Nº 12-001-X au catalogue
ISSN 1712-5685

Techniques d'enquête

Échantillonnage fondé sur des registres pour les panels auprès des ménages

par Jan A. van den Brakel

Date de diffusion : le 22 juin 2016

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 STATCAN.infostats-infostats.STATCAN@canada.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

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 « Contactez-nous » > « Normes de service à la clientèle ».

Note de reconnaissance

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

Signes conventionnels dans les tableaux

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)

Une version HTML est aussi disponible.

This publication is also available in English.

Techniques d'enquête, juin 2016
Vol. 42, nº 1, pp. 147-172
Statistique Canada, nº 12-001-X au catalogue

Échantillonnage fondé sur des registres pour les panels auprès des ménages

Jan A. van den Brakel[1]

Résumé

Aux Pays-Bas, les données statistiques sur le revenu et le patrimoine reposent sur deux grands panels auprès des ménages qui sont entièrement dérivés de données administratives. L'utilisation de ménages comme unités d'échantillonnage dans les plans de sondage des panels pose problème en raison de l'instabilité de ces unités au fil du temps. Les changements dans la composition des ménages influent sur les probabilités d'inclusion nécessaires aux méthodes d'inférence fondées sur le plan et assistées par modèle. Dans les deux panels auprès des ménages susmentionnés, ces problèmes sont surmontés par la sélection de personnes que l'on suit au fil du temps. À chaque période, les membres des ménages auxquels appartiennent les personnes choisies sont inclus dans l'échantillon. Il s'agit d'une méthode équivalente à un échantillonnage selon des probabilités proportionnelles à la taille du ménage, selon laquelle les ménages peuvent être sélectionnés plus d'une fois jusqu'à concurrence du nombre de membres du ménage. Dans le présent article, nous décrivons les propriétés de ce plan d'échantillonnage et les comparons avec la méthode généralisée du partage des poids pour l'échantillonnage indirect (Lavallée 1995, 2007). Les méthodes sont illustrées au moyen d'une application à la *Dutch Regional Income Survey*.

Mots-clés : Probabilités proportionnelles à la taille; échantillonnage indirect; pondération cohérente des personnes et des ménages; *Regional Income Survey*; méthode généralisée du partage des poids.

1 Introduction

Statistics Netherlands réalise deux grandes enquêtes par sondage pour recueillir des données sur le revenu et le patrimoine de la population néerlandaise. D'abord, la *Regional Income Survey* (RIS) brosse un portrait de la situation du revenu et du patrimoine à un niveau régional très détaillé. Des données exactes sur les répartitions des revenus par personne et par ménage au niveau des quartiers sont publiées annuellement, à partir d'un grand échantillon fondé sur un petit ensemble des principales composantes du revenu dérivées de manière assez directe à partir des données fiscales. Ensuite, des données sur le revenu annuel et les caractéristiques du patrimoine de la population néerlandaise sont publiées chaque année dans le cadre l'*Income Panel Survey* (IPS), à un niveau régional plus agrégé. Cette enquête est fondée sur un grand ensemble de variables faisant appel à toutes les composantes du revenu des ménages qu'il est possible de dériver à partir des données administratives disponibles aux Pays-Bas. Comme la dérivation des variables aux fins de cette enquête exige plus de temps, son échantillon est beaucoup plus petit que celui de la RIS. Les deux enquêtes sont conçues sous forme de panels auprès des ménages et permettent d'observer les variables sur le revenu et le patrimoine des personnes et des ménages.

Les ménages sont souvent considérés comme les unités d'échantillonnage dans les panels mis sur pied pour recueillir de l'information au niveau des ménages et des personnes (Lynn 2009; Smith, Lynn et Elliot 2009). De tels panels servent à la réalisation d'analyses longitudinales ainsi qu'à la production d'estimations transversales. L'utilisation des ménages comme unités d'échantillonnage dans le cadre d'une enquête par panel présente toutefois des inconvénients majeurs, en raison de l'instabilité des ménages au fil

1. Jan A. van den Brakel, Département des méthodes statistiques, Statistics Netherlands, C.P. 4481, 6401 CZ Heerlen, Pays-Bas et Département d'économie quantitative, Maastricht University School of Business and Economics, C.P. 616, 6200 MD, Maastricht, Pays-Bas. Courriel : ja.vandenbrakel@cbs.nl.

du temps. De fait, les ménages peuvent se défaire, se fusionner ou se séparer, de nouveaux membres peuvent s'y ajouter et d'autres les quitter pour différentes raisons. Kalton et Brick (1995) expliquent que ces changements peuvent influer sur les probabilités de sélection des ménages dans l'échantillon. La reconstruction des bonnes probabilités d'inclusion des unités d'échantillonnage est essentielle pour déterminer les poids à utiliser aux fins d'analyse, particulièrement si le panel sert à produire des estimations transversales.

Supposons un panel où les ménages sont sélectionnés par échantillonnage aléatoire simple, par exemple au moment $t = 0$. Dans beaucoup de panels, les personnes qui se joignent à un ménage échantillonné à une date ultérieure sont aussi incluses dans le panel. Lavallée (1995) appelle ces personnes des cohabitants. Au fil du temps, de plus en plus de cohabitants sont inclus dans l'échantillon et perturbent le plan d'échantillonnage à probabilités égales utilisé pour sélectionner l'échantillon initial (Kalton et Brick 1995). Prenons l'exemple du ménage A, sélectionné dans l'échantillon au moment de l'établissement du panel au temps $t = 0$. Si après un certain temps ce ménage fusionne avec le ménage B, qui n'a pas été sélectionné initialement pour le panel au temps $t = 0$, la probabilité de sélection de ce nouveau ménage correspond maintenant à la somme des probabilités de sélection des ménages A et B au temps $t = 0$. Le fait de ne pas corriger pour les différences dans les probabilités de sélection à cause de la croissance graduelle de la part des cohabitants dans l'échantillon donne lieu à une inférence biaisée. Ernst (1989) propose la méthode du partage des poids pour résoudre ce problème. Lavallée (1995) élargit cette solution à la méthode généralisée du partage des poids afin de faire des inférences à propos des populations cibles échantillonnées au moyen d'une base de sondage se rapportant à une population différente.

La RIS et l'IPS sont toutes deux des enquêtes par panel et sont réalisées afin de recueillir des données sur les ménages et les personnes. Afin d'éviter les problèmes associés au fait que les panels utilisent des ménages comme unités d'échantillonnage, un plan de sondage différent est utilisé. Au lieu de sélectionner des ménages, on sélectionne plutôt selon un plan d'échantillonnage à probabilités égales des « personnes principales », que l'on suit au fil du temps. Tous les membres du ménage auquel appartient une personne principale à chaque période particulière sont inclus dans l'échantillon. On obtient ainsi un plan d'échantillonnage en vertu duquel les ménages sont tirés proportionnellement à la taille du ménage et peuvent être sélectionnés plus d'une fois, jusqu'à concurrence du nombre de personnes dans le ménage. Ce plan est une application de l'échantillonnage indirect (Lavallée 1995, 2007; Deville et Lavallée 2006).

Le présent article décrit un plan d'échantillonnage assorti d'une technique d'estimation utile pour les panels recueillant des données au niveau de la personne et du ménage. La méthodologie employée est particulièrement utile dans le cas de l'échantillonnage fondé sur des registres, puisque les personnes principales sont incluses dans l'échantillon pendant une période indéterminée. Ce plan d'échantillonnage peut aussi servir aux panels en ligne, moyennant une forme quelconque de renouvellement afin de remédier au problème de l'attrition des participants. Cela signifie que les unités d'échantillonnage peuvent s'ajouter au panel, être observées plusieurs fois puis quitter le panel selon un schéma prédéterminé (Smith et coll. 2009). La principale contribution du présent article concerne la dérivation d'expressions explicites pour la variance des paramètres cibles à l'aide d'espérances d'inclusion plutôt que de probabilités d'inclusion en vertu du plan d'échantillonnage susmentionné. Une mesure de l'exactitude minimale pour une répartition

estimée du revenu est proposée et des expressions explicites pour déterminer la taille minimale d'échantillon sont dérivées. On utilise la RIS pour illustrer les techniques d'échantillonnage décrites.

L'article se présente comme suit. Le plan d'échantillonnage de la RIS est présenté à la section 2. À la section 3, on introduit le concept d'espérances d'inclusion comme solution de rechange pratique aux probabilités d'inclusion. Ensuite, les espérances d'inclusion de premier et de deuxième ordre sont dérivées pour le plan d'échantillonnage proposé. Ces espérances d'inclusion sont nécessaires pour construire l'estimateur π ou estimateur de Horvitz-Thompson (HT) (Narain 1951; Horvitz et Thompson 1952). On montre aussi que les mêmes poids peuvent être dérivés comme cas particulier de la méthode généralisée du partage des poids pour l'échantillonnage indirect (Lavallée 1995, 2007). Les variables cibles clés pour la RIS sont les répartitions estimées du revenu. À la section 4, les formules relatives à la taille minimale d'échantillon requise sont dérivées d'après une mesure de précision pour les répartitions estimées du revenu. Comme les ménages peuvent être sélectionnés plus d'une fois, une expression pour le nombre prévu de ménages uniques est dérivée à la section 4. La méthode d'estimation utilisée pour la RIS, fondée sur une pondération linéaire à l'aide de l'estimateur de régression généralisée (GREG) (Särndal, Swensson et Wretman 1992), est décrite à la section 5. La méthode de pondération intégrée de Lemaître et Dufour (1987), Nieuwenbroek (1993) et Steel et Clark (2007) est appliquée pour obtenir des poids égaux pour les personnes appartenant au même ménage. À la section 6, on dérive les approximations des variances pour l'estimateur GREG en vertu du plan d'échantillonnage proposé. Une application à la RIS est présentée à la section 7. L'article se termine à la section 8 par une discussion.

2 Plan d'échantillonnage

La population de la RIS comprend toutes les personnes physiques résidant aux Pays-Bas. La base de sondage est un registre de toutes les personnes physiques de 15 ans ou plus résidant aux Pays-Bas selon les dossiers du Bureau de l'impôt. À partir de ce registre, on tire un échantillon aléatoire simple stratifié de personnes dites « principales », selon une fraction de sondage de 0,16. Les quartiers servent de variable de stratification. Bien que l'on utilise un plan d'échantillonnage à probabilités égales, l'échantillonnage stratifié est utile pour éliminer la variation entre les strates et pour satisfaire aux exigences minimales en matière de précision pour chaque strate. Les Pays-Bas sont divisés en quelque 2 830 quartiers d'en moyenne 5 000 personnes de 15 ans ou plus.

