Journal de la Société Française de Statistique
Vol. 155 No. 4 (2014)

Sondage dans des registres de population et de ménages en Suisse : coordination d'enquêtes, pondération et imputation

Title: Sampling from population and household registers in Switzerland: sample coordination, estimation and imputation

Éric Graf[1] et Lionel Qualité[2]

Résumé : L'Office Fédéral de la Statistique harmonise ses enquêtes par échantillonnage auprès des personnes et des ménages en Suisse. Dans cet article, nous présentons un aperçu des méthodes actuellement utilisées. Les échantillons sont sélectionnés de manière coordonnée afin de répartir au mieux la charge d'enquête sur les ménages et les personnes. Le calcul des pondérations, dont on présente les principales étapes, est adapté aux différents besoins et aux différentes situations rencontrées. L'Office se base sur les recommandations internationales, dont il participe à l'élaboration, pour le traitement des données d'enquête et les imputations. La précision des estimateurs est systématiquement évaluée en tenant compte des traitements réalisés.

Abstract: The Swiss Federal Statistical Office is currently harmonizing its population and household survey operations. In this paper, we give an overview of its production methods. Samples are selected with coordination so as to spread the survey burden over the population. The computation of extrapolation weights adapted to different cases and needs is presented with its main steps. The Office relies on international recommendations for data editing and imputation, and contributes to their elaboration. The precision of estimators is consistently evaluated, according to the different treatments and methods involved in their construction.

Mots-clés : recensement, calage, variance, segmentation, valeurs aberrantes
Keywords: census, calibration, variance, segmentation, outliers
Classification AMS 2000 : 62D05

1. Introduction

L'Office Fédéral de la Statistique (OFS) achève d'harmoniser ses enquêtes par échantillonnage auprès des personnes et des ménages en Suisse. Dans cet article, nous présentons un aperçu des méthodes actuellement utilisées par l'office. Il ne s'agit en aucun cas d'un manuel de bonnes pratiques, mais de l'exposé de solutions et compromis qui ont été réalisés pour les besoins spécifiques de l'OFS. L'uniformisation des méthodes d'échantillonnage, de pondération, et de traitement des données d'enquête est l'occasion de réévaluer toutes ces pratiques pour en dégager un corpus de méthodes solide et applicable aux divers cas rencontrés par l'office.

Dans une première partie, nous décrivons les enquêtes auprès de la population et auprès des ménages réalisées par l'OFS. On y présente les pendants suisses de grandes enquêtes européennes

[1] Université de Neuchâtel.
E-mail : eric.graf@unine.ch
[2] Université de Neuchâtel et Office Fédéral de la Statistique.
E-mail : lionel.qualite@unine.ch

 ISSN: 2102-6238

ainsi que le nouveau système de recensement, en partie basé sur des enquêtes annuelles par échantillonnage, qui remplace le recensement décennal de la population.

Dans une deuxième partie, nous présentons la méthode de tirage avec coordination utilisée pour les enquêtes du nouveau système de recensement et qui sera utilisée à terme pour toutes les enquêtes de l'OFS. Le tirage coordonné d'échantillons permet de mieux répartir la charge d'enquête sur la population.

Dans la troisième partie, nous présentons les méthodes d'estimation et de pondération utilisées pour les enquêtes de l'OFS. Les estimateurs utilisés sont en principe des estimateurs par dilatation, c'est-à-dire des sommes pondérées de valeurs observées (ou imputées). Nous présentons les étapes-clés de la construction des pondérations : calcul des probabilités d'inclusion dans l'échantillon, estimation des probabilités de réponse, partage des poids, combinaison d'échantillons, calages.

Dans la quatrième partie, nous présentons le principe de traitement des données et d'imputation appliqué à l'OFS. L'Office suit les recommandations issues de plusieurs grands projets européens quant au traitement des valeurs manquantes, aberrantes, erronées ou extrêmes. Nous mentionnons les principales techniques d'imputation utilisées en production.

Dans la cinquième partie, nous traitons les méthodes d'estimation de la variance utilisées à l'OFS. Celles-ci sont pensées pour tenir le mieux possible compte des principaux facteurs influençant la variance des estimateurs produits : le plan de sondage, la non-réponse et son traitement, les partages ou combinaisons de poids, enfin les calages et imputations. Dans la pratique, un compromis doit être réalisé car la prise en compte de tous ces aspects rend le problème extrêmement complexe. Des approximations simples sont donc utilisées.

Dans la sixième partie, nous donnons un aperçu sur les évolutions planifiées dans un futur proche : le traitement de la non-réponse par calage généralisé, l'estimation de la variance due à l'imputation ainsi que les dernières améliorations prévues quant à la coordination des échantillons.

2. Les différentes enquêtes auprès de la population des personnes et des ménages résidant en Suisse

2.1. Le nouveau système de recensement de la population suisse

Le traditionnel recensement décennal de la population suisse est remplacé, depuis 2010, par la création et l'exploitation par l'OFS d'un cadre d'échantillonnage [1] de population et de ménages appelé SRPH (Stichprobenrahmen für Personen und Haushaltserhebungen), et un système d'enquêtes annuelles. L'office ne gère pas de registre centralisé de la population, mais utilise différents registres pour créer le fichier SRPH. Celui-ci est essentiellement construit en agrégeant les registres administratifs du contrôle des habitants des communes et des cantons. L'inscription dans le registre de la commune de résidence est en effet obligatoire en Suisse, à quelques exceptions près, parmi lesquelles on peut citer les personnes ayant le statut de fonctionnaire international. Le fichier ainsi obtenu est encore enrichi par l'utilisation de registres fédéraux : le registre des demandeurs d'asile, le registre central des étrangers, le registre des fonctionnaires internationaux et le registre informatisé de l'état civil Infostar. La livraison des données par les communes et cantons, et la création du cadre d'échantillonnage, sont effectuées quatre fois par an, en dates de

[1] La locution "cadre d'échantillonnage" utilisée en Suisse est équivalente à "base de sondage". Il s'agit de la traduction du mot allemand "Stichprobenrahmen".

