Composante du produit nº 12-001-X
Division des méthodes d'enquêtes auprès des entreprises

Article

Combiner des cohortes dans les enquêtes longitudinales

par Iván A. Carrillo et Alan F. Karr

Juin 2013

Statistique Canada Statistics Canada

Canada

Comment obtenir d'autres renseignements

Pour toute demande de renseignements au sujet de ce produit ou sur l'ensemble des données et des services de Statistique Canada, visiter notre site Web à www.statcan.gc.ca.

Vous pouvez également communiquer avec nous par :

Courriel à infostats@statcan.gc.ca

Téléphone entre 8 h 30 et 16 h 30 du lundi au vendredi aux numéros sans frais suivants :

• Service de renseignements statistiques	1-800-263-1136
• Service national d'appareils de télécommunications pour les malentendants	1-800-363-7629
• Télécopieur	1-877-287-4369

Programme des services de dépôt

Service de renseignements	1-800-635-7943
Télécopieur	1-800-565-7757

Comment accéder à ce produit

Le produit n° 12-001-X au catalogue est disponible gratuitement sous format électronique. Pour obtenir un exemplaire, il suffit de visiter notre site Web à www.statcan.gc.ca et de parcourir par « Ressource clé » > « Publications ».

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois. À cet égard, notre organisme s'est doté de normes de service à la clientèle que les employés observent. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1-800-263-1136. Les normes de service sont aussi publiées sur le site www.statcan.gc.ca sous « À propos de nous » > « Notre organisme » > « Offrir des services aux Canadiens ».

Publication autorisée par le ministre responsable de Statistique Canada

This publication is also available in English.

Note de reconnaissance

Le succès du système statistique du Canada repose sur un partenariat bien établi entre Statistique Canada et la population du Canada, ses entreprises, ses administrations et les autres établissements. Sans cette collaboration et cette bonne volonté, il serait impossible de produire des statistiques exactes et actuelles.

Signes conventionnels

Les signes conventionnels suivants sont employés dans les publications de Statistique Canada :

.	indisponible pour toute période de référence
..	indisponible pour une période de référence précise
...	n'ayant pas lieu de figurer
0	zéro absolu ou valeur arrondie à zéro
0^{s}	valeur arrondie à 0 (zéro) là où il y a une distinction importante entre le zéro absolu et la valeur arrondie
^{p}	provisoire
^{r}	révisé
x	confidentiel en vertu des dispositions de la *Loi sur la statistique*
^{E}	à utiliser avec prudence
F	trop peu fiable pour être publié
*	valeur significativement différente de l'estimation pour la catégorie de référence (p<0,05)

Techniques d'enquête, juin 2013
Vol. 39, N° 1, pp. 169-206
Statistique Canada, N° 12-001-X au Catalogue

Combiner des cohortes dans les enquêtes longitudinales

Iván A. Carrillo et Alan F. Karr[1]

Résumé

Une question fréquente concernant les enquêtes longitudinales est celle de savoir comment combiner les différentes cohortes. Dans le présent article, nous présentons une nouvelle méthode qui permet de combiner différentes cohortes et d'utiliser toutes les données à notre disposition dans une enquête longitudinale pour estimer les paramètres d'un modèle semi-paramétrique qui relie la variable réponse à un jeu de covariables. La procédure s'appuie sur la méthode des équations d'estimation généralisées pondérées pour traiter les données manquantes pour certaines vagues dans les enquêtes longitudinales. Notre méthode s'appuie, pour l'estimation des paramètres du modèle, sur un cadre de randomisation conjointe qui tient compte à la fois du modèle de superpopulation et de la sélection aléatoire selon le plan de sondage. Nous proposons aussi une méthode d'estimation de la variance sous le plan et sous randomisation conjointe. Pour illustrer la méthode, nous l'appliquons à l'enquête Survey of Doctorate Recipients réalisée par la National Science Foundation des États-Unis.

Mots clés : Paramètres de superpopulation; inférence sous randomisation conjointe; estimation de variance par rééchantillonnage; enquêtes à panel rotatif; enquête longitudinale à cohortes multiples; équations d'estimation généralisées pondérées.

1 Introduction

L'enquête Survey of Doctorate Recipients (SDR) est une enquête longitudinale menée par la National Science Foundation (NSF) dont le plan de sondage comprend des caractéristiques des panels répétés ainsi que des panels rotatifs. L'objectif de l'enquête est d'étudier les titulaires d'un doctorat en sciences, en génie ou en sciences de la santé aux États-Unis. Elle est réalisée tous les deux ans environ. Une description détaillée de la SDR peut être consultée dans NSF (2012). Dans le présent article, nous nous concentrons sur les données recueillies de 1995 à 2008 (7 vagues).

Une nouvelle cohorte est sélectionnée à l'occasion de chacune des vagues de l'enquête. La nouvelle cohorte est composée d'un échantillon de personnes récemment diplômées (au cours des deux années précédentes) tiré du Doctorate Records File, une base de données construite principalement d'après les données de l'enquête Survey of Earned Doctorates

1. Iván A. Carrillo et Alan F. Karr, National Institute of Statistical Sciences, 19 T.W. Alexander Drive, Research Triangle Park, NC 27709, États-Unis. Courriel : ivan@niss.org et karr@niss.org.

(http://www.nsf.gov/statistics/srvydoctorates/). Les personnes sélectionnées sont gardées dans l'échantillon, c'est-à-dire interviewées tous les deux ans, jusqu'à l'âge de 75 ans, à condition qu'elles vivent aux États-Unis durant la semaine de référence de l'enquête et qu'elles ne soient pas placées en établissement. Cependant, les diplômés échantillonnés qui satisfont à ces critères ne sont pas *tous* gardés indéfiniment. Certaines personnes, plutôt que des cohortes entières, sont supprimées de l'échantillon afin a) d'inclure les nouveaux diplômés dans de nouvelles cohortes et b) de maintenir une taille d'échantillon relativement constante d'une vague à l'autre. À la section 2.2, nous décrivons comment sont sélectionnées les personnes qui sont supprimées de l'échantillon.

Des poids de sondage existent déjà pour les analyses transversales des données de la SDR, mais non pour les analyses longitudinales. Au lieu de nécessiter un *nouveau* poids longitudinal pour *toutes* les données, la méthode proposée ici permet d'utiliser les poids transversaux existants pour les analyses longitudinales en n'ignorant aucune donnée. Nous nous concentrons sur l'estimation des paramètres des modèles statistiques de l'effet des covariables sur une réponse d'intérêt, mais la méthode peut également être utilisée pour estimer des quantités de population finie (Carrillo et Karr 2012). Notre analyse est axée sur la SDR, mais notre méthode est applicable à n'importe quelle enquête à panel fixe, à panel fixe plus « nouvelles unités », à panel répété, à panel rotatif, à panel divisé ou à renouvellement de l'échantillon, à condition qu'il existe pour chaque vague un poids transversal pour représenter la population d'intérêt de la vague en question. Voir Smith, Lynn et Elliot (2009), Hirano, Imbens, Ridder et Rubin (2001), et Nevo (2003) pour les définitions de tous ces types de plans de sondage longitudinaux.

La SDR est un hybride de plan à panel répété et de plan à panel rotatif. Il ne s'agit pas purement d'un plan à panel répété, à cause de la suppression de certains sujets à chaque vague. Il ne s'agit pas purement d'un plan à panel rotatif, parce que l'on supprime des personnes et *non* des panels (ou cohortes) complets; en outre, la composition de la population finie d'intérêt évolue au cours du temps, ce qui n'est pas le cas d'une enquête à panel rotatif.

Diggle, Heagerty, Liang et Zeger (2002) et Hedeker et Gibbons (2006) font remarquer que dans les enquêtes longitudinales, contrairement aux études transversales, il est possible d'isoler l'effet de l'âge (changement réel chez les sujets au cours du temps) et l'effet de cohorte (différences entre les unités au début de la période d'étude).

Hedeker et Gibbons (2006) laissent aussi entendre que, puisque les études longitudinales permettent de mesurer des variables explicatives (covariables) variant au cours du temps, les inférences statistiques au sujet de la relation dynamique entre le résultat d'intérêt (réponse) et ses covariables sont beaucoup plus puissantes que celles fondées sur des études transversales.

Si nous nous intéressons à la moyenne marginale d'une variable, éventuellement en conditionnant sur certaines covariables, plutôt qu'à la mesure du changement, une étude longitudinale n'est pas nécessaire; une étude transversale suffit. Cependant, même dans ce cas, une étude longitudinale a tendance à être plus puissante, parce que chaque sujet peut être utilisé comme son propre contrôle pour toute caractéristique non mesurée (Diggle et coll. 2002).

Notre approche diffère des solutions décrites dans la littérature, lesquelles présentent certaines limites pour l'analyse de ce genre de données, et en particulier pour l'application à la SDR. Par exemple, dans Berger (2004a) et Berger (2004b), l'estimation du changement est examinée en détail en utilisant des échantillons rotatifs, mais en posant que la composition de la population finie ne varie pas au cours du temps, ce qui n'est pas le cas de la SDR. Cette hypothèse ne tient pas non plus dans de nombreuses autres enquêtes à grande échelle. En outre, la méthodologie proposée par Berger n'est pas facilement généralisable à plus de deux vagues d'enquête. Similairement, Qualité et Tillé (2008) supposent que la population finie est fixe au cours du temps. Hirano et coll. (2001) et Nevo (2003) présentent diverses méthodes d'estimation en supposant que le plan est à panel fixe avec rafraîchissement pour tenir compte de l'attrition, mais émettent aussi l'hypothèse que la composition de la population finie est fixe au cours du temps.

McLaren et Steel (2000), et Steel et McLaren (2007) utilisent une approche fondée sur des séries chronologiques pour estimer la variation et la tendance dans les données d'enquête. Bien

que leur approche permette d'intégrer l'association intrasujet dans les estimations ponctuelles, ils ne considèrent pas de covariables dans leur modèle (sauf les covariables temporelles implicites). En outre, ils discutent uniquement de l'estimation du changement pour les variables continues.

Une autre option pour analyser des données longitudinales consiste à considérer la population finie d'intérêt comme étant fixe, sauf peut-être en ce qui concerne les décès, qui pourraient être permis. Les études de ce genre sont celles pour lesquelles des données n'existent que pour une seule cohorte. Par exemple, Vieira et Skinner (2008), Carrillo, Chen et Wu (2010), et Carrillo, Chen et Wu (2011) illustrent certaines options de modélisation en se basant sur des données d'enquête recueillies auprès d'une seule cohorte. Cependant, pour procéder à ce genre d'analyse sur les données d'enquête à plusieurs cohortes, on doit ignorer certaines (ou de nombreuses) données existantes, par exemple celles recueillies auprès des sujets qui ne sont pas présents à toutes les vagues. Un exemple de procédure de pondération de ce genre est décrit dans Ardilly et Lavallée (2007).

Enfin, l'approche de Larsen, Qing, Zhou et Foulkes (2011) est séduisante en principe, parce qu'il s'agit de la façon dont procèdent généralement les praticiens des sondages. Un poids initial est ajusté, entre autres par calage sur des totaux connus, ici sur des totaux par vague d'enquête. Néanmoins, pour les panels rotatifs, cette méthode en est encore à ses balbutiements; la manière d'exécuter certains de ses éléments n'est pas entièrement claire. Ainsi, le choix du poids initial n'est pas évident : un poids constant ?, le premier poids disponible ?, la moyenne des poids disponibles pour chaque cas ?, ou le dernier poids disponible ? En outre, en cas de décrochages, comme il en existe dans la SDR, les auteurs ne précisent pas comment procéder à un ajustement pour la non-réponse. De surcroît, on s'explique mal pourquoi un ajustement pour la non-réponse des décrocheurs, disons, à la vague 4 devrait avoir une influence sur les observations à la vague 3, comme le permet cette méthode, puisqu'elle comporte un poids unique pour chaque sujet. De plus, les auteurs mentionnent qu'ils ont estimé les erreurs-types, mais ils n'indiquent pas comment tenir compte de toutes les caractéristiques du plan de sondage, telles que les

modifications apportées au cours du temps à la stratification et aux classes pour l'ajustement des pondérations de la SDR. En revanche, notre méthode utilise uniquement des pondérations et des méthodes d'estimation de variance transversales, qui ont été étudiées en profondeur dans la documentation et auxquelles on a facilement accès pour la SDR.

La présentation du reste de l'article est la suivante. À la section 2, nous décrivons le plan de sondage de la SDR. À la section 3, nous proposons une nouvelle approche pour l'analyse longitudinale des modèles de moyenne marginale dans le cas d'enquêtes à plusieurs cohortes. À la section 4, nous présentons l'application de la méthode à la SDR. Enfin, à la section 5, nous offrons quelques points de discussion.

2 Le plan de sondage de la SDR

2.1 Population finie

La population finie d'intérêt de la SDR peut être représentée comme au tableau 2.1. À la vague 1, c'est-à-dire la première période d'intérêt, il existe un ensemble fini, $U_{1(1)} = U_1$, de $N_{1(1)} = N_1$ titulaires d'un doctorat, obtenu récemment ou non, qui satisfont aux exigences de la SDR.

Tableau 2.1
Population finie de la SDR

j :	**1**		**2**		**3**		**…**		$J-1$		J
	$U_{1(1)}$	$\supseteq$	$U_{2(1)}$	$\supseteq$	$U_{3(1)}$	$\supseteq$	$\cdots$	$\supseteq$	$U_{J-1(1)}$	$\supseteq$	$U_{J(1)}$
	$N_{1(1)}$	$\geq$	$N_{2(1)}$	$\geq$	$N_{3(1)}$	$\geq$	$\cdots$	$\geq$	$N_{J-1(1)}$	$\geq$	$N_{J(1)}$
			$U_{2(2)}$	$\supseteq$	$U_{3(2)}$	$\supseteq$	$\cdots$	$\supseteq$	$U_{J-1(2)}$	$\supseteq$	$U_{J(2)}$
			$N_{2(2)}$	$\geq$	$N_{3(2)}$	$\geq$	$\cdots$	$\geq$	$N_{J-1(2)}$	$\geq$	$N_{J(2)}$
						$\ddots$			$\vdots$		$\vdots$
									$U_{J-1(J-1)}$	$\supseteq$	$U_{J(J-1)}$
									$N_{J-1(J-1)}$	$\geq$	$N_{J(J-1)}$
											$U_{J(J)}$
											$N_{J(J)}$
	U_1		U_2		U_3		$\cdots$		U_{J-1}		U_J
	N_1		N_2		N_3		$\cdots$		N_{J-1}		N_J

À la vague 2, un sous-ensemble seulement des sujets compris dans $U_{1(1)}$ satisfont encore aux exigences de la SDR; nous appelons ce sous-ensemble de $N_{2(1)}$ sujets, $U_{2(1)}$. En outre, il existe un ensemble de nouveaux titulaires d'un doctorat, qui ont obtenu leur diplôme depuis la vague 1, et qui satisfont aussi aux autres exigences de l'enquête. Cet ensemble de nouveaux diplômés dans le champ de l'enquête est appelé $U_{2(2)}$ et est de taille $N_{2(2)}$. Par conséquent, à la vague 2, il y a un total de $N_2 = N_{2(1)} + N_{2(2)}$ sujets dans la population d'intérêt $U_2 = U_{2(1)} \cup U_{2(2)}$.