La RIS est réalisée sous forme de panel depuis 1994. Pour garantir l'exactitude de l'inférence transversale à partir de ce panel, il faut d'abord établir avec justesse les espérances d'inclusion de premier et de deuxième ordre pour les unités d'échantillonnage; ces espérances sont dérivées à la section 3. Il faut ensuite s'assurer que le panel demeure représentatif de la population cible. Pour ce faire, on détermine annuellement quelle partie de la population est entrée dans la population cible de la RIS soit par naissance, soit par immigration. À partir de cette sous-population, on sélectionne un échantillon aléatoire simple stratifié de personnes principales selon une fraction de sondage de 0,16. Ces personnes principales sont ajoutées au panel de la RIS dans le but de maintenir un échantillon représentatif.

Les quartiers sont le niveau de publication le plus détaillé pour la RIS et sont donc utilisés comme strates. À la section 4, on dérive les expressions pour les tailles minimales des échantillons en fonction des

exigences de précision. Les personnes principales font partie du panel pendant une période indéterminée. À chaque période d'enquête, tous les membres des ménages des personnes principales sont aussi inclus dans l'échantillon. Les personnes qui quittent le ménage d'une personne principale sont éliminées du panel. Les nouvelles personnes qui se joignent au ménage d'une personne principale sont suivies dans le panel tant qu'elles font partie du ménage de la personne principale. Les données sur la composition des ménages des personnes principales sont obtenues auprès de la *Municipal Basis Administration* (MBA), le registre du gouvernement néerlandais recensant tous les résidents du pays. Les citoyens néerlandais sont tenus par la loi de déclarer aux municipalités tout changement d'ordre démographique. La MBA est utilisée conjointement avec les données provenant des autorités fiscales afin d'identifier les membres du ménage des personnes principales de l'échantillon.

Le plan d'échantillonnage permet d'établir un échantillon de ménages sélectionnés selon des probabilités proportionnelles au nombre de personnes de 15 ans ou plus appartenant au ménage à ce moment-là. Les ménages peuvent être sélectionnés plus d'une fois, jusqu'à concurrence du nombre de membres du ménage de 15 ans ou plus. Dans le présent article, l'expression « personne principale » désigne toute personne qui faisait partie de l'échantillon initial et fait l'objet d'un suivi au fil du temps dans le panel. Le mot « personnes » désigne l'échantillon obtenu si tous les membres du ménage à une période particulière sont inclus dans l'échantillon.

L'IPS repose sur un plan d'échantillonnage similaire, mais assorti d'une fraction de sondage beaucoup plus faible. La RIS et l'IPS sont toutes deux fondées sur des échantillons constitués à partir de registres, ce qui signifie que pour chaque personne incluse dans l'échantillon, les données nécessaires pour les variables de la RIS sont obtenues à partir des registres du Bureau de l'impôt. Les personnes principales et les membres de leurs ménages ne savent donc pas qu'elles font partie des échantillons. Cette méthode a pour avantage de ne poser aucun problème de non-réponse sélective ou d'attrition des participants au panel. Elle permet aussi d'inclure les personnes principales sur une période indéterminée. Dans le cas d'un panel où les unités d'échantillonnage doivent répondre à un questionnaire, il faut mettre en place un plan quelconque de renouvellement afin d'éliminer le biais de sélection attribuable à l'attrition. Par ailleurs, la méthode employée élimine les problèmes de biais de mesure associés à la collecte de données par questionnaire. Bien sûr, d'autres types d'erreurs de mesure se produisent lorsque l'enquête repose sur des registres (Wallgren et Wallgren 2007). Cela suppose notamment que tous les renseignements requis à propos du revenu pour estimer les paramètres cibles de la RIS et de l'IPS figurent dans les registres. Comme tous les renseignements requis se trouvent dans un registre, un dénombrement complet de la population est possible. Toutefois, par le passé, l'infrastructure de TI ne permettait pas de produire rapidement des données statistiques régionales sur le revenu pour toute la population des Pays-Bas. En conséquence, la RIS est réalisée par tradition sur un grand échantillon de personnes principales selon une fraction de sondage de 0,16. Pour la même raison, l'IPS repose par tradition sur un échantillon d'environ 80 000 personnes principales. Avec la capacité de calcul actuelle, un dénombrement complet serait possible mais tout de même très exigeant. La principale raison justifiant la réalisation de l'enquête à partir d'un échantillon est le maintien du panel aux fins des analyses longitudinales couvrant des périodes antérieures où un recensement n'était pas possible.

3 Poids d'inclusion

3.1 Pondération selon les espérances d'inclusion

Pour l'inférence fondée sur le plan de sondage, on a besoin des probabilités d'inclusion de premier et de deuxième ordre pour les ménages et les personnes. Soit M le nombre de ménages de la population, N le nombre de personnes de 15 ans et plus dans la population et g_k le nombre de personnes de 15 ans et plus appartenant au k^{e} ménage. Selon le plan d'échantillonnage décrit à la section 2, le ménage k peut être inclus plus d'une fois, jusqu'à concurrence de g_k fois. Cela complique la dérivation des probabilités d'inclusion, car la probabilité de sélectionner le ménage k est égale à la probabilité de sélection de l'union des membres du ménage (k, j) de 15 ans ou plus. Cette probabilité est définie comme suit :

$$\begin{aligned} P(k \in s) = P\left(\bigcup_{j=1}^{g_k}[(k, j) \in s]\right) &= \sum_{j=1}^{g_k} P((k, j) \in s) \\ &\quad - \sum_{j=1}^{g_k}\sum_{j'=j+1}^{g_k} P([(k, j) \cap (k, j')] \in s) \\ &\quad + \sum_{j=1}^{g_k}\sum_{j'=j+1}^{g_k}\sum_{j''=j+j'+1}^{g_k} P([(k, j) \cap (k, j') \cap (k, j'')] \in s) - \ldots \end{aligned}$$

On peut éviter ce type de calcul en utilisant le concept d'espérances d'inclusion, plutôt que les probabilités d'inclusion. Bethlehem (2009, chapitre 2) généralise l'estimateur HT au concept d'espérance d'inclusion pour un échantillonnage avec remise. Soit a_k le nombre de fois que le ménage k est sélectionné dans l'échantillon. Selon le plan d'échantillonnage proposé, $a_k \in [0, 1, \ldots, g_k]$. Soit $\text{E}(.)$ l'espérance relative au plan d'échantillonnage. Maintenant, $\pi_k = \text{E}(a_k)$ désigne l'espérance d'inclusion de l'unité d'échantillonnage k. Puisque a_k peut être plus grand que un, π_k peut aussi prendre une valeur supérieure à un et ne peut donc plus être interprétée comme une probabilité d'inclusion. Elle peut toutefois être interprétée comme une espérance.

Le paramètre d'intérêt est le total de population, défini par

$$t_y = \sum_{k=1}^{M}\sum_{j=1}^{N_k} y_{kj} \equiv \sum_{k=1}^{M} y_k. \tag{3.1}$$

L'estimateur HT du total de population en (3.1) peut être défini par

$$\hat{t}_y = \sum_{k=1}^{M} \frac{a_k y_k}{\pi_k}. \tag{3.2}$$

Puisque $\text{E}(a_k) = \pi_k$, il s'ensuit que cet estimateur HT est sans biais par rapport au plan. Soit $\pi_{kk'}$ l'espérance d'inclusion des unités k et k', c'est-à-dire $\pi_{kk'} = \text{E}(a_k a_{k'})$. Par définition, la variance de l'estimateur HT est égale à

$$\begin{aligned}
\mathrm{V}\left(\hat{t}_y\right) &= \sum_{k=1}^{M}\sum_{k'=1}^{M}\mathrm{Cov}\left(a_k a_{k'}\right)\frac{y_k}{\pi_k}\frac{y_{k'}}{\pi_{k'}} \\
&= \sum_{k=1}^{M}\sum_{k'=1}^{M}\left[\mathrm{E}\left(a_k a_{k'}\right) - \mathrm{E}\left(a_k\right)\mathrm{E}\left(a_{k'}\right)\right]\frac{y_k}{\pi_k}\frac{y_{k'}}{\pi_{k'}} \\
&= \sum_{k=1}^{M}\sum_{k'=1}^{M}\left(\pi_{kk'} - \pi_k\pi_{k'}\right)\frac{y_k}{\pi_k}\frac{y_{k'}}{\pi_{k'}}.
\end{aligned}$$

Soulignons que dans le cas d'un échantillonnage sans remise, a_k est une variable factice qui prend la valeur zéro ou un pour indiquer si l'unité k est sélectionnée dans l'échantillon. Dans ce cas, π_k et $\pi_{kk'}$ sont les probabilités d'inclusion de premier et de deuxième ordre habituelles. Cela montre que l'estimateur HT standard, fondé sur les probabilités d'inclusion, peut facilement être élargi aux espérances d'inclusion. Dans le cas des plans d'échantillonnage en vertu desquels les unités peuvent être sélectionnées plus d'une fois, il est plus commode de travailler avec des espérances d'inclusion, puisqu'elles peuvent être dérivées relativement facilement. Dans le reste de la présente sous-section, on dérive les espérances d'inclusion de premier et de deuxième ordre pour le plan d'échantillonnage décrit à la section 2.

Les personnes principales sont tirées à l'aide d'un échantillonnage aléatoire simple stratifié. Comme la stratification repose sur des régions géographiques, tous les membres d'un ménage k appartiennent à la même strate h au moment du tirage des personnes principales. Soit N_h le nombre de personnes de 15 ans ou plus dans la population de la strate h, n_h le nombre de personnes principales sélectionnées dans l'échantillon de la strate h et g_k le nombre de personnes de 15 ans ou plus appartenant au ménage k. Enfin, a_{jk} correspond à un indicateur égal à un si la personne j du ménage k est sélectionnée dans l'échantillon, et à zéro dans le cas contraire. L'espérance d'inclusion de premier ordre du k^{e} ménage égale

$$\pi_{kh} = \mathrm{E}\left(a_k\right) = E\left(\sum_{j=1}^{g_k} a_{jk}\right) = \sum_{j=1}^{g_k} E\left(a_{jk}\right) = g_k\,\frac{n_h}{N_h}. \tag{3.3}$$

Les espérances d'inclusion de deuxième ordre pour les ménages k et k' pour $k \neq k'$ appartenant à la même strate h égalent

$$\pi_{kk'} = \mathrm{E}\left(a_k a_{k'}\right) = E\left(\sum_{j=1}^{g_k} a_{jk}\sum_{j'=1}^{g_{k'}} a_{j'k'}\right) = \sum_{j=1}^{g_k}\sum_{j'=1}^{g_{k'}} E\left(a_{jk}a_{j'k'}\right) = g_k g_{k'}\,\frac{n_h\left(n_h-1\right)}{N_h\left(N_h-1\right)}. \tag{3.4}$$