 ISSN: 2102-6238

valeur des 31 décembre, 31 mars, 30 juin et 30 septembre. Les données sont disponibles environ six semaines après ces dates de référence.

Le fichier SRPH contient des informations démographiques basiques sur les personnes habitant en Suisse : sexe, date de naissance, état civil, permis de séjour, nationalité, date d'arrivée en Suisse et au lieu de résidence actuel, ainsi que l'adresse, le statut de résidence (résidence principale dans un ménage privé, résidence secondaire, ménage collectif, etc.). L'OFS a la conviction que ce répertoire offre une très bonne couverture de la population résidant de manière permanente en Suisse, l'inscription au registre de la commune étant nécessaire pour bénéficier d'un certain nombre de prestations, et déterminante pour le recouvrement de l'impôt sur le revenu. La pertinence de ce cadre de sondage pour les opérations statistiques est garantie par l'utilisation permanente du numéro unique de sécurité sociale comme identifiant à toutes les étapes de sa création, ainsi que par des procédures de contrôle de la qualité, depuis la livraison des données par les communes jusqu'à la constitution finale du fichier SRPH.

Depuis la fin décembre 2012, les communes ont l'obligation légale de fournir le numéro de bâtiment et de logement de chaque personne. Ces numéros, qui permettent de connaître la constitution des ménages, sont extraits du registre des bâtiments et logements entretenu à l'OFS. Cette obligation était déjà largement respectée depuis 2010, mais avec une certaine hétérogénéité selon les cantons et communes.

Le premier élément du nouveau système de recensement est le résultat de l'exploitation, chaque année, du répertoire de population du 31 décembre précédent. Les données utilisées pour constituer le fichier SRPH du 31 décembre font l'objet de traitements supplémentaires pour produire la statistique de la population et des ménages (Statpop). Cette exploitation permet de fournir une première série de résultats, et également la population de référence pour les enquêtes complémentaires du système de recensement. Les données démographiques individuelles sont utilisées. Le taux de couverture de cette statistique sera évaluée sur la base d'une enquête réalisée en 2013. La constitution des ménages n'a pas été exploitée pour ces premiers résultats de comptages, tant que l'obligation légale de renseigner n'était pas en vigueur. Elle est par contre utilisée dans les processus de production de l'OFS au moment de l'échantillonnage.

Le deuxième élément du nouveau système de recensement est une enquête annuelle, par échantillonnage aléatoire, de grande ampleur : l'enquête structurelle. L'échantillon de base, financé par la confédération helvétique, est dimensionné de manière à obtenir approximativement 200'000 questionnaires renseignés par des personnes ayant leur résidence permanente en Suisse, et âgées de 15 ans ou plus. Cela représente un taux de sondage de 3% environ. Les cantons et communes peuvent financer une augmentation de l'échantillon les concernant, jusqu'à le doubler. Ponctuellement, en 2010, ils avaient la possibilité de quadrupler l'échantillon. Le questionnaire de cette enquête reprend en grande partie les questionnaires des recensements précédents en omettant les informations disponibles dans le cadre d'échantillonnage SRPH. Les thèmes abordés sont, entre autres, le niveau d'éducation, le statut d'activité et la branche d'activité, les trajets domicile-travail, la langue employée, la composition du ménage et les relations familiales, ainsi que le statut d'occupation du logement et le loyer s'il y a lieu. Le taux de réponse observé à cette enquête obligatoire est de 90% environ. Le mode de collecte est en principe le questionnaire papier à retourner ou la télédéclaration par internet, le choix étant laissé au répondant.

Le troisième élément du système de recensement suisse est une enquête thématique annuelle, avec un échantillon de base de 10'000 à 40'000 répondants selon le thème. Chacun des cinq

 ISSN: 2102-6238

thèmes suivants est traité à son tour tous les cinq ans : la mobilité et les transports, la formation initiale et continue, la santé, la famille, les langues les religions et les cultures. L'échantillon de ces enquêtes est constitué de 3 à 4 blocs sélectionnés dans les différents cadres d'échantillonnage de l'année d'enquête. La collecte est principalement réalisée par entretien téléphonique assisté par ordinateur. Enfin, le quatrième élément du système de recensement est une enquête téléphonique annuelle auprès d'environ 3'000 répondants, appelée enquête Omnibus, sur des thèmes qui sont décidés chaque année.

Tous ces échantillons sont sélectionnés dans le cadre d'échantillonnage le plus récent. L'OFS fournit également des échantillons tirés dans le SRPH pour les enquêtes menées par d'autres institutions sur des thèmes d'intérêt national, dont en particulier les projets soutenus par le Fonds National de la Recherche Scientifique.

L'apparition de l'enquête structurelle, qui touche chaque année une part non négligeable de la population suisse, a motivé l'utilisation d'une procédure pour coordonner les échantillons de l'OFS. En effet, si l'on ne prenait aucune mesure pour l'éviter, des dizaines de milliers de personnes seraient sélectionnées à deux enquêtes structurelles successives ou à plusieurs enquêtes de l'OFS en même temps, sans que cela n'ait une utilité statistique. Cela pourrait conduire à une baisse des taux de réponse, à une dégradation de l'image de l'office et à une surcharge de travail liée au traitement des plaintes venant du public. On essaye donc d'éviter de solliciter de manière répétée les personnes et les ménages lorsque cela n'est pas utile. La procédure de sélection coordonnée d'échantillons utilisée à l'OFS pour répondre à ce besoin est décrite en Section 3.