À la vague suivante, la vague 3, le même processus a lieu. Certains sujets compris dans $U_{2(1)}$ ont quitté la population d'intérêt et il n'en reste que $N_{3(1)}$ dans $U_{3(1)}$. La même chose se produit avec l'ensemble $U_{2(2)}$; seulement un sous-ensemble $U_{3(2)}$ de $N_{3(2)}$ sujets satisfera encore aux exigences de la SDR. En outre, $N_{3(3)}$ diplômés récents entrent dans la population d'intérêt; cet ensemble est appelé $U_{3(3)}$. En tout, la population finie d'intérêt à la vague 3 est $U_3 = U_{3(1)} \cup U_{3(2)} \cup U_{3(3)}$, avec $N_3 = N_{3(1)} + N_{3(2)} + N_{3(3)}$ sujets.

Cette procédure de réduction des cohortes anciennes et d'ajout de nouvelles cohortes se poursuit jusqu'à la dernière vague d'intérêt, la vague J. Nous constatons que la population finie d'intérêt change à chaque vague, principalement pour deux raisons. Premièrement, certains sujets appartenant aux anciennes cohortes ne sont plus dans le champ de la vague courante et ne font pas partie de la population cible courante. Deuxièmement, des diplômés récents sont ajoutés à la population cible de la vague courante. Nous désignons par $j = 1, 2, \ldots, J$ la vague d'intérêt (hors des parenthèses) et par $j' = 1, 2, \ldots, J$ la cohorte à laquelle le sujet appartient (entre parenthèses), et par conséquent $U_{j(j')} = U_{\text{vague(cohorte)}}$.

2.2 Échantillonnage

Le plan d'échantillonnage de la SDR possède une structure similaire à celle de la population finie et est illustré au tableau 2.2. À la vague 1, un échantillon (complexe) $s_{1(1)} = s_1$ de $n_{1(1)} = n_1$ sujets est sélectionné parmi les N_1 éléments de U_1. Chaque élément i dans s_1 est interviewé et les données qu'il fournit sont recueillies; en outre, il existe un poids de sondage

$w_{i1} = 1 / \pi_{i1}$ associé à l'élément, qui est l'inverse de la probabilité d'inclusion de ce dernier dans l'échantillon à la vague 1.

Tableau 2.2
Échantillon de la SDR

j :	**1**		**2**		**3**		$\cdots$		$J-1$		J
	$s_{1(1)}$	$\supseteq$	$s_{2(1)}$	$\supseteq$	$s_{3(1)}$	$\supseteq$	$\cdots$	$\supseteq$	$s_{J-1(1)}$	$\supseteq$	$s_{J(1)}$
	$n_{1(1)}$	$\geq$	$n_{2(1)}$	$\geq$	$n_{3(1)}$	$\geq$	$\cdots$	$\geq$	$n_{J-1(1)}$	$\geq$	$n_{J(1)}$
			$s_{2(2)}$	$\supseteq$	$s_{3(2)}$	$\supseteq$	$\cdots$	$\supseteq$	$s_{J-1(2)}$	$\supseteq$	$s_{J(2)}$
			$n_{2(2)}$	$\geq$	$n_{3(2)}$	$\geq$	$\cdots$	$\geq$	$n_{J-1(2)}$	$\geq$	$n_{J(2)}$
					$s_{3(3)}$	$\supseteq$	$\cdots$	$\supseteq$	$s_{J-1(3)}$	$\supseteq$	$s_{J(3)}$
					$n_{3(3)}$	$\geq$	$\cdots$	$\geq$	$n_{J-1(3)}$	$\geq$	$n_{J(3)}$
							$\ddots$		$\vdots$		$\vdots$
									$s_{J-1(J-1)}$	$\supseteq$	$s_{J(J-1)}$
									$n_{J-1(J-1)}$	$\geq$	$n_{J(J-1)}$
											$s_{J(J)}$
											$n_{J(J)}$
	s_1		s_2		s_3		$\cdots$		s_{J-1}		s_J
	n_1		n_2		n_3		$\cdots$		n_{J-1}		n_J

À la deuxième vague, les éléments compris dans $s_{1(1)}$ qui ne sont plus dans le champ de l'enquête sont simplement supprimés de la base de sondage (mais les observations les concernant faites à la vague 1 sont gardées), et un sous-échantillon $s_{2(1)}$, de taille $n_{2(1)}$, des sujets encore dans le champ de l'enquête est sélectionné. Les membres de $s_{1(1)}$ qui sont encore dans le champ de l'enquête à la vague 2 ne sont pas tous gardés dans l'échantillon, et ce pour permettre d'ajouter l'échantillon de nouveaux titulaires d'un doctorat tout en maintenant plus ou moins la même taille d'échantillon qu'à la vague 1. Un échantillon $s_{2(2)}$ de taille $n_{2(2)}$ est tiré de $U_{2(2)}$; les sujets compris dans $s_{2(2)}$ forment la deuxième cohorte. L'échantillon total à la vague 2 est $s_2 = s_{2(1)} \cup s_{2(2)}$, dont la taille est $n_2 = n_{2(1)} + n_{2(2)}$, qui est approximativement égale à n_1. Tous les sujets compris dans s_2 sont interviewés à la vague 2. Les poids de sondage à la vague 2, $w_{i2} = 1 / \pi_{i2}$, sont tels que l'échantillon s_2 représente la population d'intérêt à la vague 2, à savoir U_2.

La même procédure est répétée à chaque vague jusqu'à la dernière (J), où un sous-échantillon des sujets restants en provenance de chacune des $J-1$ cohortes antérieures est sélectionné, et un nouvel échantillon (nouvelle cohorte) $s_{J(J)}$ de diplômés récents est tiré de $U_{J(J)}$. À la dernière vague, tous les sujets compris dans $s_J = \bigcup_{j'=1}^{J} s_{J(j')}$ sont interviewés et un poids de sondage $w_{iJ} = 1/\pi_{iJ}$ est créé pour chacun, de sorte que s_J représente la population finie U_J.

En ce qui concerne la façon dont sont sélectionnés les sujets supprimées de l'échantillon, selon NSF (2012), en 2008 par exemple, le sous-échantillon $s_{08} \setminus s_{08(08)}$ a été sélectionné en stratifiant s_{06} « en 150 strates en fonction de 3 variables : groupe démographique, domaine du diplôme et sexe. » Le rapport explique aussi que :

- l'ancienne pratique consistant à sélectionner l'échantillon avec probabilité proportionnelle à la taille s'est poursuivie, la mesure de taille étant le poids de base associé au cycle précédent de l'enquête. Pour chaque strate, l'algorithme d'échantillonnage a commencé par repérer et supprimer les cas autoreprésentatifs selon une procédure itérative. Ensuite, dans chaque strate, les cas non autoreprésentatifs ont été triés en fonction de la citoyenneté, de l'état d'incapacité, du domaine du diplôme et de l'année d'obtention du diplôme de doctorat. Enfin, le solde de l'échantillon (c'est-à-dire le total attribué à la strate moins le nombre de cas autoreprésentatifs) a été sélectionné dans chaque strate systématiquement avec probabilité proportionnelle à la taille.

Il convient de mentionner que, jusqu'à 1989, la cohorte (ou plus précisément l'année d'obtention du diplôme) faisait partie des variables de stratification (et des cellules d'ajustement des poids), mais qu'à partir de 1991, elle ne l'a plus été; elle a été remplacée par l'état d'incapacité. Pour des renseignements plus détaillés sur la procédure de sous-échantillonnage, y compris la description de la répartition de l'échantillon, voir NSF (2012) ou Cox, Grigorian, Wang et Harter (2010).

La description qui précède montre clairement que la SDR n'est pas réalisée selon un plan à panel rotatif. Outre le fait que la composition de la population finie d'intérêt évolue avec le

temps, un plan à panel rotatif donnerait lieu à la sélection, au temps j, d'une nouvelle cohorte à partir de U_j, et non à partir de $U_j \setminus U_{j-1}$ comme cela est le cas dans la SDR.

Une autre particularité de la SDR est qu'à chaque vague j, une base de sondage de diplômés récents $U_{j(j)}$ existe, de laquelle peut être tirée directement la nouvelle cohorte $s_{j(j)}$. Cependant, dans d'autres applications, le coût de la création d'une telle base de sondage, par exemple une liste de nouveaux membres, peut être excessif (particulièrement lorsqu'il est cumulé sur l'ensemble des vagues), et la nouvelle cohorte doit parfois être sélectionnée à partir de U_j (par opposition à $U_{j(j)}$). La méthode proposée dans le présent article peut également être appliquée à ce genre de cas, à condition que l'on puisse créer pour l'échantillon total à la vague j, s_j, un poids transversal pour représenter U_j. Nous discutons de cet aspect plus en détail à la section 3.2.

Soulignons que, dans la notation $s_{j(j')}$, la quantité j représente la vague à laquelle l'échantillon se rapporte, et j' désigne la cohorte de l'échantillon, c'est-à-dire la vague à laquelle l'échantillon a été sélectionné initialement. La notation pour les pondérations est w_{ij}, où le premier indice inférieur désigne le sujet et le second, la vague d'intérêt, quelle que soit la période où le sujet a été sélectionné initialement.

3 Méthodologie

3.1 Motivation

Supposons que (hors du contexte d'une enquête) l'on s'intéresse au paramètre vectoriel $\boldsymbol{\beta}$ de dimension $p \times 1$ dans le modèle suivant :

$$\xi : \begin{cases} E[Y_{ij} \mid X_{ij}] = \mu_{ij} = g^{-1}(X'_{ij}\boldsymbol{\beta}), & j = 1,2,\ldots,J, i = 1,2,\ldots \\ \mathrm{Var}[Y_{ij} \mid X_{ij}] = \phi\nu(\mu_{ij}), & j = 1,2,\ldots,J, i = 1,2,\ldots \\ \mathrm{Cov}[Y_i \mid X_i] = \Sigma_i, & i = 1,2,\ldots \\ Y_k \perp Y_l \mid X_k, X_l, & k \neq l = 1,2,\ldots; \end{cases} \tag{3.1}$$

où Y_{ij} est la variable réponse pour le sujet i à la vague j, X_{ij} est un vecteur de covariables de dimension $p \times 1$, $Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{iJ})'$, $X_i = (X_{i1}, X_{i2}, \ldots, X_{iJ})$ est une matrice de dimensions $p \times J$, $g(\cdot)$ est une « fonction de lien » monotone de type un à un différenciable, $\nu(\cdot)$ est la « fonction de variance » de forme connue, et $\phi > 0$ est le « paramètre de dispersion ». Puisqu'en général, la matrice de covariance Σ_i de dimensions $J \times J$ est difficile à spécifier, nous la modélisons sous la forme $\text{Cov}[Y_i \mid X_i] = V_i = A_i^{1/2}\mathbf{R}(\alpha)A_i^{1/2}$, une matrice de covariance « de travail », où $A_i = \text{diag}[\phi\nu(\mu_{i1}), \phi\nu(\mu_{i2}), \ldots, \phi\nu(\mu_{iJ})]$ et $\mathbf{R}(\alpha)$ est une matrice de corrélation « de travail », toutes deux de dimensions $J \times J$, et α est un vecteur qui caractérise entièrement $\mathbf{R}(\alpha)$ (voir Liang et Zeger 1986).

Pour estimer $\boldsymbol{\beta}$, nous sélectionnons un échantillon (cohorte unique) de n éléments à partir du modèle ξ et nous mesurons (avons l'intention de mesurer) chacun d'eux à J occasions. Si tous les éléments de l'échantillon répondent à chaque occasion j, la tâche peut être achevée en appliquant la méthode aux équations d'estimation généralisées (EEG) habituelles de Liang et Zeger (1986). Cependant, dans toute étude, il est rare que tous les sujets répondent à toutes les vagues. Il est plus fréquent que certains éléments de l'échantillon décrochent de l'étude.

Dans ces conditions, et en supposant que les réponses manquantes peuvent être considérées comme manquant au hasard ou MAR (pour *missing at random*) (voir Rubin 1976), en particulier que le décrochage durant une vague donnée ne dépend pas de la valeur courante (non observée), Robins, Rotnitzky et Zhao (1995) ont proposé d'estimer $\boldsymbol{\beta}$ en résolvant les équations d'estimation $\sum_{i=1}^{n}(\partial\boldsymbol{\mu}_i' / \partial\boldsymbol{\beta})V_i^{-1}\hat{\Delta}_i(\boldsymbol{y}_i - \boldsymbol{\mu}_i) = \mathbf{0}$, où $\boldsymbol{\mu}_i = (\mu_{i1}, \mu_{i2}, \ldots, \mu_{iJ})'$, $\hat{\Delta}_i = \text{diag}[R_{i1}\hat{q}_{i1}^{-1}, R_{i2}\hat{q}_{i2}^{-1}, \ldots, R_{iJ}\hat{q}_{iJ}^{-1}]$, R_{ij} est l'indicateur de réponse pour le sujet i à la vague j, et $\hat{q}_{ij}$ est une estimation de la probabilité que le sujet i soit observé durant la vague j.

Pour les applications d'enquête, on utiliserait l'équation d'estimation $\sum_{i\in s}[w_i(\partial\boldsymbol{\mu}_i' / \partial\boldsymbol{\beta})\, V_i^{-1}\hat{\Delta}_i(\boldsymbol{y}_i - \boldsymbol{\mu}_i)] = \mathbf{0}$, où w_i est le poids de sondage du sujet i. Une autre façon d'écrire cette équation est $\sum_{i\in s}(\partial\boldsymbol{\mu}_i' / \partial\boldsymbol{\beta})\, V_i^{-1}\hat{\Delta}_{wi}(\boldsymbol{y}_i - \boldsymbol{\mu}_i) = \mathbf{0}$, avec $\hat{\Delta}_{wi} = \text{diag}[w_iR_{i1}\hat{q}_{i1}^{-1}, w_iR_{i2}\hat{q}_{i2}^{-1}, \ldots, w_iR_{iJ}\hat{q}_{iJ}^{-1}]$.

Nous constatons que les éléments de la diagonale de $\hat{\Delta}_{wi}$ sont simplement égaux aux poids de sondage propres à la vague non corrigés de la non-réponse quand le sujet est observé et sont égaux à zéro quand le sujet est manquant. Cette caractéristique suggère en soi une solution au problème des cohortes multiples, qui sera présentée à la section suivante.

3.2 Une nouvelle approche pour combiner les cohortes dans les enquêtes longitudinales

Compte tenu de la discussion de la section précédente, si nous avons une enquête à panel fixe, à panel fixe plus des « unités nouvelles », à panel répété, à panel rotatif, à panel divisé ou à renouvellement de l'échantillon, nous proposons d'estimer le paramètre de superpopulation $\boldsymbol{\beta}$ dans le modèle ξ par la solution des équations d'estimation :

$$\Psi_s(\boldsymbol{\beta}) = \sum_{i \in s} \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} V_i^{-1} W_i (\boldsymbol{y}_i - \boldsymbol{\mu}_i) = \mathbf{0}, \tag{3.2}$$

où la sommation est faite sur l'échantillon s, c'est-à-dire sur les éléments sélectionnés (pour la première fois) dans l'un des échantillons $s_{1(1)}, s_{2(2)}, \ldots, s_{J(J)}$. La matrice diagonale W_i est $W_i = \text{diag}[I_i(U_1)w_{i1}, I_i(U_2)w_{i2}, \ldots, I_i(U_J)w_{iJ}]$, où w_{ij} est le poids transversal (corrigé pour la non-réponse) pour le sujet i à la vague j (à condition que le sujet i fasse partie de l'échantillon s_j) et $I_i(U_j)$ est l'indicateur signalant si le sujet i appartient ou non à la population finie U_j. À la section 3.2.1, nous présentons des arguments justifiant qu'il s'agit d'une procédure d'estimation raisonnable, et à la section 3.2.2, nous discutons du problème des valeurs manquantes.

