{(s)}\hat{y}_{t-s} = \hat{y}_t - 2\hat{y}_{t-s} + \hat{y}_{t-2s}.$$

Lorsqu'une enquête à passages répétés mensuelle ou trimestrielle a été menée durant plusieurs années, on peut produire et analyser une série chronologique. Des estimations désaisonnalisées sont souvent effectuées afin de faciliter l'interprétation de la série chronologique, donnant la série \hat{x}_t . La désaisonnalisation n'est pas un processus linéaire, mais des approximations linéaires sont disponibles. Afin d'évaluer le schéma de changement sous-jacent, on peut aussi produire des estimations de tendance, ce qui soulève la question suivante : qu'est-ce que la tendance ? Dans certains cas, on suppose qu'elle est $\Delta^{(s)}\hat{y}_t$ ou $\Delta^{(s)}\hat{x}_t$. Le Bureau australien de la statistique produit souvent des estimations de la tendance en utilisant les moyennes mobiles de Henderson appliquées à des séries désaisonnalisées (ABS, 1987). Il existe aussi d'autres possibilités.

D'après une opinion assez répandue, la première fois qu'une chose arrive, c'est une petite anomalie passagère; la deuxième, c'est une coïncidence; la troisième crée une tendance (*The Economist*, 1990). Cette affirmation reflète ce que font de nombreux utilisateurs, à savoir examiner $\Delta^{(s)}\hat{y}_t$ lorsque $s=1, 2, 3$.

Toutes ces approches de l'analyse et de l'interprétation sont des fonctions linéaires des séries chronologiques qui se terminent au temps t , $\hat{\mathbf{y}}_t$, $l^T \hat{\mathbf{y}}_t$, où l est un vecteur de coefficients. La variance d'échantillonnage associée est

$$\text{Var}(l^T \hat{\mathbf{y}}_t) = l^T \text{Var}(\hat{\mathbf{y}}_t) l.$$

Donc, en général, nous avons besoin d'une estimation de la matrice de covariance $\text{Var}(\hat{\mathbf{y}}_t)$, ce qui nécessite l'estimation des covariances (ou bien des variances des différences entre les estimations) pour un certain nombre d'estimations. Avec de telles estimations de variance, nous pouvons établir des intervalles de confiance et tester des hypothèses, qui prennent souvent la forme

$$l^T \hat{\mathbf{y}}_t > a.$$

Les schémas de renouvellement utilisés tendent à mettre l'accent sur la production d'un chevauchement d'échantillon important au décalage $s=1$ et, dans une moindre mesure, $s=2,3$ et 12 . Le schéma de renouvellement 1-2-1 (5) induit un chevauchement aux décalages 3, 6, 9, 12 et peut être utile dans certaines approches d'analyse, telles que l'examen des mouvements sur trois mois et l'utilisation de moyennes sur trois mois (voir Steel, 1997).

Lorsque nous nous penchons sur l'interprétation des changements récents, nous pouvons observer des quantités comme

$$\text{Prob}(l^T \hat{y}_t > 0 | l^T Y_t) \text{ et } \text{Prob}(l^T \hat{y}_t \text{ sig} | l^T Y_t).$$

2.3 Évaluer la puissance pour détecter les changements

Alors que les enquêtes à passages répétés se caractérisent souvent par des tailles d'échantillon importantes et un degré élevé de chevauchement entre les périodes consécutives, elles cherchent fréquemment à estimer des changements relativement faibles. Par exemple, l'écart-type des estimations du changement mensuel du taux de chômage dans la MLFS australienne est d'environ 0,1 point, même avec le chevauchement d'échantillon mensuel élevé (sept huitième) et une taille d'échantillon de plus de 30 000 ménages. Les changements sur un mois du taux de chômage sont souvent à peu près de cette ampleur et ne sont donc pas statistiquement significatifs. La valeur de l'enquête dépend de la rapidité avec laquelle elle va fournir la preuve d'un changement important.