2.2. Les enquêtes sur les conditions de vie, le budget des ménages et la population active

Outre les enquêtes du système de recensement, l'OFS réalise chaque année trois grandes enquêtes par échantillonnage auprès des ménages et des personnes résidant de manière permanente en Suisse dans des ménages privés. Il s'agit de l'enquête sur les revenus et conditions de vie SILC (Survey on Income and Living Conditions), l'Enquête sur le Budget des Ménages (EBM), et l'Enquête Suisse sur la Population Active (Espa).

L'enquête SILC est la source de référence pour les comparaisons statistiques en matière de distribution de revenu et d'exclusion sociale pour l'Union Européenne. Elle est, depuis 2007, le pendant suisse de l'enquête européenne EU-SILC réalisée dans plus de 30 pays d'Europe. Elle doit de ce fait satisfaire aux exigences et recommandations données par Eurostat dans de nombreux domaines, en particulier sur la manière de calculer les pondérations (Eurostat, 2004b, 2005b; Graf, 2008), sur les tailles minimales des échantillons de répondants, ménages et individus (Graf, 2006), et enfin sur la précision des indicateurs clés qui sont livrés chaque année à Eurostat.

L'enquête SILC utilise un panel rotatif de répondants : chaque année, un quart de l'échantillon est renouvelé, et chaque ménage enquêté est théoriquement sollicité pendant quatre années. Elle permet ainsi de produire des résultats pour une année donnée, mais aussi d'estimer des évolutions avec une bonne précision. SILC est réalisée annuellement par téléphone auprès d'environ 7'000 ménages comprenant 17'000 personnes.

L'objectif principal de l'enquête Espa est de fournir des données sur la structure de la population active et sur les comportements en matière d'activité professionnelle. Grâce à l'application stricte de définitions internationales, les données de la Suisse peuvent être comparées avec celles des

 ISSN: 2102-6238

pays de l'OCDE et de l'Union européenne. Il s'agit d'une enquête auprès des personnes, qui est réalisée chaque année depuis 1991. L'organisation de l'enquête a été revue en 2010 pour permettre d'obtenir des estimations trimestrielles, mais aussi des estimations d'évolutions trimestrielles et annuelles précises. Les unités enquêtées sont ré-interrogées après 3, 12 et 15 mois, avant de sortir de l'échantillon. À un trimestre donné, l'échantillon est constitué de quatre blocs. L'un est interrogé pour la première fois, et les autres respectivement pour la deuxième, troisième et quatrième fois (Renfer, 2009; OFS, 2012). L'échantillon de l'Espa est complété par un échantillon de personnes de nationalité étrangère sélectionné dans le registre central des étrangers (Système d'information central sur la migration - Symic). L'Espa est réalisée auprès d'un échantillon annuel d'environ 105'000 personnes, augmenté d'un sur-échantillon annuel de 21'000 personnes étrangères.

Les résultats de l'EBM permettent d'adapter chaque année la composition du panier-type pour le calcul de l'indice suisse des prix à la consommation, ainsi que d'étudier la structure et l'évolution des revenus et des dépenses des ménages. Cette enquête se base sur des fondements et définitions méthodologiques an accord avec les directives internationales, notamment la nomenclature des fonctions de consommation des ménages COICOP (Classification Of Individual COnsumption by Purpose) définie par le Bureau International du Travail (BIT). L'EBM comporte environ 3'000 ménages interviewés chaque année. L'échantillon de l'EBM est partitionné en douze sous-échantillons qui sont interrogés chacun leur tour durant un mois. Cette enquête est réalisée annuellement à l'OFS depuis 2000.

Hormis l'échantillon spécifique d'étrangers de l'Espa, les échantillons de ces trois enquêtes sont tous actuellement sélectionnés dans le registre de numéros de téléphone Castem (Cadre de sondage pour le tirage d'échantillons de ménages) constitué à partir des livraisons des opérateurs téléphoniques actifs en Suisse. La collecte de données est en principe réalisée par entretien téléphonique assisté par ordinateur, éventuellement complétée par des questionnaires papiers ou électroniques. Afin de limiter la charge d'enquête demandée à la population, les numéros de téléphone qui ont été sélectionnés pour une enquête sont écartés pour une durée déterminée de la liste des numéros sélectionnables. Cette pratique est, à l'OFS, appelée "historisation". Cette procédure n'est acceptable sur le plan méthodologique que lorsque les échantillons représentent une part négligeable et non spécifique de la population. D'ici à la fin 2014, les échantillons de ces trois enquêtes seront désormais sélectionnés dans le cadre d'échantillonnage SRPH. Elles pourront ainsi être coordonnées avec les enquêtes du système de recensement. La procédure "d'historisation" des numéros de téléphone déjà sélectionnés sera abandonnée au profit de la méthode de coordination développée pour le système de recensement.

3. Enquêtes coordonnées dans le fichier SRPH

La méthode décrite dans cette section permet la sélection d'échantillons positivement ou négativement coordonnés, c'est à dire d'échantillons dont l'intersection est volontairement importante ou au contraire petite, voire vide. Une coordination négative est utile pour répartir la charge d'enquête équitablement sur la population en évitant que les mêmes unités soient sélectionnées dans plusieurs échantillons lorsque cela n'est pas nécessaire. Elle devient particulièrement importante lorsqu'un grand nombre d'enquêtes sont réalisées, dont certaines ont de forts taux de sondage. Une coordination positive est souhaitable lorsque l'on veut mettre à jour un échantillon de panel, ou

 ISSN: 2102-6238

lorsque l'on veut pouvoir comparer avec précision les résultats d'une nouvelle enquête avec ceux d'une enquête précédente. La méthode proposée est utilisable avec des probabilités d'inclusion inégales, dans une population dynamique avec des naissances, ou entrées, et des décès, ou sorties, d'unités.