Les poids transversaux w_{ij}, dans W_i, sont tels que l'échantillon s_j représente U_j lorsqu'il est utilisé avec lesdits poids. Cela signifie que, pour chaque observation i dans l'échantillon s_j, il existe un poids de sondage w_{ij}, qui pourrait être considéré comme le nombre d'unités que cette observation représente dans U_j. Cependant, rappelons que l'échantillon s_j est composé de différents ensembles de sujets, ou différents sous-échantillons (les différentes cohortes), et que

l'intégration de ces sous-échantillons en une variable de pondération transversale unique w_{ij} n'est pas forcément une tâche facile.

Pour la SDR, la construction d'un poids transversal pour la vague j n'est pas trop compliquée, parce que les diverses cohortes sont sélectionnées indépendamment les unes des autres, à partir de populations non chevauchantes. Dans ces conditions, le poids de base est facile à calculer, et tout ce qu'il reste à faire est la conversion pour tenir compte d'aspects tels que l'attrition et le calage sur des totaux connus de la population U_j.

Par ailleurs, dans d'autres situations, par exemple lorsqu'il n'existe pas de liste des *nouveaux* membres, la nouvelle cohorte doit parfois être sélectionnée dans la population globale au moment de la vague en question, ou en se servant d'une base de sondage contenant les nouveaux membres *plus* certains anciens membres, ou à partir de bases de sondage multiples. Le cas échéant, la construction de poids transversaux n'est pas nécessairement simple, et la théorie des bases de sondage multiples peut devoir être appliquée. Nous renvoyons le lecteur aux travaux de Lohr (2007) et de Rao et Wu (2010), ainsi qu'aux références mentionnées dans ces articles, pour les cas de ce genre.

L'expression (3.2) est une généralisation de l'équation (2.25) donnée dans Vieira (2009). Cette dernière n'est applicable que si le nombre d'observations est le même pour tous les sujets ou que toute réponse manquante peut être considérée comme manquant entièrement au hasard ou MCAR (pour *missing completely at random*) (voir Rubin 1976). Comme il est discuté dans Robins et coll. (1995), l'utilisation d'une telle équation lorsque les réponses manquantes ne sont pas de type MCAR produit des estimateurs non convergents; par conséquent, sous un schéma de rotation tel que celui de la SDR, où les sujets ne sont pas tous supprimés (ou gardés) avec les mêmes probabilités, son utilisation ne serait pas appropriée. La question de l'adéquation de l'équation (3.2) dans ce cas et quand des réponses manquent est abordée aux sections 3.2.1 et 3.2.2, respectivement. Si tous les sujets possèdent des poids transversaux qui ne varient pas au

cours du temps (ou qu'ils possèdent un seul poids longitudinal), l'équation (3.2) se réduit à l'équation (2.25) donnée dans Vieira (2009).

3.2.1 Absence de biais

La propriété d'absence de biais de la fonction d'estimation est importante, parce que, comme le soutient Song (2007, section 5.4), il s'agit de l'hypothèse la plus cruciale en vue d'obtenir un estimateur convergent.

Définissons $\boldsymbol{\beta}_N$, ledit « estimateur par recensement », comme étant la solution de l'équation d'estimation en population finie suivante :

$$\Psi_U(\boldsymbol{\beta}_N) = \sum_{i \in U} \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}_N} V_i^{-1} \mathrm{I}_i(\mathrm{U})(\boldsymbol{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta}_N)) = \mathbf{0}, \tag{3.3}$$

où la somme est calculée sur U, c'est-à-dire sur tous les éléments qui sont devenus membres de la population cible dans l'une des $U_{1(1)}, U_{2(2)}, \ldots, U_{J(J)}$, et $\mathrm{I}_i(\mathrm{U}) = \mathrm{diag}[I_i(U_1), I_i(U_2), \ldots I_i(U_J)]$. Afin de montrer l'absence de biais sous le plan de la fonction d'estimation $\Psi_s(\boldsymbol{\beta})$, nous devons montrer que son espérance sous le plan est $\Psi_U(\boldsymbol{\beta})$ pour tout $\boldsymbol{\beta}$.

Les caractéristiques du plan d'échantillonnage d'une enquête longitudinale peuvent être vues comme celles d'un échantillon à plusieurs phases tel que l'ont montré Särndal, Swensson et Wretman (1992, section 9.9). Par conséquent, nous utilisons la méthodologie d'échantillonnage à plusieurs phases pour les calculs. Nous supposons, sans perte de généralité, que l'enquête ne comprend que trois vagues; les calculs pour trois vagues seulement montrent les tendances pour J, général, en ce qui concerne l'absence de biais et la variance.

Comme nous l'avons mentionné plus haut, nous supposons que w_{ij} est le poids transversal pour le sujet i à la vague j, si ce sujet appartient à s_j, et zéro autrement. Partant de la théorie de l'échantillonnage à plusieurs phases, nous avons que, pour $i \in s_{1(1)}$, $w_{i1} = \pi_{i1}^{-1}$, $w_{i2} = \pi_{i1}^{-1}\pi_{i2|s_{1(1)}}^{-1}$ et $w_{i3} = \pi_{i1}^{-1}\pi_{i2|s_{1(1)}}^{-1}\pi_{i3|s_{2(1)}}^{-1}$, pour $i \in s_{2(2)}$, $w_{i2} = \pi_{i2}^{-1}$ et $w_{i3} = \pi_{i2}^{-1}\pi_{i3|s_{2(2)}}^{-1}$, et pour $i \in s_{3(3)}$,

$w_{i3} = \pi_{i3}^{-1}$, où π_{ij} est la probabilité d'inclusion du sujet i dans l'échantillon $s_{j(j)}$ et $\pi_{ij|s_{j-1(j')}}$ est la probabilité d'inclusion conditionnelle du sujet i dans l'échantillon $s_{j(j')}$ sachant $s_{j-1(j')}$.

En utilisant $E_p(\cdot)$ pour désigner l'espérance par rapport au plan d'échantillonnage, nous avons :

$$E_p\left[\sum_{i\in s}\frac{\partial\boldsymbol{\mu}_i'}{\partial\boldsymbol{\beta}}V_i^{-1}W_i(\boldsymbol{y}_i-\boldsymbol{\mu}_i)\right] = E_p\left[\sum_{j=1}^{3}\sum_{i\in s_{j(j)}}B_iW_i\boldsymbol{e}_i\right], \tag{3.4}$$

où $B_i = (\partial\boldsymbol{\mu}_i' / \partial\boldsymbol{\beta})V_i^{-1}$ et $\boldsymbol{e}_i = \boldsymbol{y}_i - \boldsymbol{\mu}_i$. Par exemple, pour $\sum_{i\in s_{2(2)}}B_iW_i\boldsymbol{e}_i$, nous obtenons :

$$\begin{aligned} E_p\left[\sum_{i\in s_{2(2)}}B_iW_i\boldsymbol{e}_i\right] &= E\left\{E\left[\sum_{i\in U_{2(2)}}B_iD_i\boldsymbol{e}_i \mid s_{2(2)}\right]\right\} = E\left\{\sum_{i\in U_{2(2)}}B_iD_i^*\boldsymbol{e}_i\right\} \\ &= \sum_{i\in U_{2(2)}}B_iD_i^{**}\boldsymbol{e}_i \overset{\text{def}}{=} \sum_{i\in U_{2(2)}}B_i\mathrm{I}_i(\mathrm{U})\boldsymbol{e}_i, \end{aligned} \tag{3.5}$$

où $D_i = \mathrm{diag}[0, I_i(U_2)\, w_{i2}I_i(s_{2(2)}), I_i(U_3)\, w_{i3}I_i(s_{3(2)})\, I_i(s_{2(2)})]$, $D_i^* = \mathrm{diag}[0, (I_i(U_2)\, w_{i2} \times I_i(s_{2(2)})), (I_i(U_3)\pi_{i3|s_{2(2)}}I_i(s_{2(2)})) / (\pi_{i2}\pi_{i3|s_{2(2)}})]$, et $D_i^{**} = \mathrm{diag}[0, (I_i(U_2)\pi_{i2}) / \pi_{i2}, (I_i(U_3) \times \pi_{i2}) / \pi_{i2}]$; similairement, nous pouvons montrer que $E_p\left[\sum_{i\in s_{1(1)}}B_iW_i\boldsymbol{e}_i\right] = \sum_{i\in U_{1(1)}}B_i\mathrm{I}_i(\mathrm{U})\boldsymbol{e}_i$ et $E_p\left[\sum_{i\in s_{3(3)}}B_iW_i\boldsymbol{e}_i\right] = \sum_{i\in U_{3(3)}}B_i\mathrm{I}_i(\mathrm{U})\boldsymbol{e}_i$. De ces expressions et de l'équation (3.4), nous concluons que $E_p[\Psi_s(\boldsymbol{\beta})] = \Psi_U(\boldsymbol{\beta})$ pour tout $\boldsymbol{\beta}$, ce qui signifie que la fonction d'estimation $\Psi_s(\boldsymbol{\beta})$ est sans biais sous le plan pour la fonction d'estimation en population finie.

En outre, comme la cible de l'inférence est le paramètre de superpopulation, nous devons garantir que le modèle pour μ_{ij} est tel que l'expression $E_\xi(Y_{ij} - \mu_{ij}) = 0$ est satisfaite, où $E_\xi(\cdot)$ représente l'espérance sous le modèle ξ car, si cela est le cas, nous avons :

$$E_{\xi p}[\Psi_s(\boldsymbol{\beta})] \overset{\text{def}}{=} E_\xi E_p[\Psi_s(\boldsymbol{\beta})] = E_\xi[\Psi_U(\boldsymbol{\beta})] = \sum_{i\in U}\frac{\partial\boldsymbol{\mu}_i'}{\partial\boldsymbol{\beta}}V_i^{-1}\mathrm{I}_i(\mathrm{U})E_\xi(\boldsymbol{y}_i-\boldsymbol{\mu}_i) = \mathbf{0},$$

de sorte que la fonction d'estimation $\Psi_s(\boldsymbol{\beta})$ est sans biais par rapport au modèle et au plan. La contrainte $E_\xi(Y_{ij} - \mu_{ij}) = 0$ signifie que le modèle de la moyenne doit être spécifié correctement; par conséquent, il faut faire attention aux tests diagnostiques sur les résidus pour le modèle particulier qui est ajusté.

3.2.2 Une remarque concernant la non-réponse

Dans le cas de la SDR, comme dans celui de toute autre enquête (longitudinale), il y a de la non-réponse. Certains sujets échantillonnés choisissent de ne pas participer du tout, tandis que certains participent à certaines vagues, mais pas à d'autres. Dans le cas de la SDR, pour remédier à cette situation, les poids de sondage transversaux sont corrigés pour tenir compte de la non-réponse.

Supposons que la correction pour la non-réponse à la vague j est une multiplication par l'inverse de la probabilité estimée de réponse à la vague j, $\hat{\pi}_{rij}$. Par exemple, le poids corrigé de la non-réponse pour une personne qui *a répondu* à la vague 3 (et qui avait été sélectionnée initialement à la vague 2), c'est-à-dire pour $i \in r_{3(2)}$, serait $w_{ri3} = \pi_{i2}^{-1}\pi_{i3|s_{2(2)}}^{-1}\hat{\pi}_{ri3}^{-1}$.

Nous devons redéfinir l'équation d'estimation afin d'inclure uniquement les répondants comme étant $\Psi_r(\boldsymbol{\beta}) = \sum_{i\in r}(\partial\boldsymbol{\mu}_i' / \partial\boldsymbol{\beta})V_i^{-1}W_{ri}(\boldsymbol{y}_i - \boldsymbol{\mu}_i) = \mathbf{0}$, où la somme est calculée sur l'ensemble des répondants r, c'est-à-dire sur tous les éléments qui appartenaient pour la première fois à n'importe lequel des ensembles de répondants $r_{1(1)}, r_{2(2)}, \ldots, r_{J(J)}$, et la matrice W_{ri} est $W_{ri} = \text{diag}[I_i(U_1)w_{ri1}, I_i(U_2)w_{ri2}, \ldots, I_i(U_J)w_{riJ}]$. En outre, désignons par $r_{j(j')}$ l'ensemble de répondants de la cohorte j' à la vague j. Manifestement, $w_{rij} = 0$ si $i \notin r_j = \bigcup_{j'=1}^{j} r_{j(j')}$.

Si, de surcroît, on peut supposer que le mécanisme de réponse (R) est de type MAR, nous avons alors, par exemple pour $\sum_{i\in r_{2(2)}} B_i W_{ri}\boldsymbol{e}_i$:

$$E_R\left\{\sum_{i\in r_{2(2)}} B_i W_{ri}\boldsymbol{e}_i\right\} = E_R\left\{\sum_{i\in s_{2(2)}} B_i D_i\boldsymbol{e}_i\right\} = \sum_{i\in s_{2(2)}} B_i D_i^*\boldsymbol{e}_i = \sum_{i\in s_{2(2)}} B_i D_i^{**}\boldsymbol{e}_i \overset{\text{def}}{=} \sum_{i\in s_{2(2)}} B_i W_i\boldsymbol{e}_i, \quad (3.6)$$

où $D_i = \text{diag}[0, I_i(U_2)w_{ri2}I_i(r_{2(2)}), I_i(U_3)w_{ri3}I_i(r_{3(2)})]$, $D_i^* = \text{diag}[0, (I_i(U_2)\pi_{ri2}) / (\pi_{i2}\hat{\pi}_{ri2}), (I_i(U_3)\pi_{ri3}) / (\pi_{i2}\pi_{i3|s_{2(2)}}\hat{\pi}_{ri3})]$, et $D_i^{**} = \text{diag}[0, I_i(U_2)w_{i2}, I_i(U_3)w_{i3}]$. La troisième égalité dans (3.6) requiert que le modèle de non-réponse utilisé pour $\hat{\pi}_{rij}$ satisfasse $E_R[I_i(r_{j(j')})] \overset{\text{def}}{=} \pi_{rij} = \hat{\pi}_{rij}$. Cela signifie que, dans le modèle pour $\hat{\pi}_{rij}$, nous devons inclure le

plus possible d'information possible considérée comme ayant une influence sur la propension à répondre, pour que cette hypothèse (c'est-à-dire l'hypothèse MAR) tienne. Par exemple, si l'on pense que la non-réponse est indépendante d'une vague à l'autre, on doit inclure dans le modèle pour $\hat{\pi}_{rij}$ autant de variables que possible provenant de la vague correspondante. Si, par ailleurs, il est raisonnable de supposer que la propension à répondre à une vague donnée dépend des réponses précédentes (et éventuellement de l'historique des réponses), ces réponses doivent être incluses dans le modèle de réponse, et ainsi de suite.

L'absence de biais par rapport au plan ainsi que l'absence de biais par rapport au modèle et au plan découle directement de (3.6) ainsi que de la section précédente. Par conséquent, dans la suite de l'exposé, nous ignorons la question de la non-réponse pour simplifier la notation.