Lors d'une évaluation des options pour la production des estimations mensuelles de la population active de Grande-Bretagne, Steel (1996, 1997) a examiné les écarts-types et la puissance associés aux différentes options. Il a considéré une enquête fondée sur un échantillon de 60 000 ménages, un schéma de renouvellement prévoyant l'inclusion pour 6 mois ou 1-2-1(5), ainsi qu'une analyse basée sur les estimations mensuelles et les moyennes trimestrielles. Le tableau 1 présente les écarts-types pour les estimations de changements pour différents décalages. À l'époque, le niveau de chômage était d'environ 2,4 m et les changements mensuels moyens étaient d'à peu près 30 000. La première colonne de chiffres montre l'incidence du chevauchement d'échantillon élevé sur la réduction des écarts-types pour les changements sur un mois, et l'augmentation de l'écart-type lorsque les changements sur une plus longue période sont étudiés, à cause du chevauchement d'échantillon moindre et du déclin de la corrélation au niveau unitaire. La deuxième colonne de chiffres montre l'effet du chevauchement d'échantillon qui se réalise seulement à des multiples de trois mois. Si l'analyse se concentre sur les changements sur trois mois, alors cette option génère des écarts-types plus faibles. Les deux dernières colonnes présentent les résultats correspondants lorsque l'analyse est basée sur les moyennes trimestrielles, que certains utilisateurs adoptent pour gérer approximativement la volatilité des estimations mensuelles. Ici, l'option 1-2-1(5) est supérieure à l'option prévoyant l'inclusion pour 6 mois parce qu'elle nécessite l'établissement d'une moyenne à partir d'estimations non corrélées, alors que l'option prévoyant l'inclusion pour 6 mois requiert l'établissement d'une moyenne à partir d'estimations positivement corrélées.

Tableau 1: Écarts-types (en milliers) des changements sur 6 mois, estimations mensuelles et moyennes trimestrielles, $n=60\ 000$, chômage

s	Estimations mensuelles		Moyennes trimestrielles	
	<i>Inclusion pour 6 mois</i>	1-2-1(5)	<i>Inclusion pour 6 mois</i>	1-2-1(5)
1	28	42	12	11
2	33	42	20	15
3	36	33	24	19
4	39	42	30	20
5	41	42	32	20
6	42	37	34	21
7	42	42	34	22

Steel (1996) a aussi considéré la question de la puissance qu'aurait chacune des options. Il a envisagé une situation hypothétique où le chômage augmentait de manière constante au rythme de 30 000 par mois jusqu'à avril, pour ensuite décliner au même taux. Le tableau 2 donne la probabilité d'obtenir une estimation du bon signe et aussi un changement statistiquement significatif, en utilisant un test unilatéral à 5 % pour observer différents décalages à chacun des mois suivant avril. Par exemple, si nous étudions le changement entre mai et avril, le schéma de

renouvellement prévoyant l'inclusion pour 6 mois a une probabilité de 86 % d'indiquer un changement du bon signe, mais seulement une probabilité de 26 % de relever un changement statistiquement significatif. En juillet, c'est-à-dire sur 3 mois, les probabilités augmentent respectivement à 100 % et 80 %. Ces résultats sont cohérents avec l'opinion des analystes exprimée dans la section 2.2 : on ne saurait être raisonnablement confiant qu'après trois changements. En observant les résultats pour le schéma de renouvellement 1-2-1(5), nous constatons que les écarts-types plus grands pour les décalages 1 et 2 engendrent des probabilités plus faibles, mais que les écarts-types moindres pour les résultats du décalage 3 résultent en des probabilités plus élevées à partir de juin.

Tableau 2: Puissance de l'analyse pour des différences sur s mois – point d'inflexion en avril, changement mensuel de 30 000, chômage

Renouvellement	<i>Inclusion pour 6 mois</i>			1-2-1(5)		
	1	2	3	1	2	3
Décalage s						
Mois de l'analyse	Prob (%) que le changement estimé ait le bon signe					
Mai	86	50	21	76	50	18
Juin		97	80		92	82
Juillet			100			100
Mois de l'analyse	Prob (%) que le changement estimé soit stat. signif. (test unilatéral 5 %)					
Mai	26	5	1	18	5	1
Juin		57	21		41	24
Juillet			80			87

La production de tableaux comme le tableau 2 selon différentes situations pourrait aider les analystes à interpréter la capacité de l'enquête à les informer de changements importants.

3. PRÉSENTATION GRAPHIQUE DES INTERVALLES DE CONFIANCE DE CHANGEMENTS IMPORTANTS

3.1 Inférences concernant les changements

En théorie, nous pouvons calculer une estimation de la variance pour toute fonction linéaire et la présenter, par exemple, sous forme de tableau. Cependant, des problèmes d'inférences simultanées surviennent. Il serait utile d'utiliser des méthodes graphiques simples pour expliquer aux utilisateurs et aux analystes comment interpréter les changements dans la série d'estimations.

Supposons qu'on insiste principalement sur $\Delta^{(s)}\hat{y}_t$. Un intervalle de confiance approximatif à 95 % pour \hat{y}_t peut être construit sous la forme $\hat{y}_t - 2\sqrt{\hat{Var}(\hat{y}_t)}$ à $\hat{y}_t + 2\sqrt{\hat{Var}(\hat{y}_t)}$, où $\hat{Var}(\hat{y}_t)$ est une estimation de la variance de la prise en compte du plan d'échantillonnage. Un intervalle de confiance similaire peut être construit pour \hat{y}_{t-s} .