La coordination des échantillons est un problème qui a et qui continue d'être largement étudié, et pour lequel de nombreuses solutions ont été développées en particulier par les Instituts Nationaux de Statistique (INS). On peut entre autres se référer à Patterson (1950); Keyfitz (1951); Kish and Scott (1971); Rosén (1997a,b); Brewer et al. (1972); De Ree (1983); Van Huis et al. (1994a,b); Cotton and Hesse (1992a,b); Rivière (1998, 1999, 2001a,b); Ohlsson (1995); Ernst (1996) et Kröger et al. (1999). Les méthodes de coordination développées dans les INS sont essentiellement prévues pour fonctionner avec des plans de sondage aléatoires simples ou stratifiés. Il est toutefois apparu très difficile de développer une méthode de tirage coordonnée qui respecte exactement ces plans de sondage lorsque la population et la définition des strates changent entre les enquêtes. A l'exception de l'enquête Espa qui contient un sur-échantillon d'étrangers, les plans de sondage pour les enquêtes de l'OFS ont tous des spécifications très semblables : les probabilités d'inclusion sont uniformes au sein des cantons ou des communes pour toutes les personnes résidentes permanentes âgées de 15 ans ou plus. Cependant, les échantillons que l'OFS fournit à ses partenaires pour leurs propres enquêtes peuvent avoir des caractéristiques très diverses. Ces instituts peuvent par exemple souhaiter cibler des sous-populations spécifiques identifiables grâce aux données contenues dans le SRPH. Il a donc été nécessaire de développer un système très souple qui permette de se départir d'une stratification fixe.

Brewer et al. (1972) ont inventé l'approche dite des numéros aléatoires permanents. Ce sont des méthodes qui reposent sur l'utilisation d'un même jeu de nombres aléatoires pour sélectionner les différents échantillons au cours du temps. Dans certains cas, dont celui qui est présenté ici, un nombre aléatoire entre 0 et 1 est attribué à chaque unité lorsqu'elle entre dans la population ou dans le cadre de sondage, et ce nombre lui reste attaché jusqu'à ce qu'elle sorte de la population ou du cadre. D'autres méthodes nécessitent de permuter les nombres aléatoires entre unités de la population, suivant les tirages passés (Cotton and Hesse, 1992a; Rivière, 2001a). Une des caractéristiques de la méthode de Brewer et al. (1972) est qu'elle permet de choisir librement les probabilités d'inclusion de chaque unité de la population à chaque enquête. Elle ne requiert donc pas de stratification fixe de la population et possède la flexibilité nécessaire à l'OFS. Une présentation des méthodes à nombres aléatoires permanent se trouve dans Ohlsson (1995).

3.1. Coordination d'échantillons

Il existe différentes définitions de ce qu'est la coordination d'échantillons. Le fait que les processus aléatoires qui permettent de sélectionner les échantillons de deux enquêtes soient ou non indépendants n'est pas ce qui est réellement intéressant. Une définition utile conduit à minimiser ou bien à maximiser, et généralement à contrôler le nombre d'unités communes à deux échantillons. La notion que nous utilisons ici est légèrement différente, et plus adaptée au cas des enquêtes à probabilités inégales : nous travaillons sur les probabilités, pour chaque unité, d'appartenir simultanément à deux ou une certaine collection d'enquêtes. Considérons un plan de sondage pour T enquêtes, c'est à dire une loi de probabilité

$$\mathrm{P}(s^1, s^2, \ldots, s^T),$$

 ISSN: 2102-6238

sur les T échantillons sélectionnables $s^1, s^2, \ldots, s^T$. Ces échantillons ne sont pas nécessairement tous sélectionnés dans la même population. Les distributions marginales d'ordre 1 de $\mathrm{P}(\cdot)$ sont les plans de sondages (appelés transversaux) des T enquêtes considérées. La probabilité qu'une unité k soit sélectionnée dans un échantillon s^i est désignée par

$$\pi_k^i = \mathrm{P}(s^i \ni k),$$

et la probabilité qu'une unité k soit sélectionnée dans l'échantillon s^i et dans l'échantillon s^j est notée

$$\pi_k^{i,j} = \mathrm{P}(s^i \ni k \ \& \ s^j \ni k).$$

Définition 3.1. *On dit qu'il y a une coordination positive pour l'unité k entre les enquêtes i et j si*

$$\pi_k^{i,j} > \pi_k^i \cdot \pi_k^j,$$

et qu'il y a une coordination négative si au contraire

$$\pi_k^{i,j} < \pi_k^i \cdot \pi_k^j.$$

La coordination, positive ou négative, est dite optimale pour l'unité k si le maximum ou minimum théorique de $\pi_k^{i,j}$, qui vaut respectivement $\min(\pi_k^i, \pi_k^j)$ *et* $\max(0, \pi_k^i + \pi_k^j - 1)$*, est atteint.*

Si la coordination est optimale et positive (resp. négative) entre les enquêtes i et j, au sens de la définition 3.1, pour toutes les unités, alors la taille espérée de l'intersection des échantillons s^i et s^j,

$$n^{i,j} = \mathrm{E}\left[\#(s^i \cap s^j)\right] = \sum_k \pi_k^{i,j},$$

est maximale (resp. minimale). Cela signifie que parmi toutes les distributions de probabilité $\mathrm{P}(\cdot)$ sur les couples d'échantillons, ayant des marginales fixées $\mathrm{P}(s^i)$ et $\mathrm{P}(s^j)$, la valeur de $n^{i,j}$ est maximale (resp. minimale) lorsque $\mathrm{P}(\cdot)$ fournit une coordination positive optimale (resp. négative optimale) pour toutes les unités k selon la définition 3.1. Cela n'implique pas que la taille $\#(s^i \cap s^j)$ de l'intersection des échantillons s^i et s^j est constante, maximale ou minimale, et égale à $n^{i,j}$.