3.3 Variance et estimation de la variance

Nous développons maintenant une linéarisation (développement en série de Taylor) pour la variance de l'estimateur proposé. La technique de base a été élaborée par Binder (1983). Pour simplifier les calculs et la notation, nous divisons tous les termes par N; nous redéfinissons

$$\Psi_s(\boldsymbol{\beta}) = N^{-1}\sum_{i\in s}\frac{\partial\boldsymbol{\mu}_i'}{\partial\boldsymbol{\beta}}V_i^{-1}W_i(\boldsymbol{y}_i - \boldsymbol{\mu}_i) \text{ et } \Psi_U(\boldsymbol{\beta}) = N^{-1}\sum_{i\in U}\frac{\partial\boldsymbol{\mu}_i'}{\partial\boldsymbol{\beta}}V_i^{-1}\mathrm{I}_i(\mathrm{U})(\boldsymbol{y}_i - \boldsymbol{\mu}_i),$$

où $N = \sum_{j=1}^{J} N_j$. Soit $\hat{\boldsymbol{\beta}}$ notre estimateur, qui satisfait $\Psi_s(\hat{\boldsymbol{\beta}}) = \mathbf{0}$, et soit $\boldsymbol{\beta}_N$ l'« estimateur par recensement », qui satisfait $\Psi_U(\boldsymbol{\beta}_N) = \mathbf{0}$. Supposons que $\boldsymbol{\beta}_N - \boldsymbol{\beta} = O_P(1/\sqrt{N_m})$ et $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N = O_P(1/\sqrt{n_m})$, avec $N_m = \min\{N_1, N_2, \ldots, N_J\}$ et $n_m = \min\{n_1, n_2, \ldots, n_J\}$. Nous pouvons écrire l'erreur totale de $\hat{\boldsymbol{\beta}}$ sous la forme $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N) + (\boldsymbol{\beta}_N - \boldsymbol{\beta})$ = erreur d'échantillonnage + erreur du modèle. Après certains calculs simples, la variance totale, ou plus précisément l'EQM totale, peut être décomposée comme il suit :

$$V_{\text{Tot}} = E_{\xi p}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})' = V_{\text{Éch}} + 2 \otimes C_{\text{Éch}-\text{Mod}} + o(1/n_m), \tag{3.7}$$

où $2 \otimes A = A + A'$ pour toute matrice A, $V_{\text{Éch}} = E_\xi V_p$ est la composante de « variance d'échantillonnage », $2 \otimes C_{\text{Éch}-\text{Mod}}$ est la composante croisée « variance d'échantillonnage-

modèle », $V_p = E_p[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N)']$, $C_{\text{Éch}-\text{Mod}} = E_p C_\xi$, et $C_\xi = E_\xi(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\boldsymbol{\beta}_N - \boldsymbol{\beta})'$. En outre, par développements en série de Taylor, nous pouvons obtenir les approximations suivantes : $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N = [H(\boldsymbol{\beta}_N)]^{-1}\Psi_s(\boldsymbol{\beta}_N) + o_P(1/\sqrt{n_m})$, $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = [\hat{H}(\boldsymbol{\beta})]^{-1}\Psi_s(\boldsymbol{\beta}) + o_P(1/\sqrt{n_m})$ et $\boldsymbol{\beta}_N - \boldsymbol{\beta} = [H(\boldsymbol{\beta})]^{-1}\Psi_U(\boldsymbol{\beta}) + o_P(1/\sqrt{N_m})$, où nous définissons $H(\boldsymbol{\beta}) = N^{-1}\sum_{i\in U}(\partial\boldsymbol{\mu}_i'/\partial\boldsymbol{\beta})V_i^{-1}\mathrm{I}_i(\mathrm{U})(\partial\boldsymbol{\mu}_i/\partial\boldsymbol{\beta})$ et $\hat{H}(\boldsymbol{\beta}) = N^{-1}\sum_{i\in s}(\partial\boldsymbol{\mu}_i'/\partial\boldsymbol{\beta})V_i^{-1}W_i(\partial\boldsymbol{\mu}_i/\partial\boldsymbol{\beta})$.

Nous obtenons alors, pour V_p et C_ξ dans (3.7),

$$V_p = [H(\boldsymbol{\beta}_N)]^{-1}\mathrm{Var}_p[\Psi_s(\boldsymbol{\beta}_N)][H(\boldsymbol{\beta}_N)]^{-1} + o_P(1/n_m), \tag{3.8}$$

$$\begin{aligned} C_\xi &= [\hat{H}(\boldsymbol{\beta})]^{-1}E_\xi[\Psi_s(\boldsymbol{\beta})\Psi_U'(\boldsymbol{\beta})][H(\boldsymbol{\beta})]^{-1} + o_P(1/n_m) \\ &= N^{-1}[\hat{H}(\boldsymbol{\beta})]^{-1}\hat{H}_{\Sigma V}(\boldsymbol{\beta})[H(\boldsymbol{\beta})]^{-1} + o_P(1/n_m), \end{aligned} \tag{3.9}$$

où $\mathrm{Var}_p[\Psi_s(\boldsymbol{\beta}_N)] = E_p[\Psi_s(\boldsymbol{\beta}_N)\Psi_s'(\boldsymbol{\beta}_N)]$ et $\hat{H}_{\Sigma V}(\boldsymbol{\beta}) = N^{-1}\sum_{i\in s}[(\partial\boldsymbol{\mu}_i'/\partial\boldsymbol{\beta})V_i^{-1}W_i\Sigma_i \times V_i^{-1}(\partial\boldsymbol{\mu}_i/\partial\boldsymbol{\beta})]$; la dérivation de l'expression (3.9) est donnée en annexe.

En conclusion, jusqu'à présent, nous avons trouvé que :

$$\begin{aligned} V_{\text{Tot}} &= E_\xi V_p + 2 \otimes E_p C_\xi + o(1/n_m) \\ &= E_\xi\left\{[H(\boldsymbol{\beta}_N)]^{-1}\mathrm{Var}_p[\Psi_s(\boldsymbol{\beta}_N)][H(\boldsymbol{\beta}_N)]^{-1}\right\} \\ &\quad + 2 \otimes N^{-1}E_p\left\{[\hat{H}(\boldsymbol{\beta})]^{-1}\hat{H}_{\Sigma V}(\boldsymbol{\beta})[H(\boldsymbol{\beta})]^{-1}\right\} + o(1/n_m). \end{aligned} \tag{3.10}$$

Dans (3.10), tous les termes peuvent être estimés en « insérant » l'estimation $\hat{\boldsymbol{\beta}}$, sauf pour le terme $\mathrm{Var}_p[\Psi_s(\boldsymbol{\beta}_N)]$; celui-ci est le sujet de la section suivante.

Si la fraction d'échantillonnage est faible, c'est-à-dire que $n \ll N$, le premier terme de l'expression (3.10) est une bonne approximation de la variance totale; autrement, l'expression pour V_{Tot} est simplement $E_\xi V_p$ (et les termes d'ordre inférieur). Si, au contraire, la fraction d'échantillonnage est grande, les deux termes de (3.10) sont requis.

3.3.1 Variance sous le plan de la fonction d'estimation

Afin d'obtenir une expression pour $\mathrm{Var}_p[\Psi_s(\boldsymbol{\beta}_N)]$, nous supposons que $J = 3$, comme auparavant. La méthodologie est celle de l'échantillonnage à deux phases (plus précisément, l'échantillonnage a plusieurs phases), comme il est discuté au chapitre 9 de Särndal et coll.

(1992). Après certains calculs (voir l'annexe), et en définissant $B_i = (\partial \boldsymbol{\mu}_i' / \partial \boldsymbol{\beta}) \mid_{\boldsymbol{\beta}=\boldsymbol{\beta}_N}$ $V_i^{-1}, \boldsymbol{e}_i = \boldsymbol{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta}_N), \boldsymbol{e}_{i(1\cdots 3)} = \boldsymbol{e}_i, \boldsymbol{e}_{i(2\cdots 3)} = (0, e_{i2}, e_{i3})'$, et $\boldsymbol{e}_{i(3\cdots 3)} = (0, 0, e_{i3})'$, nous obtenons :

$$\mathrm{Var}_p[\Psi_s(\boldsymbol{\beta}_N)] = \sum_{j=1}^{3} D_{(j)} = \sum_{j=1}^{3}\sum_{k=j}^{3} D_{(j)k}, \tag{3.11}$$

où $D_{(j)} \overset{\text{def}}{=} N^{-2}\mathrm{Var}_p\left(\sum_{i\in s_{j(j)}} B_i W_i \boldsymbol{e}_i\right) = \sum_{k=j}^{3} D_{(j)k}$, pour $j = 1, 2, 3,$

$$N^2 D_{(j)j} \overset{\text{def}}{=} \mathrm{Var}\left[\sum_{i\in s_{j(j)}} w_{ij} B_i \mathrm{I}_i(\mathrm{U}) \boldsymbol{e}_{i(j\cdots 3)}\right], \text{ pour } j = 1, 2, 3,$$

$$N^2 D_{(j-1)j} \overset{\text{def}}{=} E\left\{\mathrm{Var}\left[\sum_{i\in s_{j(j-1)}} w_{ij} B_i \mathrm{I}_i(\mathrm{U}) \boldsymbol{e}_{i(j\cdots 3)} \mid s_{j-1(j-1)}\right]\right\}, \text{ pour } j = 2, 3,$$

$$N^2 D_{(1)3} \overset{\text{def}}{=} E\left\{E\left[\mathrm{Var}\left(\sum_{i\in s_{3(1)}} w_{i3} B_i \mathrm{I}_i(\mathrm{U}) \boldsymbol{e}_{i(3\cdots 3)} \mid s_{2(1)}, s_{1(1)}\right) \mid s_{1(1)}\right]\right\},$$

et, en annexe, nous montrons que :

$$N^2 D_{(j)k} = \mathrm{Var}\left[\sum_{i\in s_{k(j)}} w_{ik} B_i \mathrm{I}_i(\mathrm{U}) \boldsymbol{e}_{i(k\cdots 3)}\right] - \mathrm{Var}\left[\sum_{i\in s_{k-1(j)}} w_{i,k-1} B_i \mathrm{I}_i(\mathrm{U}) \boldsymbol{e}_{i(k\cdots 3)}\right],$$

pour $j = 1, 2, 3$ et $3 \geq k > j$. De manière générale, nous avons prouvé ce qui suit.

Propriété 3.1 La variance (sous le plan) de $\Psi_s(\boldsymbol{\beta}_N)$ peut être décomposée en :

$$\begin{aligned}
&\mathrm{Var}_p[\Psi_s(\boldsymbol{\beta}_N)] \\
&= \frac{1}{N^2}\sum_{j'=1}^{J}\sum_{j=j'}^{J}\left\{\mathrm{Var}_p\left[\sum_{i\in s_{j(j')}} w_{ij} B_i \mathrm{I}_i(U)\mathbf{e}_{i(j\cdots J)}\right] - \mathrm{Var}_p\left[\sum_{i\in s_{j-1(j')}} w_{i,j-1} B_i \mathrm{I}_i(U)\mathbf{e}_{i(j\cdots J)}\right]\right\} \quad (3.12) \\
&= \frac{1}{N^2}\sum_{j=1}^{J}\left\{\mathrm{Var}_p\left[\sum_{i\in s_j} w_{ij} B_i \mathrm{I}_i(U)\mathbf{e}_{i(j\cdots J)}\right] - \mathrm{Var}_p\left[\sum_{i\in s_{j-1}} w_{i,j-1} B_i \mathrm{I}_i(U)\mathbf{e}_{i(j\cdots J)}\right]\right\}, \quad (3.13)
\end{aligned}$$

où nous posons que $w_{i,j-1} = 0$ quand $j = j', w_{i0} = 0,$ et pour obtenir (3.13), nous avons changé les variable et utilisé la propriété d'indépendance entre les cohortes.

Dans (3.11), (3.12) et (3.13), nous avons supposé que les cohortes sont indépendantes sous le plan. Cependant, dans certains cas, cette hypothèse ne tient pas; un exemple est celui d'une base

de sondage multiple dont nous avons discuté dans la première partie de la section 3.2. Une autre situation dans laquelle il pourrait ne pas être approprié de supposer que les cohortes sont indépendantes est celle où les ajustements de pondération recoupent les cohortes, ce qui est le cas de la SDR; nous discutons de ce problème à la section 5. Les calculs pour le cas des trois cohortes, fournis en annexe, montrent que l'équation (3.13) est vérifiée pour les termes de variance, même sans indépendance. Nous précisons aussi dans l'annexe les conditions sous lesquelles il s'agit d'une bonne approximation pour les termes de covariance.

3.3.2 Estimation

L'estimation de V_{Tot} dans (3.10) peut être effectuée comme il suit. $H(\boldsymbol{\beta}_N), \hat{H}(\boldsymbol{\beta})$ et $H(\boldsymbol{\beta})$ peuvent être estimés par $\hat{H}(\hat{\boldsymbol{\beta}})$. $\hat{H}_{\Sigma V}(\boldsymbol{\beta})$ peut être estimé par $\hat{H}_{\Sigma V}(\hat{\boldsymbol{\beta}})$, où $\Sigma_i = \text{Cov}[Y_i \mid X_i]$ peut être estimé par $\hat{\boldsymbol{e}}_i \hat{\boldsymbol{e}}_i'$.

Nous utilisons (3.13) dans la propriété 3.1 pour estimer $\text{Var}_p[\Psi_s(\boldsymbol{\beta}_N)]$. À condition qu'il existe une méthode pour estimer la variance des estimateurs (transversaux) de Horvitz-Thompson (H-T), l'expression (3.13) peut être utilisée. Si nous définissons $Z_{ij} = B_i \text{I}_i(\text{U}) \boldsymbol{e}_{i(j\cdots J)}$, nous constatons que chaque terme intervenant dans le calcul de (3.13) tel que $\text{Var}_p\left[\sum_{i\in s_j} w_{ij} Z_{ij}\right]$ est simplement la variance d'un estimateur H-T de la vague j. De toute évidence, la méthode d'estimation de la variance doit prendre en considération à la fois le plan d'échantillonnage et toute correction pour tenir compte de la non-réponse et du calage, mais cela ne présente aucune difficulté de plus que celle posée par tout problème transversal, car tous les éléments sont appliqués transversalement. Dans le cas de la SDR, les variances des estimateurs transversaux sont estimées par rééchantillonnage, mais toute méthode d'estimation de la variance sous le plan peut être utilisée.

Nous utilisons les poids de rééchantillonnage transversaux fournis par le programme de la SDR, mais nous ne réestimons pas le paramètre d'intérêt pour chaque réplique. Premièrement, notons que nous ne devons effectuer le rééchantillonnage que pour l'estimation de la « partie

substantielle » ($\mathrm{Var}_p[\Psi_s(\boldsymbol{\beta}_N)]$) de la variance sous le plan ($E_\xi V_p$). Deuxièmement, bien que $\hat{\boldsymbol{\beta}}$ ne figure pas dans l'expression de l'estimateur H-T dont la variance doit être calculée (et recalculée à chaque réplique), les travaux de Roberts, Binder, Kovačević, Pantel et Phillips (2003), qui appliquent la méthode du « bootstrap de la fonction d'estimation » (Hu et Kalbfleisch 2000) à des données d'enquête, montrent que dans des conditions telles que les nôtres, il n'est pas nécessaire de recalculer l'estimateur à chaque réplique, mais que l'estimateur sur l'échantillon complet suffit. Cette simplification accélère le calcul des estimations répétées.