L'espérance d'inclusion de deuxième ordre pour le ménage $k = k'$ pour la même strate h est donnée par

$$\begin{aligned}
\pi_{kk} &= \mathrm{E}\left(a_k a_k\right) = E\left(\sum_{j=1}^{g_k} a_{jk}\sum_{j'=1}^{g_k} a_{j'k}\right) = E\left(\sum_{j=1}^{g_k} a_{jk} + \sum_{j=1}^{g_k}\sum_{j'\neq j=1}^{g_k} a_{jk}a_{j'k}\right) \\
&= \sum_{j=1}^{g_k} E\left(a_{jk}\right) + \sum_{j=1}^{g_k}\sum_{j'\neq j=1}^{g_k} E\left(a_{jk}a_{j'k}\right) = g_k\,\frac{n_h}{N_h} + g_k\left(g_k-1\right)\frac{n_h\left(n_h-1\right)}{N_h\left(N_h-1\right)}.
\end{aligned} \tag{3.5}$$

Les espérances d'inclusion de deuxième ordre pour les ménages k et k' pour $k \neq k'$ appartenant à deux strates différentes h et h' égalent

$$\pi_{kk'} = \mathrm{E}\left(a_k a_{k'}\right) = \mathrm{E}\left(\sum_{j=1}^{g_k} a_{jk} \sum_{j'=1}^{g_{k'}} a_{j'k'}\right) = \sum_{j=1}^{g_k} \sum_{j'=1}^{g_{k'}} E\left(a_{jk} a_{j'k'}\right) = g_{kh} g_{k'h'} \frac{n_h n_{h'}}{N_h N_{h'}}. \tag{3.6}$$

Une autre preuve reposant sur la définition d'une espérance, qui ne fait pas appel à la règle voulant que l'espérance de la somme de variables mutuellement dépendantes soit égale à la somme des espérances de ces variables, est donnée par van den Brakel (2013).

Au fil du temps, la composition des ménages des personnes principales change, ce qui influe sur les espérances d'inclusion des ménages dans l'échantillon. Si les fractions de sondage diffèrent d'une strate à l'autre, les espérances d'inclusion (3.3) à (3.6) se complexifient et exigent de l'information sur l'appartenance à la strate pour toutes les personnes des ménages des personnes principales. On élimine cet inconvénient en choisissant un plan d'échantillonnage autopondéré. Dans ce cas, chaque membre du ménage d'une personne principale a la même probabilité d'inclusion et le seul renseignement propre au ménage nécessaire pour dériver les espérances d'inclusion du ménage est le nombre de personnes de 15 ans ou plus dans le ménage de la personne principale.

Comme tous les membres d'un ménage sélectionné sont inclus dans l'échantillon, il s'ensuit que les espérances d'inclusion de premier ordre pour les personnes appartenant au ménage k sont égales aux espérances d'inclusion de premier ordre du ménage k définies en (3.3). Les espérances d'inclusion de deuxième ordre pour les personnes appartenant à deux ménages différents k et k' sont égales à (3.4) si les deux ménages appartiennent à la même strate ou à (3.6) s'ils appartiennent à des strates différentes. Les espérances d'inclusion de deuxième ordre pour les personnes d'un même ménage sont établies par (3.5).

Les examinateurs du comité de lecture ont soulevé la question à savoir si les espérances d'inclusion elles-mêmes avaient une variance dont il faut tenir compte dans la variance des estimateurs HT ou GREG lorsque ceux-ci sont fondés sur des espérances d'inclusion plutôt que sur des probabilités d'inclusion. Dans la population finie, chaque personne et chaque ménage a une espérance d'inclusion prédéterminée. Dans le cas des ménages observés dans l'échantillon, ces espérances peuvent être calculées avec exactitude et sans incertitude, puisqu'on dispose de tous les renseignements nécessaires pour évaluer leur vraie valeur. La substitution des probabilités d'inclusion par des espérances n'introduit donc pas une variance supplémentaire.

3.2 Méthode généralisée du partage des poids

Le plan d'échantillonnage décrit à la section 2 peut être considéré comme un cas particulier d'échantillonnage indirect (Lavallée 2007). L'échantillonnage indirect s'entend d'une situation où la population d'intérêt est échantillonnée à partir d'une base de sondage se rapportant à une population différente. Lavallée (1995) a mis au point la méthode généralisée du partage des poids pour établir les poids

à utiliser dans une telle situation; cette méthode peut servir à dériver les poids de sondage pour les ménages et les personnes du plan d'échantillonnage décrit à la section 2.

Selon la notation de Lavallée (1995) pour les cas d'échantillonnage indirect, il y a une population U^A de taille N^A à partir de laquelle un échantillon s^A de taille n est tiré selon les probabilités de sélection π_i^A. De plus, il y a une population cible U^B de taille N^B. Cette population peut être divisée en M^B grappes. Chaque grappe k contient N_k^B unités, de sorte que $N^B = \sum_{k=1}^{M^B} N_k^B$. La situation du plan d'échantillonnage décrit à la section 2 est illustrée à la figure 3.1. Les grappes sont les ménages, U^A est la population des personnes de 15 ans ou plus et U^B est la population de toutes les personnes résidant aux Pays-Bas. Les personnes appartenant à U^A et à U^B sont représentées par des cercles et les ménages appartenant à U^B, par des carrés gris; les cercles se trouvant dans un carré gris représentent les personnes appartenant à un même ménage. La figure 3.1 montre respectivement un ménage d'une seule personne, un ménage de deux personnes formé par exemple d'un parent divorcé et d'un enfant de moins de 15 ans, un ménage de deux personnes formé de deux adultes sans enfant et un ménage de quatre personnes formé de deux parents et de deux enfants, l'un ayant moins de 15 ans, et l'autre, plus de 15 ans. Les flèches représentent les liens entre les unités de U^A et U^B. Dans le plan d'échantillonnage présenté à la section 2, chaque unité de U^A a exactement un seul lien avec une unité de U^B. Les grappes de U^B ont au moins un lien avec les unités de U^A. Les liens sont définis par une variable indicatrice

$$l_{ij} = \begin{cases} 1 & \text{s'il y a un lien entre } i \in U^A \text{ et } j \in U^B \\ 0 & \text{s'il n'y a pas de lien entre } i \in U^A \text{ et } j \in U^B. \end{cases}$$

Si une unité i de U^A est sélectionnée dans l'échantillon, toute la grappe k à laquelle cette unité appartient est incluse dans l'échantillon. Le paramètre d'intérêt est le total de population de U^B; il se compare à (3.1) et est défini par $t_y = \sum_{k=1}^{M^B} \sum_{j=1}^{N_k^B} y_{kj}$. Un estimateur de t_y est défini par

$$\hat{t}_y = \sum_{k=1}^{m} \sum_{j=1}^{N_k^B} w_{kj} y_{kj}, \tag{3.7}$$

avec m le nombre de grappes uniques (ménages) incluses dans l'échantillon et w_{kj} le poids associé à chaque unité j de la grappe k. En règle générale, on se sert de l'inverse des probabilités de sélection des unités (k, j) observées dans l'échantillon pour établir les poids dans l'estimateur HT. Dans ce cas, les unités de l'échantillon n'ont pas toutes une probabilité d'inclusion connue. D'abord, les unités de U^B n'ont pas toutes un lien avec une unité de U^A. Ensuite, la composition des ménages change au fil du temps en raison des mariages, des divorces, du départ des enfants et de la cohabitation. En conséquence, au fil du temps, les unités liées à U^A sont intégrées dans les grappes de l'échantillon même si elles ne faisaient pas initialement partie de l'échantillon tiré à partir de U^A. Même si leurs probabilités d'inclusion ne sont pas nécessairement connues, ces unités influent sur les espérances d'inclusion des grappes dans l'échantillon. Pour reconstruire les probabilités d'inclusion, il faut disposer de données sur les probabilités de sélection de toutes les unités de la population au moment du tirage de l'échantillon. Dans la pratique, on ne dispose généralement pas de cette information.

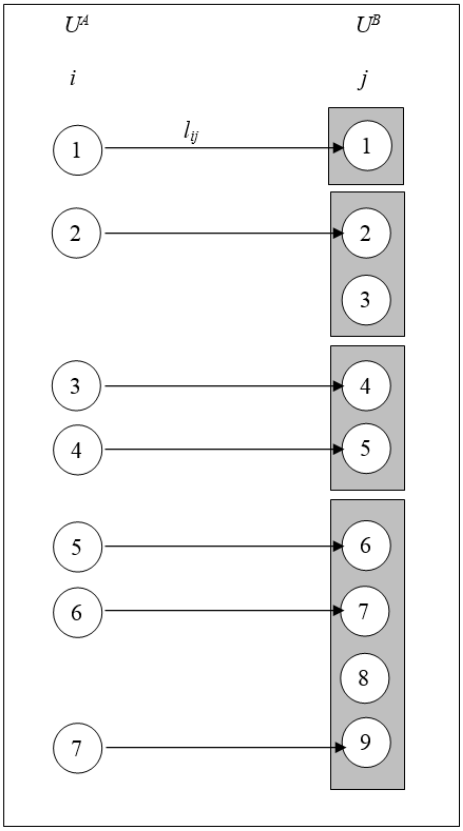

Figure 3.1 Liens entre les unités de la base de sondage et les unités de la population cible.

La méthode généralisée du partage des poids peut servir à dériver des poids non nuls pour toutes les unités de l'échantillon. Pour ce faire, il faut d'abord dériver les poids initiaux, définis par

$$w_{kj}^{*} = \begin{cases} \dfrac{\delta_i^A}{\pi_i^A} & \text{si } (k, j) \text{ a un lien avec } i \in U^A \\ 0 & \text{autrement} \end{cases},$$

avec δ_i^A une variable indicatrice égale à un si i est inclus dans l'échantillon s^A et à zéro autrement. Cette expression découle directement de Lavallée (1995), équation (2), ainsi que du fait que dans la présente application, chaque unité de U^A a exactement un seul lien avec une unité de U^B (voir la figure 3.1). Ensuite, un poids dit « de base » pour chaque grappe k est dérivé et correspond à la moyenne de tous les poids initiaux de chaque grappe :

$$w_k = \frac{\sum_{j=1}^{N_k^B} w_{kj}^{*}}{\sum_{j=1}^{N_k^B} l_{kj}},$$

qui découle de Lavallée (1995), équation (7). Enfin, toutes les personnes j appartenant au même ménage k reçoivent le même poids affecté à leur ménage, c'est-à-dire $w_{kj} = w_k$ pour tout $j \in k$. Une preuve que le recours à des poids de base en (3.7) constitue un estimateur non biaisé du total de population est aussi donnée par Lavallée (1995).