3.2. Algorithme d'échantillonnage coordonné

La méthode que nous présentons ici est une extension de la méthode de Brewer et al. (1972). Chaque unité de la population y est traitée indépendamment des autres unités. Les plans de sondage transversaux, pour chaque enquête prise séparément, sont donc des plans dits de Poisson. L'échantillonnage de Poisson, voir par exemple Tillé (2006), est le plan a probabilités inégales qui est obtenu en sélectionnant ou pas chaque unité k de la population dans l'échantillon avec une probabilité π_k, indépendamment des autres unités de la population. La loi de probabilité correspondante est donnée par

$$\mathrm{P}(s) = \prod_{k \in s} \pi_k \prod_{k \in U \setminus s} (1 - \pi_k).$$

Les sélections étant indépendantes, la taille de l'échantillon ne peut pas être imposée. Comme écrit dans Brewer et al. (1984), le plan de sondage de Poisson a trois défauts apparents par rapport

 ISSN: 2102-6238

aux plans de sondage de taille fixe. Le premier défaut, selon les auteurs, est que la probabilité de sélectionner l'échantillon vide n'est pas nulle, le second est que l'estimateur de Horvitz and Thompson (1952) a souvent une variance plus grande avec ce plan de sondage, et le troisième défaut est que l'allocation optimale de l'échantillon entre différents domaines ou strates ne peut pas être obtenue car la taille de l'échantillon dans chaque domaine est aléatoire. On peut ajouter à ces défauts que la variabilité de la taille d'échantillon ne permet pas d'avoir un contrôle total sur le budget d'enquête.

Ces défauts sont en fait présents pour tous les plans de sondage auprès des ménages et des personnes lorsque la non-réponse est possible. Le premier et le troisième point ne sont pas pertinents pour les enquêtes de l'OFS. En effet, l'échantillon est alloué, en espérance, proportionnellement à la taille de la population dans le canton ou la zone administrative. Les tailles espérées d'échantillon sont parfois augmentées dans certains cantons pour permettre d'obtenir des résultats détaillés pour ceux-ci. Aucune optimisation n'est recherchée. La publication des résultats n'est envisagée que pour des domaines où la taille espérée de l'échantillon est tellement grande que la probabilité de sélectionner l'échantillon vide est infime, et où la variabilité de la taille d'échantillon due au plan de sondage est négligeable par rapport à celle due à la non-réponse. L'aléa induit sur les coûts de collecte est également négligeable pour les enquêtes de l'OFS. Le second défaut relevé par Brewer et al. (1984) est pallié par l'utilisation d'estimateurs par le ratio (Strand, 1979; Hájek, 1981) ou d'estimateurs calés (Deville and Särndal, 1992). En effet, la non-réponse conduit systématiquement l'OFS à utiliser une méthode de calage ou apparentée. Les institutions partenaires qui commandent des échantillons à l'OFS sont informées de ces possibles problèmes et encouragées à prévoir leurs tailles d'échantillons espérées dans les domaines en fonction de ceux-ci. Il convient de noter que le principal risque pour l'enquête, que ce soit pour son budget ou pour sa qualité, est lié à l'anticipation des taux de réponse qui peut-être mauvaise et non à l'aléa contrôlé dû à la méthode de sélection.

Une présentation complète de l'algorithme de tirage coordonné utilisé à l'OFS est donnée dans Qualité (2009). La méthode peut être présentée en quelques points. Chaque unité k reçoit un nombre aléatoire permanent u_k lorsqu'elle apparaît dans la population. Ce nombre aléatoire est généré selon une loi uniforme sur $[0,1]$, indépendamment des nombres attribués aux autres unités de la population. Ces nombres aléatoires sont ensuite utilisés pour tous les tirages d'échantillons tant que l'unité fait partie de la population. Si la probabilité d'inclusion de l'unité k à l'enquête t est notée π_k^t, cette unité est sélectionnée pour cette enquête lorsque u_k est inclus dans un sous-ensemble mesurable de $[0,1]$ (en fait pour nos besoins dans une union finie d'intervalles), appelé zone de sélection, de longueur totale égale à π_k^t. De cette manière, les probabilités d'inclusion sont respectées, dans la mesure où le choix de la zone de sélection n'est pas fonction de la valeur de u_k. La coordination entre les enquêtes est alors obtenue en choisissant des zones de sélection dont l'intersection est contrôlée pour les différentes enquêtes. On obtient une coordination positive maximale entre deux enquêtes lorsque les zones de sélection correspondantes ont un recouvrement maximal, c'est-à-dire lorsqu'elles sont incluses l'une dans l'autre. On obtient au contraire une coordination négative maximale lorsque les zones de sélection se recouvrent le moins possible, c'est-à-dire, dans la mesure du possible, où elles sont disjointes. Ainsi, la construction de ces zones de sélection pour chaque enquête est l'élément essentiel de notre méthode d'échantillonnage coordonné. Son fonctionnement est aisément compris en considérant un exemple. Comme les unités de la population sont traitées de manière indépendante, il suffit de considérer une unité

 ISSN: 2102-6238

générique.