En guise d'illustration, disons que nous en sommes à la vague j, c'est-à-dire que nous estimons le j^{e} terme dans (3.13). La r^{e} réplique du premier terme est $\sum_{i\in s_j} w_{ij}^{(r)} B_i(\hat{\boldsymbol{\beta}})\mathrm{I}_i(\mathrm{U})\boldsymbol{e}_{i(j\cdots J)}(\hat{\boldsymbol{\beta}})$, où $w_{ij}^{(r)}$ est le r^{e} poids de rééchantillonnage pour le sujet i à la vague j, et la r^{e} réplique du deuxième terme est $\sum_{i\in s_{j-1}} w_{i,j-1}^{(r)} B_i(\hat{\boldsymbol{\beta}})\mathrm{I}_i(\mathrm{U})\boldsymbol{e}_{i(j\cdots J)}(\hat{\boldsymbol{\beta}})$, où $w_{i,j-1}^{(r)}$ est le r^{e} poids de rééchantillonnage pour le sujet i à la vague $j-1$.

4 Application à la SDR

Nous utilisons pour l'application le jeu de données restreint de la SDR, aux termes d'un accord de licence conclu avec la NSF. La SDR collecte des données sur, entre autres, la situation d'emploi, l'employeur principal, l'emploi principal, les emplois antérieurs, les études récentes, les caractéristiques démographiques et l'incapacité, qui varient d'une vague à l'autre. Nous n'utilisons que l'information recueillie durant toutes les vagues d'intérêt, à savoir 1995, 1997, 1999, 2001, 2003, 2006 et 2008.

Pour illustrer notre méthodologie, nous avons construit un modèle du salaire des individus au fil du temps. La réponse est le logarithme du salaire (dans l'emploi principal), avec une fonction de lien identité et plusieurs covariables; la modélisation du logarithme du salaire (par opposition au salaire) est une pratique courante. Le modèle contient des covariables indépendantes du temps (comme le sexe) ainsi que des covariables dépendant du temps (tel que le secteur d'emploi).

Nous avons quatre grandes catégories de covariables. Les *variables relatives au diplôme* sont le domaine du diplôme, le nombre d'années écoulées depuis l'obtention du diplôme et l'âge au moment de l'obtention du diplôme. Les *variables relatives à l'emploi* sont le domaine ou la catégorie d'emploi, le secteur, l'indicateur d'études postdoctorales, l'indicateur de membre auxiliaire du corps professoral, le nombre d'heures travaillées par semaine dans l'emploi principal, le nombre de semaines par année dans l'emploi principal, le lien entre l'emploi et le diplôme de doctorat, le travail à temps partiel pour différentes raisons, le nombre de mois depuis le début de l'emploi principal, le mois de début de l'emploi principal, le fait que l'employeur/le type d'emploi a changé ou non depuis la vague précédente et le fait que le changement d'employeur/de type d'emploi depuis la vague précédente était dû ou non à une mise à pied ou à une cessation d'emploi. Les *caractéristiques démographiques de la personne* sont le sexe, la citoyenneté, la race/ethnicité, la présence d'enfants dans la famille, l'état matrimonial, la situation de travail du (de la) conjoint(e). Enfin, les *variables d'« environnement »* sont le nombre d'années écoulées depuis 1995, l'État (d'emploi) et l'indice des prix à la consommation (de la région d'emploi). La liste complète des variables, des interactions et des catégories figure dans Carrillo et Karr (2011). Pour les variables catégoriques, la catégorie de référence est celle pour laquelle la fréquence est la plus élevée.

Le jeu de données pour notre modèle comprend 59 346 sujets et 190 693 observations, réparties comme il suit : $n_{95} = 30\,234$, $n_{97} = 30\,652$, $n_{99} = 26\,732$, $n_{01} = 26\,778$, $n_{03} = 24\,956$, $n_{06} = 25\,910$ et $n_{08} = 25\,431$. Ces données correspondent à des salaires non manquants compris entre 5 000 $ et 999 995 $, pour les personnes dont l'âge est cohérent d'une vague à l'autre et pour lesquelles ne manque pas la valeur de la variable indiquant si l'employeur (établissement d'enseignement postsecondaire) était un établissement public ou privé. Les poids de sondage (transversaux) moyens pour chacune de ces vagues sont : $\bar{w}_{95} = 15{,}37$, $\bar{w}_{97} = 16{,}28$, $\bar{w}_{99} = 19{,}96$, $\bar{w}_{01} = 20{,}74$, $\bar{w}_{03} = 22{,}71$, $\bar{w}_{06} = 22{,}93$ et $\bar{w}_{08} = 24{,}88$.

Les poids de sondage que nous utilisons pour chaque vague sont les poids corrigés finaux. Ces poids correspondent aux poids de sondage originaux corrigés pour la non-réponse et la poststratification. Cependant, la théorie que nous avons élaborée à la section 3 repose sur l'hypothèse que les poids sont les inverses des probabilités de sélection; autrement dit, les poids de sondage originaux. Il s'agit d'une disparité dont nous prévoyons étudier les effets dans l'avenir. Par ailleurs, les calculs de la dernière partie de l'annexe (qui ne reposent sur aucune hypothèse concernant les poids) donnent à penser que l'effet de cette disparité est faible.

Les covariables et les interactions que nous avons prises en considération ont été choisies parce qu'elles étaient suggérées par les analyses exploratoires ou par les spécialistes du domaine de la NSF. Carrillo et Karr (2011) présentent les coefficients β estimés du modèle $y_{ij} = \log(\text{SALAIRE}_{ij}) = X'_{ij}\boldsymbol{\beta} + \varepsilon_{ij}$, où X_{ij} comprend l'ordonnée à l'origine ainsi que les autres covariables. Ce $\boldsymbol{\beta}$ correspond à celui du modèle ξ, dans la formule (3.1), dont les propriétés sont discutées à la section 3. La matrice de covariance de travail est estimée être $\hat{V}_i = \hat{\phi}\mathbf{R}(\hat{\alpha})$, avec $\hat{\phi} = \hat{\sigma}^2 = \left(\sum_{i\in s}\sum_{j=95}^{08} w_{ij}\hat{e}_{ij}^2\right) \Big/ \left(\sum_{i\in s}\sum_{j=95}^{08} w_{ij} - p\right) = 0{,}196$, où $\hat{e}_{ij} = y_{ij} - X'_{ij}\hat{\boldsymbol{\beta}}$ et $p = 208$ est le nombre de covariables dans X_{ij}, w_{ij} est le poids transversal pour le sujet i à la vague j à condition que $i \in s_j$ et zéro autrement. L'estimation $\hat{\alpha}$ contient les $21 = (7\times 6)/2$ autocorrélations estimées $\hat{\alpha}_{jj'} = \hat{\alpha}_{j'j} = \left(\sum_{i\in s}\sqrt{w_{ij}}\sqrt{w_{ij'}}\hat{e}_{ij}\hat{e}_{ij'}\right) \Big/ \left(\hat{\phi}\left[\sum_{i\in s}\sqrt{w_{ij}}\sqrt{w_{ij'}} - p\right]\right)$, pour $j \neq j' = 1995,\ 1997,\ 1999,\ 2001,\ 2003,\ 2006,\ 2008$, et $\hat{\alpha}_{jj} = 1$ pour tout j. Les valeurs estimées forment la matrice d'autocorrélation :

$$\mathbf{R}(\hat{\alpha}) = \begin{pmatrix} 1 & \hat{\alpha}_{95,97} & \hat{\alpha}_{95,99} & \hat{\alpha}_{95,01} & \hat{\alpha}_{95,03} & \hat{\alpha}_{95,06} & \hat{\alpha}_{95,08} \\ & 1 & \hat{\alpha}_{97,99} & \hat{\alpha}_{97,01} & \hat{\alpha}_{97,03} & \hat{\alpha}_{97,06} & \hat{\alpha}_{97,08} \\ & & 1 & \hat{\alpha}_{99,01} & \hat{\alpha}_{99,03} & \hat{\alpha}_{99,06} & \hat{\alpha}_{99,08} \\ & & & 1 & \hat{\alpha}_{01,03} & \hat{\alpha}_{01,06} & \hat{\alpha}_{01,08} \\ & & & & 1 & \hat{\alpha}_{03,06} & \hat{\alpha}_{03,08} \\ & \text{sym} & & & & 1 & \hat{\alpha}_{06,08} \\ & & & & & & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0{,}38 & 0{,}36 & 0{,}32 & 0{,}30 & 0{,}28 & 0{,}27 \\ & 1 & 0{,}42 & 0{,}36 & 0{,}33 & 0{,}32 & 0{,}31 \\ & & 1 & 0{,}46 & 0{,}38 & 0{,}36 & 0{,}34 \\ & & & 1 & 0{,}47 & 0{,}40 & 0{,}38 \\ & & & & 1 & 0{,}49 & 0{,}44 \\ & \text{sym} & & & & 1 & 0{,}55 \\ & & & & & & 1 \end{pmatrix}.$$

Nous présentons maintenant certaines conclusions concernant les salaires de la population active possédant un doctorat fondées sur les coefficients estimés, qui figurent dans Carrillo et Karr (2011). Avant tout, une estimation raisonnable du salaire moyen prend en considération l'ordonnée à l'origine, le nombre d'heures travaillées par semaine (dont la moyenne est 47) et le nombre d'années écoulées depuis l'obtention du diplôme (en moyenne 15), de sorte qu'une estimation de la moyenne globale est $\exp(9,4 + 47 \times 0,038 - 47^2 \times 0,0003 + 15 \times 0,03 - 15^2 \times 0,0006) = 52\ 067\ \$,$ pour un sujet pour lequel toutes les autres covariables continues sont égales à zéro et dans la catégorie de référence pour toutes les variables catégoriques.

Toutes choses étant constantes par ailleurs, les salaires des femmes sont environ égaux à 93,4 % de ceux des hommes, tandis que la race ne semble pas avoir d'effet sur les salaires. L'effet du terme d'interaction du sexe et du nombre d'années écoulées depuis 1995 n'est pas significatif; par conséquent, cette différence de salaire ne varie pas au cours du temps. Soulignons qu'en n'utilisant que les données pour une seule année, nous ne serions pas capables d'évaluer l'effet du temps. Fait encore plus important, en utilisant uniquement les données d'une seule vague, disons 2008, nous ne serions pas capables de déterminer si l'effet d'être de sexe féminin varie au cours du temps.

Les titulaires d'un doctorat ayant un emploi de gestionnaire sont ceux dont les salaires sont les plus élevés, suivis par ceux appartenant aux professions du secteur de la santé; par ailleurs, ceux dont les salaires sont les plus faibles sont ceux employés dans les « autres » professions, suivis de ceux travaillant dans le domaine des sciences politiques.

Parmi les secteurs d'emploi, les salaires les plus élevés sont observés dans le secteur à but lucratif (20 % plus élevés que pour la catégorie de référence correspondant à professeur permanent dans les établissements publics offrant des programmes de quatre années), suivi, dans l'ordre, par l'administration fédérale, le secteur du travail autonome et le secteur sans but lucratif, qui offrent tous des salaires plus élevés que celui de la catégorie de référence. Les salaires les

plus faibles sont ceux offerts par les collèges avec programmes de deux ans et les établissements avec programmes de deux et de quatre ans pour lesquels la permanence ne s'applique pas.

L'effet négatif le plus important sur les salaires est également observé dans le secteur de l'enseignement. Les personnes ayant un poste de membre auxiliaire du corps professoral ont des salaires qui correspondent environ à 59 % des salaires de titulaires d'un doctorat comparable. Fait qui n'est pas étonnant, les salaires postdoctoraux ne correspondent qu'à environ 74 % des salaires de personnes comparables occupant d'autres types de poste.

Le secteur d'emploi est aussi un facteur qui contribue à la dépendance, difficile à interpréter, du salaire à l'égard du mois où a débuté l'occupation du poste courant : les salaires sont plus faibles pour les mois de début correspondant à août et à septembre. Des analyses supplémentaires montrent que l'effet du mois ne s'observe que dans le secteur de l'enseignement, où, comme nous l'avons vu, les salaires sont plus faibles que dans l'industrie ou l'administration publique, et où il est fréquent que les emplois débutent en août ou en septembre. Par conséquent, le secteur d'emploi donne partiellement, mais pas entièrement la réponse. Une ventilation plus fine du secteur de l'enseignement, en utilisant les classifications de Carnegie, réduit, mais ne supprime pas le caractère significatif de l'effet du mois. La SDR ne semble pas fournir suffisamment de données pour éliminer entièrement les effets du mois, de sorte que nous avons gardé la définition des secteurs d'emploi de la SDR.

Les personnes titulaires d'un diplôme en informatique et en sciences de l'information sont celles dont les salaires sont les plus élevés (environ 20 % plus élevés que pour les sciences biologiques), suivies par celles titulaires d'un diplôme en génie électrique et informatique et en économie (environ 16 % plus élevés). Les titulaires d'un doctorat en sciences agricoles et alimentaires, en sciences de l'environnement et de la vie, en sciences de la terre, de l'atmosphère et des océans, et en « autres » sciences sociales reçoivent les salaires les plus faibles. Les « autres » sciences sociales sont les sciences sociales à l'exclusion des sciences économiques et politiques.

Les personnes mariées sont celles dont les salaires sont les plus élevés, suivies par les personnes dans une relation de type mariage, les personnes veuves, les personnes séparées, les personnes divorcées et les personnes jamais mariées. Ces dernières ont des salaires de l'ordre de 89 % seulement de celui des personnes mariées; on pourrait soutenir qu'il existe une certaine association entre le fait de n'avoir jamais été marié et l'âge. La présence d'enfants de plus de deux ans est associée à des salaires plus élevés, mais la présence d'enfants plus jeunes ne l'est pas.

Les titulaires d'un doctorat occupant un emploi qui n'est relié que dans une certaine mesure au domaine de leur diplôme gagnent environ 93 % de ce que les personnes possédant un emploi étroitement associé à leur domaine d'études (la catégorie de référence) gagnent. Si l'emploi n'est pas relié au diplôme de doctorat à cause d'un changement de carrière ou d'intérêt professionnel, le salaire est de l'ordre de 82 % de ce gagnent les personnes occupant un emploi étroitement associé à leur domaine d'études. Par ailleurs, les personnes dont l'emploi n'est pas relié à leur domaine d'études pour d'autres raisons ne gagnent qu'environ 76 % du salaire de celles appartenant à la catégorie de référence.

On observe une augmentation de salaire d'environ 3 % pour chaque année supplémentaire depuis l'obtention du diplôme de doctorat, mais il existe un effet de diminution pour les nombres d'années élevés. Nous interprétons cela comme l'effet de l'expérience. Une petite pénalité est associée à l'obtention du doctorat plus tard dans la vie; pour chaque année d'âge supplémentaire au moment de l'obtention du diplôme, le salaire est réduit de 1 %.

Nous avons également constaté que l'indice des prix à la consommation (IPC) régional a un effet significatif. Le salaire est d'autant plus élevé que l'IPC est élevé. Nous n'avons pas pu utiliser l'IPC associé au marché du travail auquel appartient l'emploi parce que les données de la SDR n'indiquent pas l'emplacement géographique de manière plus précise que l'État. Nous avons inclus l'État dans le modèle comme mesure de substitution du coût de la vie; l'effet de l'État est fortement significatif et les coefficients de certains États comptent parmi les plus élevés

dans l'ensemble. Les salaires les plus élevés sont offerts en Californie, à Washington D.C. et ses faubourgs, ainsi que dans la ville de New York et ses faubourgs. Par ailleurs, les salaires les plus faibles sont observés à Puerto Rico, dans le Vermont, au Montana, dans le Maine, dans l'Idaho, dans le Dakota du Sud, dans le Dakota du Nord, et dans les territoires/à l'étranger.