Il est valable de représenter graphiquement ces intervalles pour chaque point chronologique afin d'indiquer l'intervalle de confiance de chaque estimation. Par contre, lorsque les utilisateurs déterminent si ces intervalles de confiance se chevauchent ou non en vue d'estimer la signification statistique du changement, ils sont alors induits en erreur, puisqu'on suppose un intervalle de confiance de

$$\hat{y}_t - \hat{y}_{t-s} - 2\left(\sqrt{\hat{Var}(\hat{y}_t)} + \sqrt{\hat{Var}(\hat{y}_{t-s})}\right) \text{ à } \hat{y}_t - \hat{y}_{t-s} + 2\left(\sqrt{\hat{Var}(\hat{y}_t)} + \sqrt{\hat{Var}(\hat{y}_{t-s})}\right)$$

plutôt qu'un intervalle approprié de

$$\hat{y}_t - \hat{y}_{t-s} - 2\left(\sqrt{\widehat{Var}(\Delta^{(s)}\hat{y}_t)}\right) \text{ à } \hat{y}_t - \hat{y}_{t-s} + 2\left(\sqrt{\widehat{Var}(\Delta^{(s)}\hat{y}_t)}\right),$$

où $Var(\Delta^{(s)}\hat{y}_t)$ est donné par (1).

Supposons que $Var(\hat{y}_t) \approx Var(\hat{y}_{t-s})$. Si nous mettons les intervalles de confiance originaux à échelle par rapport aux estimations de niveaux à l'aide d'un facteur de $\sqrt{\frac{1 - Corr(\hat{y}_t, \hat{y}_{t-s})}{2}}$, ce qui revient à utiliser

$$\hat{y}_t - \sqrt{2(1 - Corr(\hat{y}_t, \hat{y}_{t-s}))}\sqrt{\widehat{Var}(\hat{y}_t)} \text{ à } \hat{y}_t + \sqrt{2(1 - Corr(\hat{y}_t, \hat{y}_{t-s}))}\sqrt{\widehat{Var}(\hat{y}_t)},$$

alors l'intervalle de confiance pour $\hat{y}_t - \hat{y}_{t-s}$, qui était sous-entendu en observant si les intervalles se chevauchent ou non, est approprié.

Ces facteurs sont donnés dans le tableau 3, d'après les modèles de corrélation mis au point par Steel (1996) pour les variables de chômage et d'emploi. Le facteur de 0,71 correspond au cas où les estimations ne sont pas corrélées. Ces facteurs sont inférieurs à 1 et montrent qu'il est nécessaire d'effectuer l'ajustement. Pour chaque variable, le facteur dépend du décalage à l'étude. Il est plus direct de présenter les intervalles ajustés pour un décalage clé, par exemple $s=1$. Il faudrait développer des méthodes simples permettant de présenter simultanément ces intervalles ajustés pour plusieurs décalages afin d'aider les utilisateurs à interpréter les estimations de manière appropriée.

Tableau 3: Facteurs servant à ajuster les intervalles de confiance pour analyser les mouvements sous différents décalages, série originale

s	Chômage		Emploi	
	<i>Inclusion pour 6 mois</i>	1-2-1(5)	<i>Inclusion pour 6 mois</i>	1-2-1(5)
1	0,47	0,71	0,38	0,71
2	0,56	0,71	0,47	0,71
3	0,61	0,54	0,54	0,42
4	0,66	0,71	0,60	0,71
5	0,69	0,71	0,66	0,71
6	0,71	0,62	0,71	0,52
7	0,71	0,71	0,71	0,71

RÉFÉRENCES

- Australian Bureau of Statistics (1993), "A Guide to Interpreting Time Series - Monitoring "Trends", An Overview", catalogue no 1348.0, Canberra, Australia: Australian Bureau of Statistics.
- McLaren, C. H. et Steel, D. G. (2000), "L'effet de divers plans de renouvellement sur la variance d'échantillonnage des estimations désaisonnalisées et des estimations de la tendance", *Techniques d'enquête*, 26, pp. 185-195.
- Rao, J. N. K. et Graham, J. E. (1964), "Rotation Designs for Sampling on Repeated Occasions", *Journal of the American Statistical Association*, 69, pp. 492-509.
- Steel, D. (1997), "Producing Monthly Estimated of Unemployment and Employment According to the International Labour Office Definition", *Journal of the Royal Statistical Society*, 160, pp. 5-33.
- Steel, D. G. (1996) "Options for Producing Monthly Estimates of Unemployment According to the ILO Definition", rapport non publié, London, UK: Central Statistics Office.
- The Economist (1990), "Spotting a Trend", *The Economist*, édition du 17 mars, London, UK.