Supposons que cette unité a des probabilités d'inclusions égales à π_k^1, π_k^2, π_k^3, et π_k^t respectivement à la première, deuxième, troisième et $t-$ième enquête.

1. Pour la première enquête, on choisit naturellement l'intervalle $[0, \pi_k^1[$ comme zone de sélection (voir figure 1).

FIGURE 1. *Zone de sélection pour la première enquête*

2. La deuxième enquête peut être soit positivement soit négativement coordonnée avec la première. Nous n'envisageons que des cas où l'on va chercher à avoir une coordination maximale. Si la coordination voulue est positive, la zone de sélection choisie pour la deuxième enquête est l'intervalle $[0, \pi_k^2[$ (voir par exemple, si $\pi_k^2 \leq \pi_k^1$, la figure 2). Dans

FIGURE 2. *Coordination positive, avec* $\pi_k^2 \leq \pi_k^1$

la situation représentée par la figure 2, si u_k est dans l'intervalle $[0, \pi_k^2[$, alors l'unité k est sélectionnée à la première et à la deuxième enquête. Si ce nombre est entre π_k^2 et π_k^1, alors k est sélectionnée à la première mais pas à la deuxième enquête, et si u_k est plus grand que π_k^1, alors k n'est pas sélectionnée. La probabilité de sélectionner k dans les deux enquêtes est donc égale au minimum de π_k^1 et π_k^2, ce qui est le maximum théorique.

Si au contraire, la coordination souhaitée est négative, deux cas sont possibles :

– Si la probabilité d'inclusion π_k^2 est telle que $\pi_k^1 + \pi_k^2 \leq 1$, alors il est possible de faire en sorte que l'unité k ne soit prise que dans un seul des deux échantillons. Pour cela, on choisit comme zone de sélection à la deuxième enquête l'intervalle $[\pi_k^1, \pi_k^1 + \pi_k^2[$ (voir figure 3).
– Si $\pi_k^1 + \pi_k^2 > 1$, on ne pourra exiger que l'unité k ne soit jamais sélectionnée dans les deux échantillons. La probabilité que k soit prise conjointement dans les deux échantillons est alors au minimum égale à $\pi_k^1 + \pi_k^2 - 1$. Ce minimum est obtenu en utilisant la zone de sélection $[\pi_k^1, 1] \cup [0, \pi_k^1 + \pi_k^2 - 1[$ à la deuxième enquête (voir figure 4).

 ISSN: 2102-6238

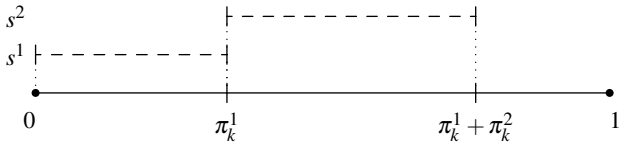

FIGURE 3. *Coordination négative lorsque* $\pi_k^1+\pi_k^2 \leq 1$

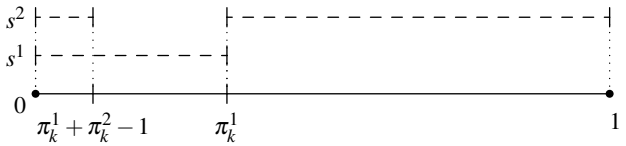

FIGURE 4. *Coordination négative lorsque* $\pi_k^1+\pi_k^2 \geq 1$

3. La troisième enquête, quant à elle, peut être positivement ou négativement coordonnée avec la première enquête, et également positivement ou négativement coordonnée avec la deuxième enquête. Ces exigences ne peuvent pas toujours être toutes satisfaites. Par exemple, si les deux premières enquêtes sont coordonnées positivement entre elles, et que l'on souhaite que la troisième soit coordonnée positivement avec l'une des deux, et négativement avec l'autre, on ne peut espérer un très bon résultat pour ces deux coordinations. Un autre exemple est le cas où $\pi_k^1 = \pi_k^2 = \pi_k^3 = 0.5$, et où l'on veut que les enquêtes soient négativement coordonnées. On pourra bien faire en sorte que k ne puisse être sélectionné simultanément dans le premier et dans le deuxième échantillon. Le recouvrement entre le troisième échantillon et l'un au moins des deux premiers est par contre inévitable. Il faut donc choisir quelle coordination l'on veut privilégier.

   La méthode utilisée à l'OFS repose sur un choix de l'ordre de priorité pour les coordinations entre une nouvelle enquête et les enquêtes passées. Un exemple suffit à comprendre son fonctionnement : supposons que les deux premières enquêtes étaient positivement coordonnées, comme dans la figure 2, et que l'on souhaite que la troisième enquête soit positivement coordonnée avec la seconde, puis, avec une priorité moindre, négativement coordonnée avec la première. Supposons de plus que $\pi_k^3 > \pi_k^2$. Le premier objectif est atteint si la zone de sélection pour la troisième enquête a le plus grand recouvrement possible avec la zone de sélection pour la deuxième enquête. Puisque $\pi_k^3 > \pi_k^2$, on va choisir d'inclure l'intervalle $[0, \pi_k^2[$ dans cette zone de sélection. Puis, il faut compléter cette zone de sélection de manière à ce que sa longueur totale soit égale à π_k^3, et à respecter au mieux les coordinations voulues. On va naturellement choisir d'ajouter l'intervalle $[\pi_k^1, \pi_k^3 + \pi_k^1 - \pi_k^2[$ comme représenté sur la figure 5.