Soit $\sum_{j=1}^{N_k^B} l_{kj} = g_k$ le nombre de personnes de 15 ans ou plus du ménage k et a_k le nombre de personnes principales du ménage k, c'est-à-dire le nombre de personnes du ménage k faisant partie de l'échantillon s^A. Puisque s^A est tiré au moyen d'un échantillonnage aléatoire simple stratifié, il s'ensuit que $\pi_i^A = n_h^A / N_h^A$ avec N_h^A le nombre de personnes de 15 ans ou plus dans la population de la strate h, et n_h^A le nombre de personnes principales sélectionnées dans l'échantillon à partir de la strate h. Il s'ensuit que

$$w_k = \frac{a_k}{g_k} \frac{N_h^A}{n_h^A}. \tag{3.8}$$

Si l'on intègre l'espérance d'inclusion de premier ordre (3.3) dans (3.2), on obtient le même estimateur HT que celui qui est dérivé à partir de la méthode généralisée du partage des poids, c'est-à-dire en intégrant (3.8) dans (3.7).

Le calcul des espérances d'inclusion à la sous-section 3.1 s'applique à l'échantillonnage stratifié des ménages avec espérances d'inclusion proportionnelles à la taille du ménage et constitue un cas particulier de la méthode généralisée du partage des poids. Le fait qu'un échantillonnage des ménages proportionnel à la taille du ménage est efficient pour les variables cibles corrélées positivement avec la taille du ménage constitue un argument en faveur de l'utilisation d'un plan comme celui qui est décrit à la section 2.

Lavallée (1995) fournit aussi des expressions de la variance pour (3.7) fondées sur la méthode généralisée du partage des poids. Ces expressions reposent sur les probabilités d'inclusion de premier et de deuxième ordre des unités de l'échantillon tirées de U^A et sur une transformation de la variable cible. Par conséquent, le fait que la probabilité de sélection des grappes soit proportionnelle à leur taille n'est pas explicite, pas plus que le fait qu'elles soient tirées partiellement avec remise. On souligne à la section 6 que les expressions de la variance de Lavallée (1995) pour cette application sont égales aux expressions de la variance fondées sur les espérances d'inclusion calculées aux expressions (3.3) à (3.6).

4 Détermination de la taille de l'échantillon

La RIS a pour objet de publier les répartitions du revenu des ménages et des personnes à différentes échelles géographiques. Les répartitions du revenu des ménages pour une région ou une zone r sont définies par

$$P_{lr} = \frac{M_{lr}}{M_{+r}}, \qquad l = 1, \ldots, L, \tag{4.1}$$

où M_{lr} correspond au nombre de ménages dans la région r appartenant à la l^{e} catégorie de revenus et $M_{+r} = \sum_l M_{lr}$, le nombre total de ménages dans la région r. Cette répartition du revenu est estimée par

$$\hat{P}_{lr} = \frac{\hat{M}_{lr}}{M_{+r}}, \qquad l = 1, \ldots, L, \tag{4.2}$$

où $\hat{M}_{lr}$ correspond à un estimateur direct approprié du nombre total de ménages dans la région r appartenant à la l^{e} catégorie de revenus. Pour le moment, l'estimateur HT est présumé être un estimateur approprié de M_{lr}, c'est-à-dire

$$\hat{M}_{lr} = \sum_{h \in r} \sum_{k=1}^{m_h} \frac{y_{khl}}{\pi_k},$$

où $y_{khl} = 1$ si le ménage k de la strate h appartient à la l^{e} catégorie de revenus et $y_{khl} = 0$ autrement, et où m_h correspond au nombre total de ménages sélectionnés dans la strate h. Pour la RIS, $L = 10$. Les répartitions du revenu des personnes sont définies et estimées comme en (4.1) et (4.2), où M_{lr} correspond au nombre de personnes de la région r appartenant à la l^{e} catégorie de revenus. L'estimateur HT pour M_{lr} correspond maintenant à

$$\hat{M}_{lr} = \sum_{h \in r} \sum_{k=1}^{m_h} \frac{1}{\pi_k} \sum_{j=1}^{N_k} y_{kjhl},$$

où $y_{kjhl} = 1$ si la personne j du ménage k et de la strate h appartient à la l^{e} catégorie de revenus et $y_{kjhl} = 0$ autrement.

Pour déterminer la taille de l'échantillon, des spécifications précises pour les répartitions estimées du revenu sont nécessaires. Dans le cas des plans d'échantillonnage stratifié, les répartitions de Neyman sont souvent considérées pour déterminer les tailles minimales des échantillons et les répartitions optimales pour satisfaire aux exigences en matière de précision aux niveaux agrégés (Cochran 1977). Les répartitions exponentielles sont utiles pour trouver le juste équilibre entre les exigences de précision pour les agrégats et les strates (Bankier 1988). Dans la présente application, la taille minimale d'échantillon est fondée sur les exigences de précision pour les strates individuelles, c'est-à-dire les quartiers, qui constituent le niveau de publication le plus détaillé.

Si les exigences de précision sont spécifiées pour les catégories distinctes des répartitions du revenu, alors la catégorie de revenu ayant la plus grande variance de population détermine la taille minimale de l'échantillon requis, ce qui donne des tailles d'échantillon inutilement grandes. Comme solution de rechange, on propose d'utiliser plutôt la racine carrée de la moyenne des variances des catégories estimées des revenus d'une répartition du revenu comme mesure de précision pour les répartitions estimées des revenus. Avec cette mesure, l'influence de la catégorie de revenus la moins précise sur la taille minimale de l'échantillon est réduite. La racine carrée de la moyenne des variances des catégories estimées des revenus d'une répartition du revenu s'appelle la mesure de l'erreur type moyenne et est définie par

$$s = \sqrt{\frac{1}{L}\sum_{l=1}^{L} V\left(\hat{P}_{lr}\right)}. \tag{4.3}$$

Dans la présente section, on calcule une expression exacte pour s et on établit une approximation qui peut servir à estimer la taille minimale d'échantillon requise qui n'exige pas que l'on dispose de données à propos des répartitions du revenu ou des variances.

Comme les quartiers sont les régions les plus détaillées pour lesquelles les répartitions du revenu sont publiées, les exigences de précision pour la détermination de la taille d'échantillon sont spécifiées à ce niveau. Puisque les quartiers sont utilisés comme variable de stratification dans le plan d'échantillonnage, les expressions pour s peuvent être dérivées en vertu d'un échantillonnage aléatoire simple sans remise des personnes principales dans chaque quartier. On trouve en annexe la preuve qu'une expression de la mesure de l'erreur type moyenne s_h en (4.3) pour une répartition du revenu est donnée par

$$s_h = \sqrt{\frac{1}{L}\frac{N_h - n_h}{n_h}\frac{1}{N_h - 1}\left(\frac{N_h}{M_h^2}\sum_{l=1}^{L}\sum_{k=1}^{M_h}\frac{y_{khl}}{g_{kh}} - \sum_{l=1}^{L}\left(\frac{M_{lh}}{M_h}\right)^2\right)}, \tag{4.4}$$

avec M_h le nombre de ménages dans la strate h et M_{lh} le nombre de ménages de la strate h appartenant à la l^{e} catégorie de revenu. Soulignons que si $g_{kh} = 1$ pour tous les ménages de la population de la strate h, il s'ensuit que $M_h = N_h$ et la formule (4.1) est simplifiée comme suit :

$$V\left(\hat{P}_{lh}\right) = \frac{N_h - n_h}{n_h}\frac{1}{N_h - 1}\left(P_{lh}\left(1 - P_{lh}\right)\right),$$

ce qui correspond à la variance d'une fraction estimée en vertu d'un échantillonnage aléatoire simple sans remise (Cochran 1977, chapitre 3).

Le calcul des exigences quant à la taille minimale d'échantillon selon (4.4) exige des données sur la répartition du revenu et ses variances des périodes antérieures. Puisque ces données ne sont généralement pas disponibles à l'étape de la conception d'un panel, il est utile de fixer une borne supérieure pour la mesure de l'erreur type moyenne pour la répartition du revenu dans l'équation (4.4). Cela revient à considérer la variance comme un paramètre défini sous forme de proportion, qui atteint un maximum lorsque la proportion est de 0,5, pour calculer la taille minimale de l'échantillon requise pour une enquête. On présente en annexe la preuve qu'une borne supérieure pour la mesure de l'erreur type moyenne s_h relative à une répartition du revenu précisée en (4.4) est donnée par

$$s_h \leq \sqrt{\frac{1}{L}\frac{N_h - n_h}{n_h}\frac{1}{N_h - 1}\left(\frac{N_h}{M_h^2}\sum_{t=1}^{T}\frac{M_{th}}{t} - \frac{1}{L}\right)}, \tag{4.5}$$

avec M_{th} le nombre de ménages de taille t dans la strate h.

Si $g_{kh} = 1$ pour tous les ménages de la population de la strate h et si le nombre de catégories de la répartition du revenu $L = 2$, il s'ensuit que l'approximation de la mesure de l'erreur type moyenne s_h en (4.5) peut être simplifiée comme suit :

$$s_h \leq \sqrt{\frac{N_h - n_h}{n_h}\frac{1}{(N_h - 1)}\frac{1}{4}},$$

ce qui est égal à la racine carrée de la variance maximale d'une fraction estimée à $\hat{P} = 0{,}5$ sous un échantillonnage aléatoire simple. Ainsi, l'approximation de la mesure de l'erreur type moyenne de l'équation (4.5) peut être interprétée comme une généralisation de l'approximation de la variance maximale d'une fraction estimée à $\hat{P} = 0{,}5,$ souvent utilisée pour la détermination de la taille d'échantillon. La mesure de l'erreur type moyenne atteint sa valeur maximale dans le cas d'une répartition égale des ménages dans les catégories de revenus, c'est-à-dire $\hat{P}_{lh} = 1/L$ pour $l = 1, \ldots, L.$ Dans ce cas, l'approximation de s_h est exacte, ce qui découle directement de l'équation (4.3).

En fixant l'expression de s_h en (4.5) à une valeur maximale prédéterminée, par exemple $\Delta_h,$ on obtient l'expression suivante pour la taille minimale d'échantillon des personnes principales :

$$n_h \geq \frac{\left(\frac{N_h}{M_h}\right)^2 \sum_{t=1}^{T} \frac{M_{th}}{t} - \frac{N_h}{L}}{(N_h - 1)L\Delta_h^2 + \frac{N_h}{M_h^2}\sum_{t=1}^{T}\frac{M_{th}}{t} - \frac{1}{L}}. \tag{4.6}$$

Les données nécessaires pour estimer la taille minimale d'échantillon sont le nombre total de personnes et le nombre total de ménages de même taille dans les quartiers. On n'a besoin d'aucun renseignement sur la répartition prévue du revenu ou sur sa variance. Des estimations plus précises de la taille minimale d'échantillon peuvent être obtenues au moyen de l'expression (4.4), mais il faut pour cela disposer de données sur les répartitions du revenu, par exemple celles des périodes antérieures.