Le fait d'avoir un emploi à temps partiel en raison de la retraite ou d'une semi-retraite a un effet significatif et figure dans plusieurs termes d'interaction significatifs. Étant donné ces résultats, nous ne pensons pas que les données à notre disposition à l'heure actuelle brossent le tableau complet de la retraite, par exemple, en ce qui concerne les personnes qui sont (semi) retraitées et continuent d'avoir un emploi à temps plein.

Enfin, nous avons analysé les résidus; les figures 4.1 et 4.2 montrent un graphique en boîtes à moustaches des résidus standardisés par année et un graphique spaghetti des résidus standardisés, respectivement.

La figure 4.1 montre que l'ajustement du modèle est raisonnablement bon pour toutes les années de référence, car la plupart des résidus standardisés sont compris entre -2 et 2. En outre, les distributions des résidus ne semblent pas différer fortement d'une année à l'autre.

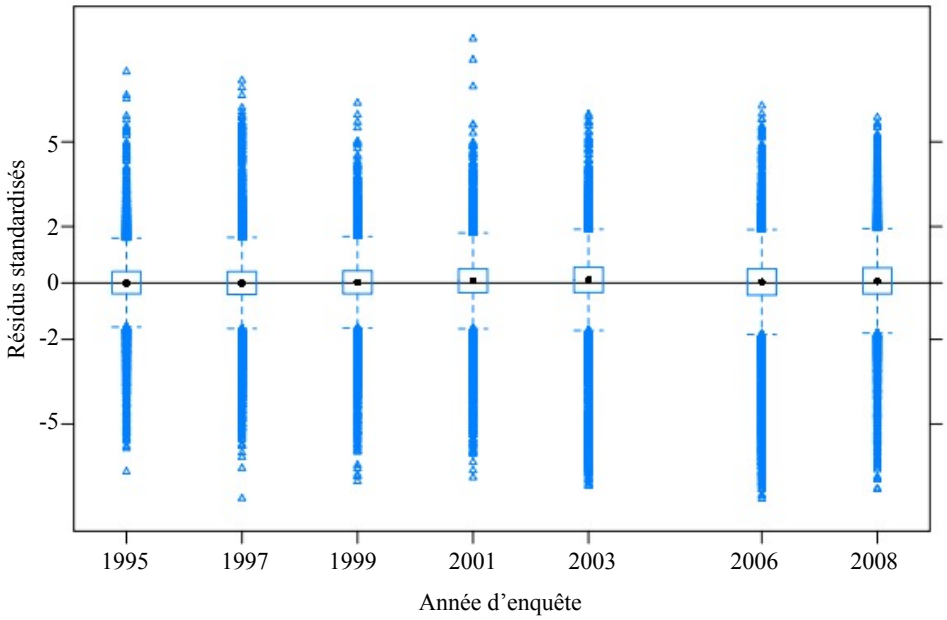

Figure 4.1 Boîtes à moustaches des résidus standardisés par année

La figure 4.2 nous permet également de conclure que l'ajustement du modèle est assez bon pour la plupart des personnes, car la plupart des lignes fluctuent entre -2 et 2. Néanmoins, pour quelques personnes, le modèle semble prédire de façon excessivement élevé fortement en 2003, ainsi qu'en 2006. Nous avons inclus plusieurs termes dans le modèle pour corriger ce problème, mais manifestement aucun ne semble le faire complètement.

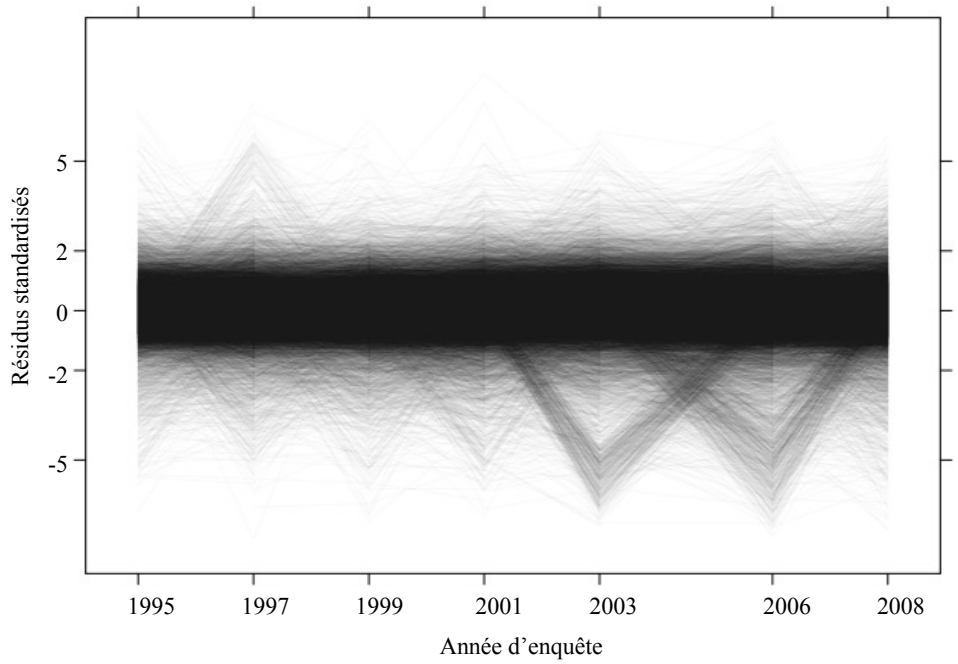

Figure 4.2 Graphique spaghetti des résidus standardisés

En dernier lieu, nous avons tenté de produire des arbres de classification exploratoires pour ces anomalies résiduelles. Nous avons constaté que, pour l'ensemble de données disponible, le seul élément relié à ces anomalies était le mode d'enquête. En 2003, les anomalies sont disproportionnellement importantes pour les réponses en ligne, et en 2006, elles sont disproportionnellement importantes pour les réponses par ITAO. Nous concluons qu'il existe un effet de mode de collecte pour ces deux années bien que, ces années-là, les répondants présentaient une caractéristique différente qui n'est pas incluse dans les variables existantes.

Enfin, la représentation graphique des valeurs prédites en fonction des valeurs observées (que l'on peut consulter dans Carrillo et Karr [2011]) montre la même chose. Pour la plupart des

observations, le modèle donne de bons résultats, sauf pour ces quelques cas en 2003 et en 2006 pour lesquels il produit une surestimation importante.

5 Conclusion et travaux de recherche à venir

Nous avons proposé une nouvelle approche pour combiner différentes cohortes d'une enquête longitudinale. La principale exigence de notre méthode est qu'il existe un poids de sondage transversal pour chaque vague, ou que l'on puisse en construire un d'après les données à notre disposition. Ce poids doit permettre de faire une inférence statistique pour la population d'intérêt durant la vague correspondante. Dans ces conditions, notre méthode devrait donner de meilleurs résultats que les procédures d'estimation habituelles (dans lesquelles l'autocorrélation n'est pas intégrée) dans de nombreuses situations pratiques, en particulier lorsque l'autocorrélation entre les réponses fournies par un même sujet est forte.

En général, les praticiens des sondages évitent autant que possible d'utiliser des poids de sondage multiples. Cependant, dans le cas des panels rotatifs, il s'agit d'une approche intéressante pour au moins deux raisons. D'une part, elle permet d'utiliser toutes les données d'une manière claire et cohérente dans une seule procédure d'analyse. D'autre part, nous avons montré comment des poids de sondage transversaux auxquels ont a facilement accès peuvent être utilisés directement pour l'analyse longitudinale, sans qu'il soit nécessaire d'élaborer, de sauvegarder et de diffuser un ou plusieurs poids longitudinaux supplémentaires.

Notre méthode est directement applicable à toute forme d'enquête longitudinale, à condition qu'il existe des poids de sondage transversaux à chaque vague (ou qu'ils puissent être créés), et que ces poids représentent la population d'intérêt durant la vague en question.

Pour la théorie que nous avons élaborée au sujet de la variance de l'estimateur proposé, nous avons utilisé les poids de sondage (transversaux) w_{ij}, qui sont les inverses des probabilités d'inclusion. Cependant, pour l'application à la SDR, nous avons utilisé les poids de sondage (transversaux) finaux dans notre modèle pour les salaires, qui ne sont pas les poids de sondage

originaux, mais les poids corrigés (de la manière habituelle). Cette disparité nécessite une étude plus approfondie.

De même, dans nos calculs de la variance, nous avons supposé que les cohortes étaient indépendantes. Cependant, la SDR ne satisfait pas entièrement cette hypothèse pour deux raisons. Premièrement, à n'importe quelle vague, la sélection de l'échantillon provenant des anciennes cohortes n'est pas effectuée indépendamment d'une cohorte à l'autre. Afin de réduire le nombre de strates, depuis 1991, la NSF les a regroupées en fonction de l'année d'obtention du diplôme pour les anciennes cohortes. En outre, les corrections pour la poststratification apportées aux poids de sondage ne sont pas conditionnées sur les cohortes non plus, et par conséquent, les poids sont partagés entre les cohortes. Ce schéma de sélection de l'échantillon et la procédure de correction de la pondération violent l'hypothèse d'indépendance entre les cohortes. Certains calculs supplémentaires (inclus en annexe) ont montré que l'indépendance entre les cohortes n'est pas une exigence vraiment cruciale pour que notre méthode d'estimation de la variance produise de bonnes approximations, comme il est expliqué à la section 3.3.1. Durant de futurs travaux de recherche, nous prévoyons évaluer plus en détail l'effet de ce problème.

Remerciements

La présente étude a été financée par la subvention SRS-1019244 accordée par la NSF au National Institute of Statistical Sciences (NISS). Les opinions, constations et conclusions ou recommandations exposées dans la présente publication sont celles des auteurs et ne reflètent pas nécessairement celles de la National Science Foundation. Les auteurs remercient Paul Biemer de RTI International, Stephen Cohen et Nirmala Kannankutty du National Center for Science and Engineering Statistics à la NSF, et Criselda Toto, anciennement du NISS, de leurs nombreuses discussions éclairantes durant l'étude. Nous remercions aussi le rédacteur associé et deux examinateurs de leurs suggestions constructives.

Annexe – Preuves

- Pour développer une expression pour C_ξ, nous commençons par simplifier $\Psi_s(\boldsymbol{\beta})\Psi'_U(\boldsymbol{\beta})$. Soit $\boldsymbol{F}_{i(k)} = B_i \mathrm{I}_i(\mathrm{U})\boldsymbol{e}_{i(k\cdots 3)}$ pour $k = 1, 2, 3,$ alors nous avons :

$$\begin{aligned} N^2\Psi_s(\boldsymbol{\beta})\Psi'_U(\boldsymbol{\beta}) &= \sum_{i\in s} B_i W_i \boldsymbol{e}_i \sum_{i\in U} \boldsymbol{e}'_i \mathrm{I}_i(\mathrm{U}) B'_i = \left[\sum_{i\in s} B_i W_i \boldsymbol{e}_i\right]\left[\sum_{i\in s} \boldsymbol{F}'_{i(1)} + \sum_{i\notin s} \boldsymbol{F}'_{i(1)}\right] \\ &= \sum_{i\in s} B_i W_i \boldsymbol{e}_i \sum_{i\in s} \boldsymbol{F}'_{i(1)} + \sum_{i\in s} B_i W_i \boldsymbol{e}_i \sum_{i\notin s} \boldsymbol{F}'_{i(1)} \\ &= \sum_{i\in s} B_i W_i \boldsymbol{e}_i \boldsymbol{e}'_i B'_i + \sum_{i\in s}\sum_{\substack{k\in s \\ k\neq i}} B_i W_i \boldsymbol{e}_i \boldsymbol{e}'_k \mathrm{I}_k(\mathrm{U}) B'_k + \mathrm{A}, \end{aligned}$$

où $\mathrm{A} = \left(\sum_{i\in s} B_i W_i \boldsymbol{e}_i\right)\left(\sum_{i\notin s} \boldsymbol{F}'_{i(1)}\right)$, et soit $\mathrm{B} = \sum_{i\in s}\sum_{\substack{k\in s \\ k\neq i}} B_i W_i \boldsymbol{e}_i \boldsymbol{e}'_k \mathrm{I}_k(\mathrm{U}) B'_k$. Les deux sommes dans A sont indépendantes du modèle, $\boldsymbol{e}_i$ et $\boldsymbol{e}'_k$ (dans B) sont deux termes dépendants du modèle, et les expressions A et B ont toutes deux une espérance nulle sous le modèle; par conséquent, $E_\xi[\Psi_s(\boldsymbol{\beta})\Psi'_U(\boldsymbol{\beta})] = N^{-2}\sum_{i\in s} B_i W_i E_\xi[\boldsymbol{e}_i \boldsymbol{e}'_i] B'_i = N^{-2}\sum_{i\in s} B_i W_i \Sigma_i B'_i = N^{-1}\hat{H}_{\Sigma V}(\boldsymbol{\beta})$; l'équation (3.9) s'ensuit.

- Nous développons maintenant l'expression pour $\mathrm{Var}_p[\Psi_s(\boldsymbol{\beta}_N)]$, la variance sous le plan de la fonction d'estimation; nous redéfinissons $B_i = (\partial\boldsymbol{\mu}'_i / \partial\boldsymbol{\beta})\,|_{\boldsymbol{\beta}=\boldsymbol{\beta}_N}\, V_i^{-1}$ et $\boldsymbol{e}_i = \boldsymbol{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta}_N)$; alors

$$\begin{aligned} \mathrm{Var}_p[\Psi_s(\boldsymbol{\beta}_N)] &= \mathrm{Var}_p\left(\frac{1}{N}\sum_{i\in s} B_i W_i \boldsymbol{e}_i\right) \\ &= \frac{1}{N^2}\mathrm{Var}_p\left(\sum_{i\in s_{1(1)}} B_i W_i \boldsymbol{e}_i + \sum_{i\in s_{2(2)}} B_i W_i \boldsymbol{e}_i + \sum_{i\in s_{3(3)}} B_i W_i \boldsymbol{e}_i\right) \\ &= \frac{1}{N^2}\mathrm{Var}_p\left(\sum_{i\in s_{1(1)}} B_i W_i \boldsymbol{e}_i\right) + \frac{1}{N^2}\mathrm{Var}_p\left(\sum_{i\in s_{2(2)}} B_i W_i \boldsymbol{e}_i\right) \\ &\quad + \frac{1}{N^2}\mathrm{Var}_p\left(\sum_{i\in s_{3(3)}} B_i W_i \boldsymbol{e}_i\right) = D_{(1)} + D_{(2)} + D_{(3)}, \qquad \text{(A.1)} \end{aligned}$$

où, pour la ligne (A.1), nous supposons que les (trois) cohortes sont indépendantes sous le plan. Maintenant, $N^2 D_{(1)} = \mathrm{Var}_p\left[\sum_{i\in U_{1(1)}} B_i W_i\, \mathrm{Diag}\{\boldsymbol{I}_{i(1)}\}\,\boldsymbol{e}_i\right] = \mathrm{Var}_p\left[\sum_{i\in U_{1(1)}} B_i W_i\, \mathrm{Diag}\{\boldsymbol{e}_i\}\,\boldsymbol{I}_{i(1)}\right]$, où $\mathrm{Diag}\{\boldsymbol{e}\}$ est, pour un vecteur colonne $\boldsymbol{e}$, une matrice diagonale dont les entrées sur la diagonale sont les éléments de $\boldsymbol{e}$, et