4. Dans le cas général, après $t-1$ enquêtes, le segment $[0,1]$ est divisé en t intervalles, qui sont les intersections des zones de sélection des précédentes enquêtes. Les intervalles

 ISSN: 2102-6238

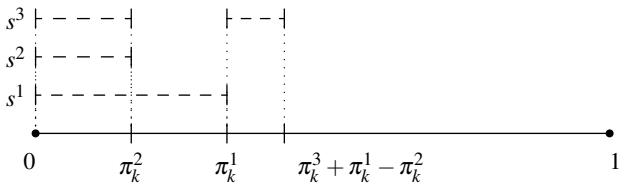

FIGURE 5. *Coordination d'un troisième échantillon*

correspondent aux séquences de sélection possibles pour l'unité k aux $t-1$ enquêtes passées. On veut sélectionner un $t-$ième échantillon qui doit être négativement coordonné avec certaines enquêtes passées et positivement coordonné avec d'autres enquêtes passées. Les exigences de coordination pour la nouvelle enquête permettent de définir un ordre sur les t intervalles. Si les coordinations avec toutes les enquêtes passées sont spécifiées alors cet ordre est total, et permet de classer les intervalles, et séquences de sélection correspondantes par ordre de compatibilité avec les règles de coordination demandées.

Une manière de calculer cet ordre est présentée dans (Qualité, 2009, p.105) : pour chaque unité k (indice omis par la suite) et chaque intervalle a_i, $i = 1,\ldots,t$, on calcule un score $S(a_i)$. Ce score est défini par :

$$S(a_i) = \sum_{j=1}^{t-1} 2^{j-1} c_{t-j}(a_i),$$

où $c_{t-j}(a_i) = 1$ si l'intervalle a_i est compatible avec la coordination souhaitée pour l'enquête qui a la $t-j$-ième priorité, et 0 sinon. Par exemple, si l'on veut une coordination positive avec l'enquête qui a la priorité 1 (la priorité la plus élevée), on définit $c_1(a_i) = 1$ si l'intervalle a_i correspond à une sélection à cette enquête, et 0 si a_i correspond à une non-sélection à cette enquête. Si au contraire on veut une coordination négative, on définit $c_1(a_i) = 0$ si l'intervalle a_i correspond à une sélection à cette enquête, et 1 si a_i correspond à une non-sélection à cette enquête. De cette manière, les intervalles a_i tels que $c_1(a_i) = 1$ auront tous un score $S(a_i)$ supérieur ou égal à 2^{t-2} et ceux tels que $c_1(a_i) = 0$ auront tous un score $S(a_i)$ strictement inférieur à 2^{t-2}. On peut montrer que la fonction score $S(\cdot)$ ainsi définie est une injection dans $\{0,\ldots,2^{t-1}-1\}$, et que l'ordre induit par ce score est compatible avec l'ordre de priorité voulu sur les coordinations. En pratique, cette méthode de calcul est assez rapidement problématique puisque les scores calculés peuvent être très grands. Il sera donc utile de trouver une méthode adaptée au langage informatique utilisé.

La zone de sélection pour la nouvelle enquête est obtenue en incluant les intervalles les mieux classés, jusqu'à dépasser une longueur totale de π_k^t. Le dernier intervalle inclus est ensuite découpé, et seulement une partie est conservée dans la zone de sélection, de manière à ce que celle-ci ait une longueur totale égale à π_k^t. Le segment $[0,1]$ est alors divisé en $t+1$ intervalles qui correspondent aux $t+1$ séquences de sélections possibles. La croissance linéaire du nombre d'intervalles à considérer est un des aspects de la méthode qui permet d'envisager son utilisation sur une période assez longue. Plus de 60 enquêtes auprès des

 ISSN: 2102-6238

environ 8 millions de personnes résidentes en Suisse ont déjà été sélectionnées en utilisant une implémentation assez grossière de cet algorithme.

Exemple. *Considérons une unité dans une population qui subit trois enquêtes : une enquête par panel s^1, une enquête unique s^2, un autre panel s^3, puis une mise à jour s^4 du premier panel et enfin une mise à jour s^5 du deuxième panel. Les probabilités d'inclusion de l'unité dans ces différents échantillons sont respectivement notées $\pi^1,\ldots,\pi^5$. Les deux panels et l'enquête ponctuelle sont, entre eux, coordonnés négativement avec des priorités définies par l'ancienneté des enquêtes. Par souci de simplicité, on suppose que la probabilité de tirage à l'ensemble de ces enquêtes est strictement inférieure à 1, que $\pi^4 > \pi^1$ et $\pi^5 < \pi^3$. Les intervalles et échantillons longitudinaux correspondants sont :*