L'expression (4.6) donne la taille minimale d'échantillon pour les personnes principales. Par la suite, tous les membres du ménage de chaque personne principale sont inclus dans l'échantillon. Un même ménage peut donc être inclus plus d'une fois dans l'échantillon et la taille d'échantillon en termes de ménages uniques et de personnes uniques est aléatoire. Pour planifier une enquête et en gérer les coûts, il faut connaître le nombre espéré de ménages et de personnes uniques que l'on obtient si on tire un échantillon de personnes principales de taille $n_h.$ On trouve en annexe la preuve montrant que le nombre espéré de ménages uniques dans un échantillon de n_h personnes principales, tiré par échantillonnage aléatoire simple sans remise à partir d'une population finie de taille $N_h,$ est donné par

$$D_h = \sum_{t=1}^{T} M_{th}\left(1 - \frac{\prod_{i=0}^{t-1}(N_h - n_h - i)}{\prod_{i=0}^{t-1}(N_h - i)}\right). \tag{4.7}$$

Le nombre espéré de personnes uniques dans un échantillon de n_h personnes principales, tiré par échantillonnage aléatoire simple sans remise à partir d'une population finie de taille $N_h,$ découle directement de l'équation (4.7) et est donné par

$$D_h^{[p]} = \sum_{t=1}^{T} tM_{th} \left(1 - \frac{\prod_{i=0}^{t-1} (N_h - n_h - i)}{\prod_{i=0}^{t-1} (N_h - i)} \right). \tag{4.8}$$

Comme les nombres espérés de ménages et de personnes uniques sont des variables aléatoires, il est utile d'avoir une mesure d'incertitude pour ces valeurs espérées. Les expressions de la variance pour (4.7) et (4.8) ne peuvent toutefois pas être déterminées simplement et doivent donc faire l'objet d'autres recherches.

Les calculs relatifs à la taille d'échantillon sont effectués au niveau des quartiers. Il a été décidé de sélectionner les personnes principales selon une fraction de sondage de 0,16. Avec un échantillon de cette taille, la valeur maximale de la mesure de l'erreur type moyenne s_h au niveau des quartiers s'établit à environ 0,01 pour les répartitions estimées du revenu des ménages. Comme la population totale est d'environ 12 millions de personnes, on obtient un échantillon d'environ 2,1 millions de personnes principales, et un échantillon espéré d'environ 4,6 millions de personnes uniques. Cet échantillon a été tiré en 1994, l'année où le panel de la RIS néerlandaise a été mis sur pied.

5 Pondération linéaire

Pour des enquêtes auprès des ménages comme la RIS, il faut établir des estimations pour les caractéristiques des personnes et pour celles des ménages. Soit t_y la valeur totale de la variable cible y. Selon une pondération linéaire, un estimateur pour une variable cible fondée sur les personnes se définit comme suit :

$$\hat{t}_y = \sum_{h=1}^{H} \sum_{k \in 1}^{m_h} \sum_{j \in k} w_{kj} y_{kjh}, \tag{5.1}$$

avec y_{kjh} la valeur de la variable cible pour les personnes (k, j, h) et w_{kj} un poids pour la personne j appartenant au ménage k. Un estimateur de la variable cible fondée sur les ménages est donné par

$$\hat{t}_y = \sum_{h=1}^{H} \sum_{k=1}^{m_h} w_k y_{kh}, \tag{5.2}$$

avec y_{kh} la valeur de la variable cible pour le ménage k de la strate h et w_k un poids pour le ménage correspondant.

Les poids sont obtenus au moyen de l'estimateur GREG afin d'utiliser les variables auxiliaires observées dans l'échantillon et pour lesquelles on connaît les totaux de population grâce à d'autres sources (Särndal et coll. 1992). En conséquence, les poids reflètent les espérances (inégales) d'inclusion des unités d'échantillonnage et un ajustement faisant en sorte que pour les variables auxiliaires, la somme des observations pondérées corresponde aux totaux de population connus. Des variables catégoriques comme le sexe, l'âge, l'état matrimonial ou la région sont souvent utilisées comme variables auxiliaires. Comme les valeurs des variables auxiliaires varient d'une personne à l'autre dans le même ménage, différents poids

peuvent être calculés pour les personnes d'un même ménage. Pour s'assurer que la relation entre les variables du ménage et les variables des personnes est prise en compte dans les totaux estimés, il convient d'appliquer une méthode de pondération qui attribue un poids de ménage unique à tous les membres du ménage. Si les poids pour les personnes d'un ménage sont les mêmes, alors les estimations des mêmes variables cibles fondées sur le ménage et sur les personnes sont cohérentes entre elles (par exemple le revenu total estimé du ménage et celui des personnes). Pour ce faire, on peut utiliser des méthodes de pondération dites « intégrées ».

Lemaître et Dufour (1987) appliquent une méthode de pondération intégrée au niveau de la personne et remplacent les variables auxiliaires originales définies à ce niveau par la moyenne du ménage correspondant. Ainsi, les membres du même ménage ont la même espérance d'inclusion et partagent les mêmes données auxiliaires, ce qui fait que les poids de régression qui en découlent sont aussi forcément les mêmes. Nieuwenbroek (1993) propose une approche légèrement plus générale et applique une méthode de pondération linéaire au niveau du ménage, en vertu de laquelle les données auxiliaires sur les caractéristiques des personnes sont agrégées au niveau du ménage. Nieuwenbroek (1993) souligne que la méthode de pondération linéaire au niveau du ménage est équivalente à la méthode de pondération linéaire de Lemaître et Dufour (1987) au niveau de la personne, si la variance résiduelle du modèle de régression au niveau du ménage est choisie proportionnelle au nombre de personnes dans le ménage. Steel et Clark (2007) et Estevao et Särndal (2006) généralisent encore davantage la pondération intégrée dans les enquêtes auprès des personnes et des ménages. Steel et Clark (2007) ont étudié la question de savoir si les avantages cosmétiques de la pondération intégrée entraînent une variance accrue par rapport au plan dans les estimations GREG. Ils montrent que pour de grands échantillons, les variances par rapport au plan obtenues lorsqu'on applique une pondération linéaire au niveau du ménage sont inférieures ou égales à la variance par rapport au plan obtenue lorsqu'on applique une pondération linéaire au niveau de la personne. Pour de petits échantillons, il arrive que la pondération intégrée entraîne une légère augmentation de la variance par rapport au plan. On perd donc peu ou pas en efficience lorsqu'on applique une méthode de pondération intégrée.

Dans la présente étude, on applique la méthode de pondération intégrée au niveau du ménage. Soit $\mathbf{x}_{kh}$ un q-vecteur comprenant q variables auxiliaires pour le ménage k de la strate h. Les caractéristiques fondées sur la personne sont agrégées aux totaux pour le ménage. L'estimateur GREG est dérivé à partir d'un modèle de régression linéaire qui précise la relation entre la variable cible et les variables auxiliaires disponibles pour lesquelles les totaux de population sont connus et est défini par

$$y_{kh} = \mathbf{x}_{kh}^t \boldsymbol{\beta} + e_{kh}, \quad \text{avec} \quad \mathrm{E}_m(e_{kh}) = 0, \quad \mathrm{V}_m(e_{kh}) = \sigma_{kh}^2. \tag{5.3}$$

Dans (5.3), $\boldsymbol{\beta}$ désigne un vecteur comprenant les q coefficients de régression de la régression de y_{kh} sur $\mathbf{x}_{kh}$, e_{kh} désigne les résidus et E_m et V_m désignent l'espérance et la variance à l'égard du modèle de régression. Dans cette application, la structure de variance est considérée proportionnelle à la taille du ménage, c'est-à-dire $\sigma_{hk}^2 = g_k \sigma^2$. Nieuwenbroek (1993) montre que dans un tel cas, la pondération appliquée au niveau du ménage est égale à celle de la méthode de Lemaître et Dufour (1987).

Les poids de régression pour les ménages sont finalement obtenus par

$$w_k = \frac{1}{\pi_k}\left(1 + \left(\mathbf{t}_{\mathbf{x}} - \hat{\mathbf{t}}_{\mathbf{x}\boldsymbol{\pi}}\right)^t \left(\sum_{k=1}^{m} \frac{\mathbf{x}_{kh}\mathbf{x}_{kh}^t}{\pi_k g_k}\right)^{-1} \frac{\mathbf{x}_{kh}}{g_k}\right),$$

avec $\mathbf{t}_{\mathbf{x}}$ un vecteur q comprenant les totaux de population connus des variables auxiliaires $\mathbf{x}$, $\hat{\mathbf{t}}_{\mathbf{x}\boldsymbol{\pi}}$ l'estimateur HT pour $\mathbf{t}_{\mathbf{x}}$. Les poids calculés au niveau du ménage peuvent servir à pondérer les caractéristiques fondées sur la personne des membres du ménage correspondant, à l'aide de la formule énoncée en (5.1) puisque $w_{kj} = w_k$ pour toutes les personnes appartenant au même ménage k.

6 Estimation de la variance

Les paramètres de la RIS sont estimés sous forme de ratio de deux totaux de population :

$$\hat{R} = \frac{\hat{t}_y}{\hat{t}_z}, \tag{6.1}$$

où $\hat{t}_y$ et $\hat{t}_z$ sont des estimateurs GREG définis par (5.1) ou (5.2) selon que les variables cibles sont fondées sur la personne ou sur le ménage, respectivement. Une approximation de la variance de (6.1) par rapport à un plan d'échantillonnage où les personnes principales sont tirées par échantillonnage aléatoire simple stratifié et où tous les membres des ménages de ces personnes principales sont inclus dans l'échantillon peut être donnée par

$$V\left(\hat{R}\right) = \frac{1}{t_z^2}\sum_{h=1}^{H} \frac{N_h^2\left(1 - f_h\right)}{n_h} \frac{1}{N_h - 1}\sum_{k=1}^{N_h}\left(\frac{e_{kh}}{g_k} - \frac{1}{N_h}\sum_{k'=1}^{N_h}\frac{e_{k'h}}{g_{k'}}\right)^2, \tag{6.2}$$

où $f_h = n_h / N_h$, $e_{kh} = \left(y_{kh} - \mathbf{x}_{kh}^t \mathbf{b}_y\right) - R\left(z_{kh} - \mathbf{x}_{kh}^t \mathbf{b}_z\right)$, et $\mathbf{b}_y$ et $\mathbf{b}_z$ sont les coefficients de régression dans la population finie de la régression de y_{kh} et z_{kh}, respectivement, sur $\mathbf{x}_{kh}$. Un estimateur de la variance calculée par (6.2) est donné par

$$\hat{V}\left(\hat{R}\right) = \frac{1}{\hat{t}_z^2}\sum_{h=1}^{H}\left(1 - f_h\right)\frac{n_h}{n_h - 1}\sum_{k=1}^{n_h}\left(w_k \hat{e}_k - \frac{1}{n_h}\sum_{k'=1}^{n_h} w_{k'}\hat{e}_{k'h}\right)^2, \tag{6.3}$$

où $\hat{e}_{kh} = \left(y_{kh} - \mathbf{x}_{kh}^t \hat{\mathbf{b}}_y\right) - \hat{R}\left(z_{kh} - \mathbf{x}_{kh}^t \hat{\mathbf{b}}_z\right)$ et $\hat{\mathbf{b}}_y$ et $\hat{\mathbf{b}}_z$ sont les estimateurs de type HT pour $\mathbf{b}_y$ et $\mathbf{b}_z$. Ces résultats découlent directement de l'intégration des espérances d'inclusion de premier et de deuxième ordre calculées par les équations (3.3) à (3.6) dans l'approximation générale de la variance du ratio de deux estimateurs GREG et de son estimateur (Särndal et coll. 1992, section 7.13).