$\boldsymbol{I}_{i(1)} = \left(I_i(s_{1(1)}),\ I_i(s_{2(1)})I_i(s_{1(1)}),\ I_i(s_{3(1)})I_i(s_{2(1)})\, I_i(s_{1(1)}) \right)'$. Similairement, nous pouvons obtenir $N^2 D_{(2)} = \mathrm{Var}_p \left[\sum_{i \in U_{2(2)}} B_i W_i \, \mathrm{Diag}\{\boldsymbol{e}_i\} \, \boldsymbol{I}_{i(2)} \right]$ et $N^2 D_{(3)} = \mathrm{Var}_p \left[\sum_{i \in U_{3(3)}} B_i W_i \, \mathrm{Diag}\{\boldsymbol{e}_i\} \, \boldsymbol{I}_{i(3)} \right]$, où $\boldsymbol{I}_{i(2)} = \left(0, I_i(s_{2(2)}), I_i(s_{3(2)})I_i(s_{2(2)}) \right)'$, et $\boldsymbol{I}_{i(3)} = \left(0, 0, I_i(s_{3(3)}) \right)'$. Maintenant, concentrons-nous sur $D_{(1)}$; en posant que $C_i = B_i W_i \mathrm{Diag}\{\boldsymbol{e}_i\}$, nous avons :

$$
\begin{aligned}
N^2 D_{(1)} &= \mathrm{Var}_p \left(\sum_{i \in U_{1(1)}} C_i \boldsymbol{I}_{i(1)} \right) = \mathrm{Var} \left\{ E \left[\sum_{i \in U_{1(1)}} C_i \boldsymbol{I}_{i(1)} \mid s_{1(1)} \right] \right\} \\
&\quad + E \left\{ \mathrm{Var} \left[\sum_{i \in U_{1(1)}} C_i \boldsymbol{I}_{i(1)} \mid s_{1(1)} \right] \right\} \\
&= \mathrm{Var} \left\{ E \left[E \left(\sum_{i \in U_{1(1)}} C_i \boldsymbol{I}_{i(1)} \mid s_{2(1)}, s_{1(1)} \right) \mid s_{1(1)} \right] \right\} \\
&\quad + E \left\{ \mathrm{Var} \left[E \left(\sum_{i \in U_{1(1)}} C_i \boldsymbol{I}_{i(1)} \mid s_{2(1)}, s_{1(1)} \right) \mid s_{1(1)} \right] \right. \\
&\qquad \left. + E \left[\mathrm{Var} \left(\sum_{i \in U_{1(1)}} C_i \boldsymbol{I}_{i(1)} \mid s_{2(1)}, s_{1(1)} \right) \mid s_{1(1)} \right] \right\} \\
&= N^2 D_{(1)1} + N^2 D_{(1)2} + N^2 D_{(1)3}. \qquad \text{(A.2)}
\end{aligned}
$$

Occupons-nous de chacun des termes de (A.2) tour à tour; en commençant par $N^2 D_{1(1)}$, nous avons :

$$
E \left(\sum_{i \in U_{1(1)}} B_i W_i \mathrm{Diag}\{\boldsymbol{e}_i\} \boldsymbol{I}_{i(1)} \mid s_{2(1)}, s_{1(1)} \right) = \sum_{i \in U_{1(1)}} B_i W_i \mathrm{Diag}\{\boldsymbol{e}_i\} \boldsymbol{I}^{(1)}_{i(1)},
$$

où $\boldsymbol{I}^{(1)}_{i(1)} = (I_i(s_{1(1)}), I_i(s_{2(1)})I_i(s_{1(1)}), \pi_{i3|s_{2(1)}} I_i(s_{2(1)})I_i(s_{1(1)}))'$, alors

$$
\begin{aligned}
&E \left[E \left(\sum_{i \in U_{1(1)}} C_i \boldsymbol{I}_{i(1)} \mid s_{2(1)}, s_{1(1)} \right) \mid s_{1(1)} \right] \\
&\qquad = \sum_{i \in U_{1(1)}} C_i \boldsymbol{I}^{(2)}_{i(1)} = \sum_{i \in U_{1(1)}} B_i \mathrm{I}^{(1)}_i(\mathrm{U}) \mathrm{Diag}\{\boldsymbol{I}^{(2)}_{i(1)}\} \boldsymbol{e}_i \\
&\qquad = \sum_{i \in U_{1(1)}} F_i \left[\frac{I_i(s_{1(1)})}{\pi_{i1}}, \frac{I_i(s_{1(1)})}{\pi_{i1}}, \frac{I_i(s_{1(1)})}{\pi_{i1}} \right]' \\
&\qquad = \sum_{i \in U_{1(1)}} F_i \mathbf{1}_3 \frac{I_i(s_{1(1)})}{\pi_{i1}} = \sum_{i \in U_{1(1)}} w_{i1(1)} \boldsymbol{F}_{i(1)} I_i(s_{1(1)}),
\end{aligned}
$$

où $\boldsymbol{I}_{i(1)}^{(2)} = (I_i(s_{1(1)}),\ \pi_{i2|s_{1(1)}} I_i(s_{1(1)}),\ \pi_{i3|s_{2(1)}}\pi_{i2|s_{1(1)}} I_i(s_{1(1)}))'$, $\mathrm{I}_i^{(1)}(\mathrm{U}) = \mathrm{diag}[I_i(U_1)/\pi_{i1},\ I_i(U_2)/(\pi_{i1}\pi_{i2|s_{1(1)}}), I_i(U_3)/(\pi_{i1}\pi_{i2|s_{1(1)}}\pi_{i3|s_{2(1)}})]$, $F_i = B_i \mathrm{I}_i(\mathrm{U})\mathrm{Diag}\{\boldsymbol{e}_i\}$, et $\mathbf{1}_3 = (1,1,1)'$; cela implique que $N^2 D_{(1)1} = \mathrm{Var}\left[\sum_{i\in U_{1(1)}} w_{i1} B_i \mathrm{I}_i(\mathrm{U})\boldsymbol{e}_i I_i(s_{1(1)})\right] = \mathrm{Var}\left[\sum_{i\in s_{1(1)}} w_{i1}\boldsymbol{F}_{i(1)}\right]$.

Pour $N^2 D_{(1)2}$, nous avons :

$$
\begin{aligned}
&E\left(\sum_{i\in U_{1(1)}} C_i \boldsymbol{I}_{i(1)} \mid s_{2(1)}, s_{1(1)}\right) \\
&\qquad = \sum_{i\in U_{1(1)}} B_i W_i \mathrm{Diag}\{\boldsymbol{e}_i\}\boldsymbol{I}_{i(1)}^{(1)} = \sum_{i\in U_{1(1)}} B_i \mathrm{I}_i^{(1)}(\mathrm{U})\mathrm{Diag}\{\boldsymbol{I}_{i(1)}^{(3)}\}\boldsymbol{e}_i \\
&\qquad = \sum_{i\in U_{1(1)}} B_i \mathrm{I}_i(\mathrm{U})\mathrm{Diag}\{\boldsymbol{e}_i\}\left[\frac{I_i(s_{1(1)})}{\pi_{i1}}, \frac{I_i(s_{2(1)})I_i(s_{1(1)})}{\pi_{i1}\pi_{i2|s_{1(1)}}}, \frac{I_i(s_{2(1)})I_i(s_{1(1)})}{\pi_{i1}\pi_{i2|s_{1(1)}}}\right]' \\
&\qquad = \sum_{i\in s_{1(1)}} w_{i2} B_i \mathrm{I}_i(\mathrm{U})\mathrm{Diag}\{\boldsymbol{e}_i\}[\pi_{i2|s_{1(1)}}, I_i(s_{2(1)}), I_i(s_{2(1)})]',
\end{aligned}
$$

où $\boldsymbol{I}_{i(1)}^{(3)} = (I_i(s_{1(1)}), I_i(s_{2(1)})I_i(s_{1(1)}), \pi_{i3|s_{2(1)}} I_i(s_{2(1)})I_i(s_{1(1)}))'$; alors,

$$
\begin{aligned}
&\mathrm{Var}\left[E\left(\sum_{i\in U_{1(1)}} C_i \boldsymbol{I}_{i(1)} \mid s_{2(1)}, s_{1(1)}\right) \mid s_{1(1)}\right] \\
&\qquad = \mathrm{Var}\left[\sum_{i\in s_{1(1)}} w_{i2} B_i \mathrm{I}_i(\mathrm{U})\mathrm{Diag}\{\boldsymbol{e}_i\}\boldsymbol{I}_{i(1)}^{(4)} \mid s_{1(1)}\right] \\
&\qquad = \mathrm{Var}\left[\sum_{i\in s_{1(1)}} w_{i2} B_i \mathrm{I}_i(\mathrm{U})\mathrm{Diag}\{\boldsymbol{e}_i\}[0, I_i(s_{2(1)}), I_i(s_{2(1)})]' \mid s_{1(1)}\right] \\
&\qquad = \mathrm{Var}\left[\sum_{i\in s_{1(1)}} w_{i2} B_i \mathrm{I}_i(\mathrm{U})\mathrm{Diag}\{\boldsymbol{e}_i\} I_i(s_{2(1)})\mathbf{1}_{02} \mid s_{1(1)}\right] \\
&\qquad = \mathrm{Var}\left[\sum_{i\in s_{2(1)}} w_{i2} B_i \mathrm{I}_i(\mathrm{U})\boldsymbol{e}_{i(2\cdots 3)} \mid s_{1(1)}\right], \qquad \text{(A.3)}
\end{aligned}
$$

où $\boldsymbol{I}_{i(1)}^{(4)} = [\pi_{i2|s_{1(1)}}, I_i(s_{2(1)}), I_i(s_{2(1)})]', \mathbf{1}_{02} = (0,1,1)'$, et la ligne (A.3) est obtenue parce que, conditionnellement à $s_{1(1)}, \pi_{i2|s_{1(1)}}$ est constant et par conséquent la variance de cette composante est nulle. Cela signifie que :

$$
\begin{aligned}
N^2 D_{(1)2} &= E\left\{ \mathrm{Var}\left[E\left(\sum_{i\in U_{1(1)}} C_i \boldsymbol{I}_{i(1)} \mid s_{2(1)}, s_{1(1)} \right) \mid s_{1(1)} \right] \right\} \\
&= E\left\{ \mathrm{Var}\left[\sum_{i\in s_{2(1)}} w_{i2} \boldsymbol{F}_{i(2)} \mid s_{1(1)} \right] \right\} \\
&= \mathrm{Var}\left[\sum_{i\in s_{2(1)}} w_{i2} \boldsymbol{F}_{i(2)} \right] - \mathrm{Var}\left\{ E\left[\sum_{i\in s_{2(1)}} w_{i2} \boldsymbol{F}_{i(2)} \mid s_{1(1)} \right] \right\} \\
&= \mathrm{Var}\left[\sum_{i\in s_{2(1)}} w_{i2} \boldsymbol{F}_{i(2)} \right] - \mathrm{Var}\left\{ E\left[\sum_{i\in s_{2(1)}} w_{i2|s_{1(1)}} w_{i1} \boldsymbol{F}_{i(2)} \mid s_{1(1)} \right] \right\} \\
&= \mathrm{Var}\left[\sum_{i\in s_{2(1)}} w_{i2} \boldsymbol{F}_{i(2)} \right] - \mathrm{Var}\left\{ \sum_{i\in s_{1(1)}} w_{i1} \boldsymbol{F}_{i} \right\}.
\end{aligned}
$$

Nous pouvons, similairement, montrer que :

$$
\begin{aligned}
N^2 D_{(1)3} &= E\left\{ E\left[\mathrm{Var}\left(\sum_{i\in U_{1(1)}} B_i W_i \mathrm{Diag}\{\boldsymbol{e}_i\} \boldsymbol{I}_{i(1)} \mid s_{2(1)}, s_{1(1)} \right) \mid s_{1(1)} \right] \right\} \\
&= E\left\{ E\left[\mathrm{Var}\left(\sum_{i\in s_{3(1)}} w_{i3} I_i(s_{2(1)}) I_i(s_{1(1)}) \boldsymbol{F}_{i(3)} \mid s_{2(1)}, s_{1(1)} \right) \mid s_{1(1)} \right] \right\} \\
&= E\left\{ \mathrm{Var}\left[\sum_{i\in s_{3(1)}} w_{i3} I_i(s_{2(1)}) \boldsymbol{F}_{i(3)} \mid s_{1(1)} \right] \right. \\
&\qquad \left. -\mathrm{Var}\left[E\left(\sum_{i\in s_{3(1)}} w_{i3} I_i(s_{2(1)}) \boldsymbol{F}_{i(3)} \mid s_{2(1)}, s_{1(1)} \right) \mid s_{1(1)} \right] \right\} \\
&= E\left\{ \mathrm{Var}\left[\sum_{i\in s_{3(1)}} w_{i3} \boldsymbol{F}_{i(3)} \mid s_{1(1)} \right] - \mathrm{Var}\left[\sum_{i\in s_{2(1)}} w_{i2} \boldsymbol{F}_{i(3)} \mid s_{1(1)} \right] \right\} \\
&= \mathrm{Var}\left[\sum_{i\in s_{3(1)}} w_{i3} \boldsymbol{F}_{i(3)} \right] - \mathrm{Var}\left[E\left(\sum_{i\in s_{3(1)}} w_{i3} \boldsymbol{F}_{i(3)} \mid s_{1(1)} \right) \right] \\
&\quad -\mathrm{Var}\left[\sum_{i\in s_{2(1)}} w_{i2} \boldsymbol{F}_{i(3)} \right] + \mathrm{Var}\left[E\left(\sum_{i\in s_{2(1)}} w_{i2} \boldsymbol{F}_{i(3)} \mid s_{1(1)} \right) \right] \\
&= \mathrm{Var}\left[\sum_{i\in s_{3(1)}} w_{i3} \boldsymbol{F}_{i(3)} \right] - \mathrm{Var}\left[\sum_{i\in s_{1(1)}} w_{i1} \boldsymbol{F}_{i(3)} \right] \\
&\quad -\mathrm{Var}\left[\sum_{i\in s_{2(1)}} w_{i2} \boldsymbol{F}_{i(3)} \right] + \mathrm{Var}\left[\sum_{i\in s_{1(1)}} w_{i1} \boldsymbol{F}_{i(3)} \right].
\end{aligned}
$$

Au moyen de calculs similaires, nous obtenons les expressions correspondantes pour $N^2D_{(2)}, N^2D_{(2)2}, N^2D_{(2)3}$ et $N^2D_{(3)} = N^2D_{(3)3}$.