Les mêmes expressions pour la variance peuvent être dérivées des expressions de la variance proposées pour la méthode généralisée du partage des poids dans le cas d'un échantillonnage indirect. Lavallée (1995) a établi pour l'estimateur HT des expressions de la variance fondées sur le plan d'échantillonnage utilisé pour sélectionner l'échantillon s^A de n unités de la population U^A avec des variables cibles transformées, par exemple z_i. Dans cette application, chaque unité de U^A a exactement un lien avec une unité de U^B. Par conséquent, la variable z_i de Lavallée (1995) est dans ce cas définie comme la somme des variables

cibles de tous les éléments de la grappe k, divisée par le nombre d'unités de la grappe k ayant un lien avec la population U^A, c'est-à-dire $z_i = y_k / g_k$ pour tout $i \in U^A$ ayant un lien avec la grappe $k \in U^B$. En insérant les probabilités d'inclusion de premier et de deuxième ordre pour l'échantillonnage aléatoire simple stratifié sans remise et les variables transformées z_i (où la variable cible y_k est remplacée par le résidu de la régression sur les totaux de grappe e_k) dans la formule de variance pour un ratio, on obtient (6.2). L'expression (6.3) est obtenue de la même manière.

7 Application

Dans le cadre de la RIS, les personnes principales sont sélectionnées dans la population de 15 ans ou plus au moyen d'un échantillonnage aléatoire simple stratifié sans remise selon une fraction de sondage de 0,16. Dans la présente application, les résultats sont présentés pour une grande municipalité (Rotterdam), une municipalité de taille moyenne (Enschede) et une petite municipalité (Sevenum), pour trois années consécutives (2006, 2007 et 2008). Les tailles des populations et des échantillons pour ces trois municipalités sont indiquées dans le tableau 7.1.

Tableau 7.1
Taille de population et d'échantillon de la RIS pour trois municipalités néerlandaises

Municipalité	Population		Échantillon		
	Ménages	Personnes de 15 ans ou plus	Personnes principales	Ménages uniques	Personnes uniques
Rotterdam	293 400	484 000	73 000	67 600	171 400
Enschede	74 200	128 000	19 300	17 600	46 300
Sevenum	2 950	6 100	870	750	2 500

Les variables cibles d'intérêt pour la RIS sont les suivantes :

- Répartition du revenu des ménages dans dix catégories fondées sur des quantiles de dix points de pourcentage (déciles) de la répartition nationale, selon le revenu du ménage normalisé (abrégée RépRevMén);
- Revenu moyen normalisé du ménage (abrégé RevMén);
- Revenu disponible moyen des personnes ayant un revenu durant les 52 semaines de l'année (abrégé RevP).

Le revenu disponible d'une personne s'entend du revenu total d'une personne après impôt. Le revenu total comprend la rémunération, les profits, le revenu tiré du capital et de l'épargne, ainsi que les avantages sociaux et autres avantages. Le revenu normalisé du ménage s'entend du revenu disponible total d'un ménage, corrigé pour tenir compte des différences dans la taille et la composition du ménage. Dans les ouvrages publiés, on parle aussi de revenu disponible équivalent (OECD 2013).

Les estimations destinées aux publications officielles relatives à la RIS sont obtenues à l'aide de l'estimateur GREG selon la méthode de Lemaître et Dufour (1987). Comme l'enquête n'est pas touchée par la non-réponse, les données auxiliaires sont utilisées dans l'estimation pour réduire la variance et pour la

cohérence entre les cellules marginales des différents tableaux publiés. Les espérances d'inclusion sont fondées sur les formules dérivées à la sous-section 3.1. Pour chaque municipalité, on applique le schéma de pondération suivant dans l'estimateur GREG :

$$\text{Âge}\,(7) \times \text{Sexe} + \text{Âge}\,(4) \times \text{Sexe} \times \text{État matrimonial}\,(2) + \text{Adresse}\,(2) \times \text{Taille du ménage}\,(5).$$

Toutes les variables auxiliaires sont catégoriques. Les chiffres entre parenthèses correspondent au nombre de catégories. L'état matrimonial fait la distinction entre les personnes mariées et les autres états matrimoniaux. L'adresse fait la distinction entre les adresses où réside une famille et les autres types d'adresse. La taille du ménage fait la distinction entre les ménages comptant une, deux, trois, quatre et cinq personnes ou plus. Les estimations pour RevMén et RevP incluant les erreurs types fondées sur l'estimateur HT, l'estimateur GREG et l'estimateur GREG selon la méthode de Lemaître et Dufour (1987) sont données au tableau 7.2. À la figure 7.1, les répartitions du revenu RépRevMén estimées à l'aide de l'estimateur HT, de l'estimateur GREG et de l'estimateur GREG selon la méthode de Lemaître et Dufour (1987) sont représentées selon un intervalle de confiance à 95 % pour Rotterdam et Sevenum en 2008. Les erreurs types pour ces estimations sont comparées dans un histogramme distinct. À la figure 7.2, la RépRevMén pour Rotterdam et Sevenum estimée selon la méthode de Lemaître et Dufour (1987) est donnée pour 2006, 2007 et 2008. Voir van den Brakel (2013) pour en savoir davantage sur les répartitions du revenu.

Tableau 7.2
Résultats des estimations dans le cadre de la RIS pour Rotterdam (grande ville), Enschede (ville de taille moyenne) et Sevenum (petit village); erreurs types entre parenthèses

	Variable	Année	HT		GREG		GREG convergent (L et D)	
Rotterdam	RevMén	2006	19 790	(83)	20 134	(80)	20 161	(76)
		2007	22 306	(73)	22 950	(64)	22 866	(64)
		2008	23 750	(78)	24 511	(69)	24 410	(68)
	RevP	2006	22 074	(94)	22 219	(84)	22 233	(93)
		2007	24 094	(82)	24 362	(75)	24 432	(78)
		2008	25 325	(84)	25 625	(75)	25 705	(78)
Enschede	RevMén	2006	19 810	(128)	20 353	(111)	20 300	(107)
		2007	20 878	(128)	21 716	(107)	21 753	(105)
		2008	22 254	(148)	23 235	(125)	23 237	(123)
	RevP	2006	20 402	(102)	20 608	(92)	20 590	(92)
		2007	21 387	(115)	21 751	(103)	21 852	(106)
		2008	22 235	(123)	22 659	(110)	22 724	(114)
Sevenum	RevMén	2006	25 696	(799)	25 698	(734)	25 968	(711)
		2007	28 207	(618)	28 901	(520)	29 026	(490)
		2008	31 466	(795)	32 372	(715)	32 536	(694)
	RevP	2006	21 328	(466)	21 680	(428)	21 712	(428)
		2007	24 056	(456)	24 219	(396)	24 459	(393)
		2008	24 980	(468)	25 482	(426)	25 644	(455)

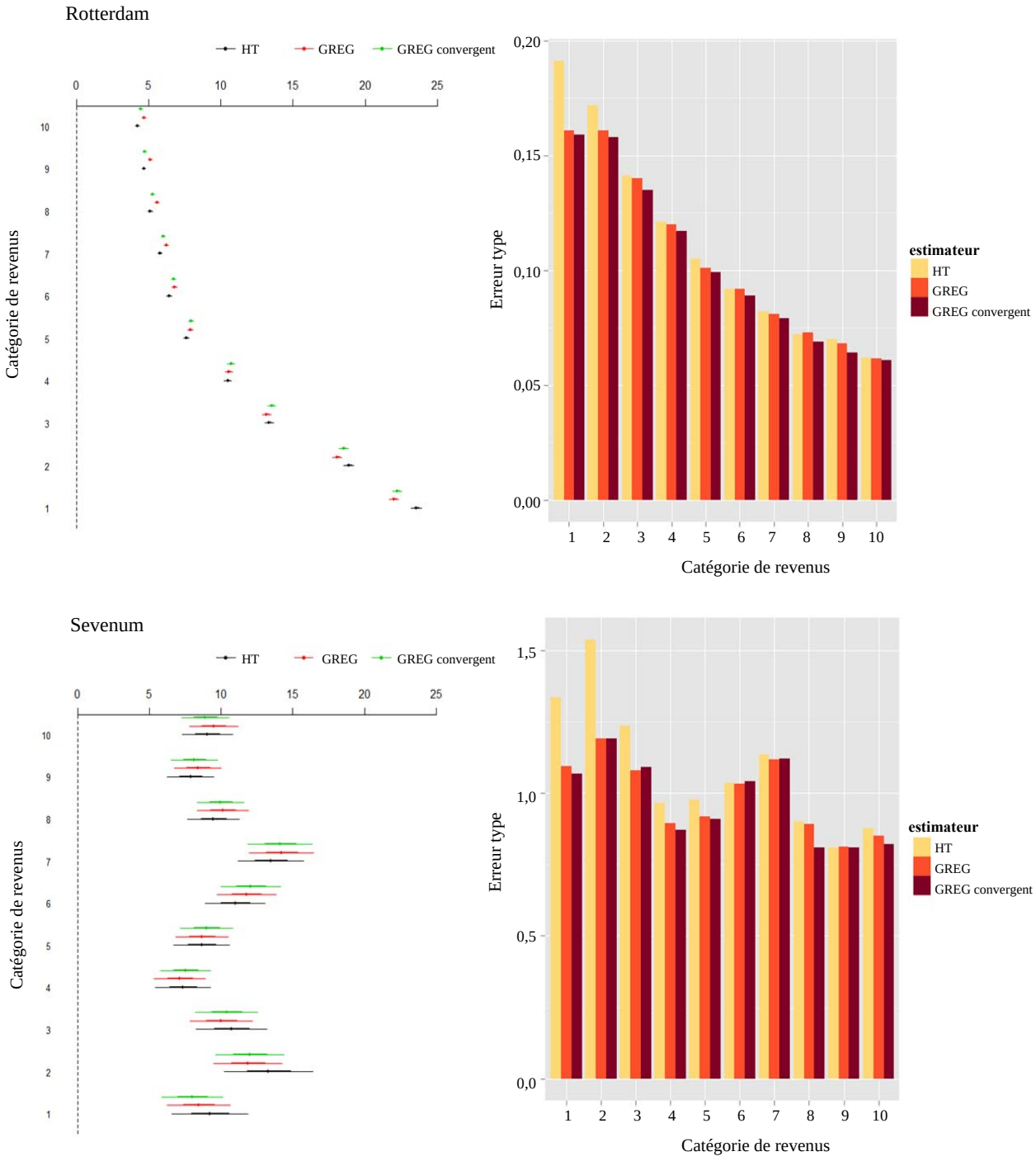

Figure 7.1 RépRevMén en pourcentage pour Rotterdam et Sevenum (à gauche) selon l'estimateur de Horvitz-Thompson, l'estimateur GREG et l'estimateur GREG intégré (GREG convergent), avec intervalles de confiance à 95 %; les erreurs types des estimateurs correspondants sont illustrées à droite.