- Enfin, nous esquissons le développement d'une expression pour $\mathrm{Var}_p[\Psi_s(\boldsymbol{\beta}_N)]$ sans supposer que les cohortes sont indépendantes. Premièrement, notons que $\Psi_s(\boldsymbol{\beta}_N)$ peut s'écrire :

$$\sum_{i\in s_1} B_i \mathrm{I}_i(\mathrm{U}) \begin{bmatrix} w_{i1} & & \mathrm{O} \\ & w_{i2} & \\ \mathrm{O} & & w_{i3} \end{bmatrix} \begin{bmatrix} e_{i1} \\ e_{i2} \\ e_{i3} \end{bmatrix} + \sum_{i\in s_2} B_i \mathrm{I}_i(\mathrm{U}) \begin{bmatrix} 0 & & \mathrm{O} \\ & w_{i2} & \\ \mathrm{O} & & w_{i3} \end{bmatrix} \begin{bmatrix} 0 \\ e_{i2} \\ e_{i3} \end{bmatrix}$$

$$- \sum_{i\in s_1} B_i \mathrm{I}_i(\mathrm{U}) \begin{bmatrix} 0 & & \mathrm{O} \\ & w_{i2} & \\ \mathrm{O} & & w_{i3} \end{bmatrix} \begin{bmatrix} 0 \\ e_{i2} \\ e_{i3} \end{bmatrix}$$

$$+ \sum_{i\in s_3} B_i \mathrm{I}_i(\mathrm{U}) \begin{bmatrix} 0 & & \mathrm{O} \\ & 0 & \\ \mathrm{O} & & w_{i3} \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ e_{i3} \end{bmatrix} - \sum_{i\in s_2} B_i \mathrm{I}_i(\mathrm{U}) \begin{bmatrix} 0 & & \mathrm{O} \\ & 0 & \\ \mathrm{O} & & w_{i3} \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ e_{i3} \end{bmatrix}$$

$$= \sum_{i\in s_1} w_{i1} B_i \mathrm{I}_i(\mathrm{U}) \boldsymbol{e}_i - \sum_{i\in s_1} w_{i1} B_i \mathrm{I}_i(\mathrm{U}) \begin{bmatrix} 0 \\ e_{i2} \\ e_{i3} \end{bmatrix} + \sum_{i\in s_2} w_{i2} B_i \mathrm{I}_i(\mathrm{U}) \begin{bmatrix} 0 \\ e_{i2} \\ e_{i3} \end{bmatrix} - \sum_{i\in s_2} w_{i2} B_i \mathrm{I}_i(\mathrm{U}) \begin{bmatrix} 0 \\ 0 \\ e_{i3} \end{bmatrix}$$

$$+ \sum_{i\in s_3} w_{i3} B_i \mathrm{I}_i(\mathrm{U}) \begin{bmatrix} 0 \\ 0 \\ e_{i3} \end{bmatrix};$$

en posant que $\boldsymbol{z}_i = B_i \mathrm{I}_i(\mathrm{U})\, \boldsymbol{e}_i, \boldsymbol{z}_{i(2\cdots3)} = B_i \mathrm{I}_i(\mathrm{U})[0, e_{i2}, e_{i3}]'$ et $\boldsymbol{z}_{i(3\cdots3)} = B_i \mathrm{I}_i(\mathrm{U})[0, 0, e_{i3}]'$, $\mathrm{Var}_p[\Psi_s(\boldsymbol{\beta}_N)]$ peut être développée comme il suit :

$$\mathrm{Var}_p\left[\sum_{i\in s_1} w_{i1}\boldsymbol{z}_i\right] + \mathrm{Var}_p\left[\sum_{i\in s_1} w_{i1}\boldsymbol{z}_{i(2\cdots3)}\right]$$

$$+ \mathrm{Var}_p\left[\sum_{i\in s_2} w_{i2}\boldsymbol{z}_{i(2\cdots3)}\right] + \mathrm{Var}_p\left[\sum_{i\in s_2} w_{i2}\boldsymbol{z}_{i(3\cdots3)}\right]$$

$$+ \mathrm{Var}_p\left[\sum_{i\in s_3} w_{i3}\boldsymbol{z}_{i(3\cdots3)}\right] - 2\mathrm{Cov}_p\left[\sum_{i\in s_1} w_{i1}\boldsymbol{z}_i, \sum_{i\in s_1} w_{i1}\boldsymbol{z}_{i(2\cdots3)}\right]$$

$$+ 2\mathrm{Cov}_p\left[\sum_{i\in s_1} w_{i1}\boldsymbol{z}_i, \sum_{i\in s_2} w_{i2}\boldsymbol{z}_{i(2\cdots3)}\right]$$

$$- 2\mathrm{Cov}_p\left[\sum_{i\in s_1} w_{i1}\boldsymbol{z}_i, \sum_{i\in s_2} w_{i2}\boldsymbol{z}_{i(3\cdots3)}\right] + 2\mathrm{Cov}_p\left[\sum_{i\in s_1} w_{i1}\boldsymbol{z}_i, \sum_{i\in s_3} w_{i3}\boldsymbol{z}_{i(3\cdots3)}\right]$$

$$
\begin{aligned}
&- 2\mathrm{Cov}_p\left[\sum_{i\in s_1} w_{i1}\mathbf{z}_{i(2\cdots3)}, \sum_{i\in s_2} w_{i2}\mathbf{z}_{i(2\cdots3)}\right] + 2\mathrm{Cov}_p\left[\sum_{i\in s_1} w_{i1}\mathbf{z}_{i(2\cdots3)}, \sum_{i\in s_2} w_{i2}\mathbf{z}_{i(3\cdots3)}\right] \\
&- 2\mathrm{Cov}_p\left[\sum_{i\in s_1} w_{i1}\mathbf{z}_{i(2\cdots3)}, \sum_{i\in s_3} w_{i3}\mathbf{z}_{i(3\cdots3)}\right] - 2\mathrm{Cov}_p\left[\sum_{i\in s_2} w_{i2}\mathbf{z}_{i(2\cdots3)}, \sum_{i\in s_2} w_{i2}\mathbf{z}_{i(3\cdots3)}\right] \\
&+ 2\mathrm{Cov}_p\left[\sum_{i\in s_2} w_{i2}\mathbf{z}_{i(2\cdots3)}, \sum_{i\in s_3} w_{i3}\mathbf{z}_{i(3\cdots3)}\right] - 2\mathrm{Cov}_p\left[\sum_{i\in s_2} w_{i2}\mathbf{z}_{i(3\cdots3)}, \sum_{i\in s_3} w_{i3}\mathbf{z}_{i(3\cdots3)}\right] \\
= \ &\mathrm{Var}_p\left[\sum_{i\in s_1} w_{i1}\mathbf{z}_i\right] + \mathrm{Var}_p\left[\sum_{i\in s_2} w_{i2}\mathbf{z}_{i(2\cdots3)}\right] - \mathrm{Var}_p\left[\sum_{i\in s_1} w_{i1}\mathbf{z}_{i(2\cdots3)}\right] \\
&+ \mathrm{Var}_p\left[\sum_{i\in s_3} w_{i3}\mathbf{z}_{i(3\cdots3)}\right] - \mathrm{Var}_p\left[\sum_{i\in s_2} w_{i2}\mathbf{z}_{i(3\cdots3)}\right] \\
&+ 2\mathrm{Cov}_p\left[\sum_{i\in s_1} w_{i1}\mathbf{z}_{i(1\cdots1)}, \sum_{i\in s_2} w_{i2}\mathbf{z}_{i(2\cdots3)}\right] - 2\mathrm{Cov}_p\left[\sum_{i\in s_1} w_{i1}\mathbf{z}_{i(1\cdots1)}, \sum_{i\in s_1} w_{i1}\mathbf{z}_{i(2\cdots3)}\right] \\
&+ 2\mathrm{Cov}_p\left[\sum_{i\in s_1} w_{i1}\mathbf{z}_{i(1\cdots2)}, \sum_{i\in s_3} w_{i3}\mathbf{z}_{i(3\cdots3)}\right] - 2\mathrm{Cov}_p\left[\sum_{i\in s_1} w_{i1}\mathbf{z}_{i(1\cdots2)}, \sum_{i\in s_2} w_{i2}\mathbf{z}_{i(3\cdots3)}\right] \qquad \text{(A.4)} \\
&+ 2\mathrm{Cov}_p\left[\sum_{i\in s_1} w_{i1}\mathbf{z}_{i(2\cdots2)}, \sum_{i\in s_2} w_{i2}\mathbf{z}_{i(3\cdots3)}\right] - 2\mathrm{Cov}_p\left[\sum_{i\in s_1} w_{i1}\mathbf{z}_{i(2\cdots2)}, \sum_{i\in s_3} w_{i3}\mathbf{z}_{i(3\cdots3)}\right] \\
&+ 2\mathrm{Cov}_p\left[\sum_{i\in s_2} w_{i2}\mathbf{z}_{i(2\cdots2)}, \sum_{i\in s_3} w_{i3}\mathbf{z}_{i(3\cdots3)}\right] - 2\mathrm{Cov}_p\left[\sum_{i\in s_2} w_{i2}\mathbf{z}_{i(2\cdots2)}, \sum_{i\in s_2} w_{i2}\mathbf{z}_{i(3\cdots3)}\right].
\end{aligned}
$$

Dans cette dernière expression, la première chose que nous constatons est que *tous* les éléments diagonaux dans *tous* les termes de covariance sont exactement égaux à zéro; cela signifie que l'expression (3.13) est exacte pour les termes de variance, que les cohortes soient indépendantes ou non les unes des autres.

Pour analyser l'importance des termes de covariance, nous nous concentrons sur le terme de la ligne (A.4); la conclusion pour les autres termes est la même; notons que ce terme peut s'écrire :

$$
2\mathrm{Cov}_p\left[\sum_{i\in s_1} w_{i1}\frac{\partial\boldsymbol{\mu}_i'}{\partial\boldsymbol{\beta}}V_i^{-1}\begin{pmatrix} e_{i1}\\ e_{i2}\\ 0\end{pmatrix}, \sum_{i\in s_3} w_{i3}\frac{\partial\boldsymbol{\mu}_i'}{\partial\boldsymbol{\beta}}V_i^{-1}\begin{pmatrix} 0\\ 0\\ e_{i3}\end{pmatrix} - \sum_{i\in s_2} w_{i2}\frac{\partial\boldsymbol{\mu}_i'}{\partial\boldsymbol{\beta}}V_i^{-1}\begin{pmatrix} 0\\ 0\\ e_{i3}\end{pmatrix}\right];
$$

La propriété 3.1 énonce que, si les *cohortes* sont indépendantes sous le plan, tous les termes de covariance sont exactement égaux à zéro. En outre, de cette dernière

expression, nous concluons, trivialement, que si les *vagues* sont indépendantes sous le plan, tous les termes de covariance sont égaux à zéro également. Cette formule pour le terme de la ligne (A.4) implique aussi que, si les poids individuels ne varient pas fortement d'une vague à la suivante, et que le chevauchement entre vague consécutive est important, les termes de covariance ne sont pas trop grands. Enfin, si le chevauchement est faible, il est raisonnable de supposer l'indépendance des vagues sous le plan, et les termes de covariance peuvent alors être approximés sans risque par zéro.

Bibliographie

Ardilly, P., et Lavallée, P. (2007). Pondération dans les échantillons rotatifs : le cas de l'Enquête SILC en France. *Techniques d'enquête*, 33, 2, 149-156.

Berger, Y.G. (2004a). Variance estimation for change: An evaluation based upon the 2000 finnish labour force survey. Proceedings. European Conference on Quality and Methodology in Official Statistics.

Berger, Y.G. (2004b). Variance estimation for measures of change in probability sampling. *La Revue Canadienne de Statistique*, 32, 4, 451-467.

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *Revue Internationale de Statistique*, 51, 279-292.

Carrillo, I.A., Chen, J. et Wu, C. (2010). The pseudo-GEE approach to the analysis of longitudinal surveys. *La Revue Canadienne de Statistique*, 38, 4, 540-554.

Carrillo, I.A., Chen, J. et Wu, C. (2011). A pseudo-GEE approach to analyzing longitudinal surveys under imputation for missing responses. *Journal of Official Statistics*, 27, 2, 255-277.

Carrillo, I.A., et Karr, A.F. (2011). Combining cohorts in longitudinal surveys. Rapport technique 180, National Institute of Statistical Sciences, Research Triangle Park, NC. Adresse URL http://www.niss.org/sites/default/files/tr180.pdf.

Carrillo, I.A., et Karr, A.F. (2012). Estimating change with multi-cohort longitudinal surveys. En préparation.

Cox, B.G., Grigorian, K., Wang, R. et Harter, R. (2010). 2008 Survey of Doctorate Recipients Weighting Implementation Report, document préparé par la National Opinion Research Center (NORC) for the National Science Foundation (NSF).

Diggle, P., Heagerty, P., Liang, K.-Y. et Zeger, S. (2002). *Analysis of Longitudinal Data*, 2e Édition. Oxford University Press, New York.

Hedeker, D., et Gibbons, R.D. (2006). *Longitudinal Data Analysis*. Wiley Series in Probability and Statistics. New Jersey : John Wiley & Sons, Inc., Hoboken.

Hirano, K., Imbens, G.W., Ridder, G. et Rubin, D.B. (2001). Combining panel data sets with attrition and refreshment samples. *Econometrica*, 69, 6, 1645-1659.

Hu, F., et Kalbfleisch, J.D. (2000). The estimating function bootstrap (Pkg: P449-495). *La Revue Canadienne de Statistique*, 28, 3, 449-481.

Larsen, M.D., Qing, S., Zhou, B. et Foulkes, M.A. (2011). Calibration estimation and longitudinal survey weights: Application to the NSF Survey of Doctorate Recipients. *Proceedings of the Survey Research Method Section*, American Statistical Association, 1360-1374.

Liang, K.-Y., et Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73, 13-22.

Lohr, S. (2007). Recent developments in multiple frame surveys. *Joint Statistical Meeting of the American Statistical Association*, 3257-3264.

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, 2, 185-195.

National Science Foundation, National Center for Science and Engineering Statistics (2012). Survey of doctorate recipients. http://www.nsf.gov/statistics/srvydoctoratework/, consulté le 9 février 2012.

Nevo, A. (2003). Using weights to adjust for sample selection when auxiliary information is available. *Journal of Business & Economic Statistics*, 21, 1, 43-52.

Qualité, L., et Tillé, Y. (2008). Estimation de la précision d'évolutions dans les enquêtes répétées, application à l'Enquête suisse sur la valeur ajoutée. *Techniques d'enquête*, 34, 2, 193-201.

Rao, J.N.K., et Wu, C. (2010). Pseudo-empirical likelihood inference for multiple frame surveys. *Journal of the American Statistical Association*, 105, 492, 1494-1503.

Roberts, G., Binder, D., Kovačević, M., Pantel, M. et Phillips, O. (2003). Using an estimating function bootstrap approach for obtaining variance estimates when modelling complex health survey data. *Proceedings of the Survey Methods Section, Statistical Society of Canada*, Halifax.

Robins, J.M., Rotnitzky, A. et Zhao, L.P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90, 106-121.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.

Särndal, C.-E., Swensson, B. et Wretman, J. (1992). *Model Assisted Survey Sampling*. New York : Springer-Verlag.

Smith, P., Lynn, P. et Elliot, D. (2009). Sample design for longitudinal surveys. Dans *Methodology of Longitudinal Surveys*, (Éd., P. Lynn). Wiley, Chichester, Chapitre 2, 21-33.

Song, P.X.-K. (2007). *Correlated Data Analysis: Modeling, Analytics, and Applications*. Springer Series in Statistics. New York : Springer.

Steel, D., et McLaren, C. (2007). Design and analysis of repeated surveys. Discours principal, International Conference on Quality Management of Official Statistics, Corée.

Vieira, M.D.T. (2009). Analysis of Longitudinal Survey Data: Allowing for the Complex Survey Design in Covariance Structure Models. VDM Verlag.

Vieira, M.D.T., et Skinner, C.J. (2008). Estimating models for panel survey data under complex sampling. *Journal of Official Statistics*, 24, 3, 343-364.