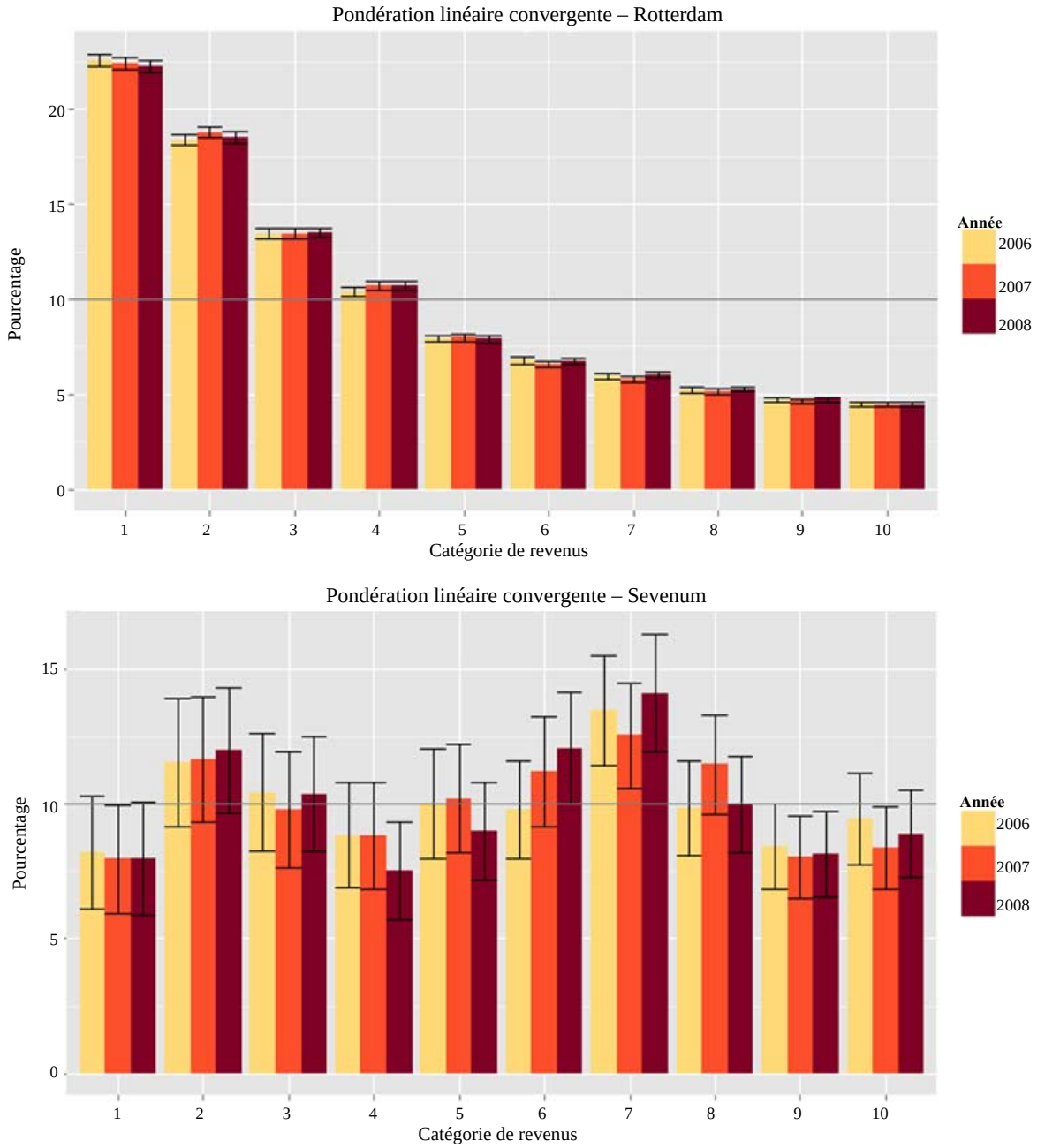

Figure 7.2 RépRevMén en pourcentage pour Rotterdam (en haut) et Sevenum (en bas) estimées selon une pondération intégrée pour 2006, 2007 et 2008 avec intervalles de confiance à 95 %; la ligne grise représente la répartition nationale du revenu.

Les répartitions du revenu observées illustrées aux figures 7.1 et 7.2 résultent de la composition démographique de chacune des deux municipalités. Rotterdam est une ville où la fraction des ménages se trouvant dans les catégories inférieures de revenus est supérieure à la moyenne nationale, les fractions des

trois premières catégories étant supérieures à 10 %. Le nombre de ménages se trouvant dans les catégories supérieures de revenus, en revanche, est inférieur à la moyenne nationale, les fractions de ces catégories étant inférieures à 10 %. Il s'agit d'une répartition type pour une grande ville universitaire où on trouve une fraction élevée d'immigrants non occidentaux. En revanche, Sevenum est un petit village situé près d'une grande ville industrielle. De tels villages ont généralement de petites fractions d'immigrants, pas d'étudiants et des fractions importantes de ménages comptant une ou deux personnes recevant un revenu 52 semaines par année. Cela explique pourquoi la fraction des ménages se trouvant dans la catégorie inférieure de revenus est sous la moyenne nationale, alors que la fraction des ménages se trouvant dans les catégories supérieures de revenus (6, 7 et 8) est supérieure à la moyenne nationale. Sevenum est un village qui n'attire pas les ménages extrêmement riches.

Comme RevMén et RevP sont fondées sur des définitions différentes du revenu et que RevP représente la moyenne des domaines des personnes qui reçoivent un revenu 52 semaines par année, les différences entre les deux moyennes varient d'une municipalité à l'autre. Dans une grande ville universitaire comme Rotterdam, le revenu moyen normalisé des ménages est généralement plus faible que la moyenne du revenu personnel disponible pour l'ensemble des personnes qui reçoivent un revenu 52 semaines par année. D'autres villes où se trouvent de grandes universités présentent un profil semblable. Dans un village petit mais riche comme Sevenum, la situation est inversée.

On remarque en outre qu'à Rotterdam et à Enschede, l'écart entre l'estimateur HT et l'estimateur GREG est relativement grand par rapport aux erreurs types. Compte tenu du grand échantillon et du fait qu'il n'y a pas de non-réponse, ces différences devraient être plus petites. Cela pourrait s'expliquer par le fait que Rotterdam et Enschede sont de grandes villes universitaires. Les étudiants sont souvent inscrits dans les registres de l'impôt (qui sont utilisés comme base de sondage) d'une manière différente que dans les registres de population (qui sont utilisés pour dériver les répartitions des variables auxiliaires dans la population), particulièrement en ce qui concerne la situation de leur ménage.

Pour chaque municipalité, on constate au fil du temps une augmentation constante de la moyenne du revenu des ménages et des personnes. De plus, les répartitions du revenu dans chaque municipalité affichent des tendances stables au fil des ans. Il s'agit là de résultats prévisibles si un panel est réalisé sur de grands échantillons pour estimer des phénomènes qui ne sont pas très volatils dans le temps.

La comparaison des estimations GREG avec et sans recours à la méthode de Lemaître et Dufour (1987) montre que les erreurs types des paramètres estimés des ménages sont plus faibles si on utilise la méthode de Lemaître et Dufour (1987). La différence est particulièrement visible lorsqu'on observe le revenu moyen des ménages dans le petit échantillon de Sevenum. En revanche, pour les paramètres estimés fondés sur la personne, la méthode de Lemaître et Dufour (1987) entraîne une erreur type légèrement plus élevée que celle de l'estimateur GREG ordinaire. Ce résultat donne à penser que la structure de variance présumée pour les résidus du modèle de régression sous-jacent dans le cas de la pondération intégrée convient mieux aux variables fondées sur le ménage qu'aux variables fondées sur la personne.

8 Discussion

En raison de leur instabilité au fil du temps, les ménages ne constituent pas des unités d'échantillonnage appropriées dans les panels visant à recueillir des données au niveau des ménages ou des personnes. Dans

le présent article, on propose un plan d'échantillonnage où les personnes sont tirées à l'aide d'un plan d'échantillonnage autopondéré. À chaque point dans le temps, les membres du ménage de ces personnes dites « principales » sont inclus dans l'échantillon. On obtient ainsi un échantillon où les ménages peuvent être tirés plus d'une fois, jusqu'à concurrence du nombre de personnes dans le ménage. Les ménages sont inclus selon une espérance proportionnelle à la taille du ménage. Les espérances d'inclusion de premier et de deuxième ordre pour les ménages sont calculées en vertu d'un plan d'échantillonnage à probabilités égales pour la sélection des personnes principales. Ces espérances d'inclusion peuvent être utilisées d'une façon comparable aux probabilités d'inclusion plus couramment employées pour l'inférence fondée sur le plan et assistée par modèle.

Le plan d'échantillonnage proposé dans la présente étude constitue un cas particulier de la méthode d'échantillonnage indirect (Lavallée 1995, 2007). Dans le cas d'un plan d'échantillonnage autopondéré, il est montré qu'il est possible de dériver d'une manière relativement simple les espérances d'inclusion de premier et de deuxième ordre pour ce plan d'échantillonnage à partir de la composition des ménages des personnes principales à chaque point dans le temps. Dans le cas des plans d'échantillonnage plus complexes, il faut employer la méthode généralisée du partage des poids (Lavallée 1995, 2007) pour établir les poids d'inclusion à chaque point dans le temps.

L'avantage du plan d'échantillonnage proposé est que la méthode d'estimation est plus simple que la méthode généralisée du partage des poids. Le plan est particulièrement utile si les personnes principales sont sélectionnées à l'aide d'un plan d'échantillonnage autopondéré. Si toutefois on a besoin, par exemple à cause des exigences minimales en matière de précision et maximales en matière de coûts, d'un plan à probabilités inégales pour la sélection des personnes principales, il faut alors utiliser la méthode généralisée du partage des poids. Comme les personnes principales demeurent dans le panel pour une période indéterminée, ce plan d'échantillonnage convient particulièrement bien aux panels auprès des ménages fondés sur des registres, en vertu desquels toute l'information requise est tirée de données administratives. Dans le cas des panels auprès des ménages fondés sur des interviews, il faut mettre en place un plan de renouvellement afin de remédier à certains problèmes comme l'attrition des participants.

La présente étude propose la mesure dite de l'erreur type moyenne, soit la racine carrée de la moyenne des variances des catégories des revenus estimés d'une répartition des revenus, comme mesure de précision pour déterminer la taille minimale d'échantillon. On montre que la valeur maximale de cette mesure de précision correspond à une distribution où les proportions dans les catégories sont égales. On montre aussi que ce résultat peut être vu comme une généralisation de la variance d'une fraction qui atteint sa valeur maximale à 0,5. On dérive ensuite une expression de la taille minimale d'échantillon requise pour satisfaire une précision prédéterminée pour les répartitions estimées. Comme un même ménage peut être inclus plus d'une fois dans l'échantillon, on dérive aussi une expression pour déterminer le nombre prévu de ménages uniques dans l'échantillon.

D'autres recherches sont nécessaires pour étudier la combinaison de cette mesure de l'erreur type moyenne avec une répartition de Neyman ou avec des répartitions exponentielles afin d'établir des expressions pour la taille minimale d'échantillon fondées sur les exigences en matière de précision pour les répartitions estimées au niveau des strates agrégées. On obtient actuellement un plan à probabilités inégales d'inclusion pour les personnes principales, pour lequel il faut employer la méthode généralisée du partage des poids afin de calculer les poids appropriés.

Dans le contexte des enquêtes et des panels auprès des ménages, il convient d'employer des méthodes de pondération qui appliquent des poids de régression égaux aux personnes d'un même ménage afin d'assurer la cohérence des estimations fondées sur les personnes et sur les ménages. Dans le cadre de la présente étude, on utilise pour la RIS une approche de pondération intégrée fondée sur les travaux de Lemaître et Dufour (1987). Les erreurs types obtenues en vertu de la méthode de Lemaître et Dufour (1987) sont plus faibles que celles qu'on obtient avec une méthode de pondération non intégrée pour les estimations fondées sur les ménages. Dans le cas des estimations fondées sur les personnes, les erreurs types peuvent être légèrement plus grandes. Ces résultats sont conformes à ceux de Steel et Clark (2007), qui montrent que la variance par rapport à un plan de sondage en grand échantillon avec pondération intégrée au niveau du ménage est plus faible ou égale à la variance par rapport au plan obtenue avec une pondération non intégrée au niveau de la personne. Ils rapportent aussi que leur simulation a donné de faibles augmentations des variances par rapport au plan à cause de la pondération intégrée utilisée sur des échantillons de petite taille.

La pondération intégrée de Lemaître et Dufour (1987) au niveau du ménage est obtenue grâce à une structure de variance pour les résidus proportionnelle à la taille du ménage (Nieuwenbroek 1993). Si les caractéristiques du ménage sont proportionnelles à la taille du ménage, on peut s'attendre à ce qu'une telle structure de variance explique mieux la variation des variables des ménages dans la population, comparativement à une structure de variance qui suppose une variance résiduelle égale pour les ménages. Dans le cas des variables fondées sur la personne, une telle structure de variance pourrait être moins efficiente, mais la pondération intégrée présente l'avantage supplémentaire que les totaux pour le revenu fondés sur les ménages et sur les personnes, qui peuvent être dérivés directement à partir des moyennes, sont cohérents.

Remerciements

Les opinions exprimées dans l'article sont celles de l'auteur et ne reflètent pas les politiques de *Statistics Netherlands*. L'auteur remercie le rédacteur en chef adjoint et les examinateurs anonymes de leurs commentaires constructifs au sujet de deux versions antérieures du présent article, ainsi que Drs. M. van den Brakel-Hofmans pour avoir mis à sa disposition les données de la RIS.

Annexe technique

Preuve de l'équation (4.4)

Une expression pour la variance de la fraction estimée des ménages de la catégorie de revenus l peut être dérivée de l'expression générale pour la variance de l'estimateur HT (Särndal et coll. 1992, section 2.8) :

$$V\left(\hat{P}_{lh}\right) = \frac{1}{M_h^2}\sum_{k=1}^{M_h}\sum_{k'=1}^{M_h}\left(\pi_{kk'h} - \pi_{kh}\pi_{k'h}\right)\frac{y_{khl}}{\pi_{kh}}\frac{y_{k'hl}}{\pi_{k'h}}. \qquad \text{(A.1)}$$

Si l'on insère les espérances d'inclusion de premier et de deuxième ordre précisées aux expressions (3.3) à (3.6) et si l'on exploite l'égalité $y_{khl} = y_{khl}^2$, puisque les valeurs de la variable cible sont limitées à zéro ou un, il s'ensuit, après quelques manipulations algébriques, que l'expression (A.1) peut être simplifiée comme suit :

$$V\left(\hat{P}_{lh}\right) = \frac{N_h - n_h}{n_h}\frac{1}{N_h - 1}\left(\frac{N_h}{M_h^2}\sum_{k=1}^{M_h}\frac{y_{khl}}{g_{kh}} - \left(\frac{M_{lh}}{M_h}\right)^2\right). \tag{A.2}$$

On obtient l'équation (4.4) en insérant (A.2) dans (4.3).

Preuve de l'équation (4.5)

La *population* des ménages de la strate h peut être divisée en T sous-populations de ménages de taille égale. Soit M_{th} le nombre de ménages de taille t dans la strate h. Il s'ensuit maintenant pour la double sommation entre parenthèses pour l'expression de s en (4.4) que

$$\sum_{l=1}^{L}\sum_{k=1}^{M_h}\frac{y_{khl}}{g_{kh}} = \sum_{l=1}^{L}\sum_{t=1}^{T}\sum_{k=1}^{M_{th}}\frac{y_{khl}}{t} = \sum_{t=1}^{T}\frac{M_{th}}{t}. \tag{A.3}$$

En vertu de l'inégalité de Cauchy-Schwartz (Cochran 1977, section 5.5), il s'ensuit pour la sommation entre parenthèses pour l'expression de s_h en (4.4) que

$$\sum_{l=1}^{L}\left(\frac{M_{lh}}{M_h}\right)^2 = \sum_{l=1}^{L}P_{lh}^2 \geq \frac{1}{L}. \tag{A.4}$$

On obtient l'équation (4.5) en insérant (A.3) et (A.4) dans l'expression de s en (4.4).

Preuve de l'équation (4.7)

Soit $\tilde{\pi}_{tkh}$ la probabilité d'inclusion pour le ménage k de la strate h de taille t. Comme les ménages de même taille ont les mêmes probabilités de premier ordre, il s'ensuit que $\tilde{\pi}_{tkh} = \tilde{\pi}_{tk'h} \equiv \tilde{\pi}_{th}$. Soit I_{tkh} une variable indicatrice, qui prend la valeur 1 si le ménage k de la strate h de taille t est inclus dans l'échantillon, et la valeur 0 autrement. Le nombre prévu de ménages uniques peut être calculé comme suit :

$$\begin{aligned} D_h &= \mathrm{E}\left(\sum_{t=1}^{T}\sum_{k=1}^{M_{th}} I_{tkh}\right) = \sum_{t=1}^{T} M_{th}\tilde{\pi}_{th} \\ &= \sum_{t=1}^{T} M_{th}\left(1 - \frac{\binom{N_h - t}{n_h}}{\binom{N_h}{n_h}}\right) = \sum_{t=1}^{T} M_{th}\left(1 - \frac{(N_h - n_h)(N_h - n_h - 1)\ldots(N_h - n_h - t + 1)}{N_h(N_h - 1)\ldots(N_h - t + 1)}\right). \end{aligned}$$

Bibliographie

Bankier, M.D. (1988). Power allocations: Determining sample sizes for subnational areas. *The American Statistician*, 42, 174-177.

Bethlehem, J.G. (2009). *Applied Survey Methods*, New Jersey: John Wiley & Sons, Inc.

Cochran, W.G. (1977). *Sampling Techniques*, New York: John Wiley & Sons, Inc.

Deville, J.-C., et Lavallée, P. (2006). Sondage indirect : les fondements de la méthode généralisée du partage des poids. *Techniques d'enquête*, 32, 2, 185-196.

Ernst, L. (1989). Weighting issues for longitudinal household and family estimates. Dans *Panel Surveys*, (Éds. D. Kasprzyk, G. Duncan, G. Kalton et M.P. Singh). New York: John Wiley & Sons, Inc., 135-159.

Estevao, V.M., et Särndal, C.-E. (2006). Survey estimates by calibration on complex auxiliary information. *Revue Internationale de Statistique*, 74, 127-147.

Horvitz, D.G., et Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.

Kalton, G., et Brick, J.M. (1995). Méthodes de pondération pour les enquêtes par panel auprès des ménages. *Techniques d'enquête*, 21, 1, 37-49.

Lavallée, P. (1995). Pondération transversale des enquêtes longitudinales menées auprès des individus et des ménages à l'aide de la méthode du partage des poids. *Techniques d'enquête*, 21, 1, 27-35.

Lavallée, P. (2007). *Indirect Sampling*, New York: Springer Verlag.

Lemaître, G., et Dufour, J. (1987). Une méthode intégrée de pondération des personnes et des familles. *Techniques d'enquête*, 13, 2, 211-220.

Lynn, P. (2009). Methods for longitudinal surveys. Dans *Methodology of Longitudinal Surveys,* (Éd., P. Lynn), Wiley, Chichester, 1-19.

Narain, R. (1951). On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 3, 169-174.

Nieuwenbroek, N.J. (1993). An integrated method for weighting characteristics of persons and households using the linear regression estimator. Document de recherche, BPA nr: 8555-93-M1-1, Statistics Netherlands, Heerlen.

OECD (2013). *OECD Framework for Statistics on the Distribution of Household Income, Consumption and Wealth*. OECD publishing, http://dx.doi.org/10.1787/9789264194830-en.

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, 21-33.

Steel, D.G., et Clark, R.G. (2007). Estimation par la régression au niveau de la personne et au niveau du ménage dans les enquêtes-ménages. *Techniques d'enquête*, 33, 1, 59-69.

van den Brakel, J.A. (2013). Sampling and estimation techniques for household panels. Document de discussion 2013-15, Statistics Netherlands, Heerlen. http://www.cbs.nl/NR/rdonlyres/B4F85FB9-52F2-4B8A-94C4-56DA43F2250D/0/201315x10pub.pdf.

Wallgren, A., et Wallgren, B. (2007). *Register-Based Statistics: Administrative Data for Statistical Purposes*. New York: John Wiley & Sons, Inc.