1. *lors du premier tirage,*
   - $a_1 = [0,\pi^1]$ *avec* $s = 1$ *et*
   - $a_2 =]\pi^1, 1]$ *avec* $s = 0$,
2. *lors du deuxième tirage,*
   - $a_1 = [0,\pi^1]$ *avec* $s = (1,0)$,
   - $a_2 =]\pi^1, \pi^1+\pi^2]$ *avec* $s = (0,1)$ *et*
   - $a_3 =]\pi^1+\pi^2, 1]$ *avec* $s = (0,0)$,
3. *lors du troisième tirage,*
   - $a_1 = [0,\pi^1]$ *avec* $s = (1,0,0)$,
   - $a_2 =]\pi^1, \pi^1+\pi^2]$ *avec* $s = (0,1,0)$,
   - $a_3 =]\pi^1+\pi^2, \pi^1+\pi^2+\pi^3]$ *avec* $s = (0,0,1)$ *et*
   - $a_4 =]\pi^1+\pi^2+\pi^3, 1]$ *avec* $s = (0,0,0)$,
4. *lors du quatrième tirage,*
   - $a_1 = [0,\pi^1]$ *avec* $s = (1,0,0,1)$,
   - $a_2 =]\pi^1, \pi^1+\pi^2]$ *avec* $s = (0,1,0,0)$,
   - $a_3 =]\pi^1+\pi^2, \pi^1+\pi^2+\pi^3]$ *avec* $s = (0,0,1,0)$,
   - $a_4 =]\pi^1+\pi^2+\pi^3, \pi^2+\pi^3+\pi^4]$ *avec* $s = (0,0,0,1)$ *et*
   - $a_5 =]\pi^2+\pi^3+\pi^4, 1]$ *avec* $s = (0,0,0,0)$,
5. *et lors du cinquième tirage,*
   - $a_1 = [0,\pi^1]$ *avec* $s = (1,0,0,1,0)$,
   - $a_2 =]\pi^1, \pi^1+\pi^2]$ *avec* $s = (0,1,0,0,0)$,
   - $a_3 =]\pi^1+\pi^2, \pi^1+\pi^2+\pi^5]$ *avec* $s = (0,0,1,0,1)$,
   - $a_4 =]\pi^1+\pi^2+\pi^5, \pi^1+\pi^2+\pi^3]$ *avec* $s = (0,0,1,0,0)$,
   - $a_5 =]\pi^1+\pi^2+\pi^3, \pi^2+\pi^3+\pi^4]$ *avec* $s = (0,0,0,1,0)$ *et*
   - $a_6 =]\pi^2+\pi^3+\pi^4, 1]$ *avec* $s = (0,0,0,0,0)$.

Les priorités définies sur les coordinations prennent toute leur importance lorsque certaines coordinations sont contradictoires (cas que nous n'avons jamais rencontré à l'OFS), ou lorsque la probabilité d'inclusion des unités dans l'ensemble des échantillons atteint 1. Considérons que l'on sélectionne un échantillon pour une sixième enquête, avec une coordination négative par ordre chronologique avec les anciennes enquêtes. L'ordre de préférence pour définir la zone de tirage dans s^6 est : a_6 (aucune sélection antérieure) puis a_2 (dernière sélection dans s^2) puis a_4 (dernière sélection dans s^3) puis a_5 et a_1 (dernière sélection dans s^4, mais une sélection de plus pour a_1), et

 ISSN: 2102-6238

enfin a_3 (dernière sélection dans s^5). Si $\pi^6 > 1 - \pi^2 - \pi^3 - \pi^4$, cet ordre de préférence permet de choisir lesquels des intervalles $a_1, \ldots, a_5$ sont inclus ou partiellement inclus dans la zone de sélection pour s^6 bien qu'ils ne respectent pas parfaitement toutes les coordinations voulues.

3.3. Sélection d'une seule personne par ménage

Afin d'obtenir un bon taux de réponse, il a été décidé à l'OFS de ne pas interroger plus d'une personne par ménage à une enquête donnée lorsque cela n'est pas nécessaire pour obtenir l'information nécessaire aux estimations. Les exceptions notables à l'OFS sont l'enquête sur le budget des ménages (EBM) et l'enquête sur les revenus et conditions de vie (SILC), où, au contraire, tous les membres des ménages sélectionnés sont interrogés. Jusqu'à fin 2013, SILC et l'EBM n'étaient pas sélectionnées dans le fichier SRPH et nous n'avons pas été confrontés à ce dernier problème.

La méthode de tirage avec coordination décrite dans la section 3.2 ne permet pas d'obtenir directement ce résultat. En effet, elle repose sur des sélections indépendantes entre les unités, et ne permet donc pas d'exclure la sélection simultanée de deux personnes d'un même ménage, ou au contraire de forcer la sélection simultanée de toutes les personnes d'un même ménage. Les enquêtes pour lesquelles on s'est interdit de sélectionner plusieurs personnes par ménage sont l'enquête structurelle, l'enquête Omnibus et les enquêtes thématiques. Les échantillons des deux premières sont sélectionnés en une fois dans un seul cadre de sondage tandis que les enquêtes thématiques sont sélectionnées par blocs dans trois ou quatre fichiers SRPH successifs. Pour ces dernières, il faut ainsi considérer la difficulté supplémentaire que la composition des ménages peut changer entre les tirages des différents blocs.

Pour obtenir le résultat voulu, on a choisi comme principe de sélectionner les échantillons selon une procédure de tirage en plusieurs phases. La première phase consiste en la sélection coordonnée d'un échantillon de personnes, selon la méthode de la section 3.2, et avec des probabilités d'inclusion à déterminer. Les phases suivantes consistent à éliminer les cas indésirables en :

- ne retenant qu'une unité dans chaque ménage où plusieurs ont été sélectionnées en première phase,
- et pour les enquêtes thématiques en éliminant toutes les sélections dans des ménages dont un membre participe déjà à l'enquête.

Afin de rendre ce procédé le plus simple possible, le choix de la personne retenue lors de la deuxième phase est effectué à probabilités égales parmi les unités d'un même ménage présélectionnées en première phase. Les probabilités d'inclusion à la première phase sont calculées de manière à obtenir les probabilités d'inclusion finales souhaitées sur l'ensemble des deux ou trois phases.

3.3.1. Enquête structurelle et enquête Omnibus

Les personnes d'un même ménage qui sont sélectionnables à ces enquêtes reçoivent des probabilités de sélection égales. En effet, les probabilités d'inclusion dans les enquêtes de l'OFS ne sont fonctions que du canton ou de la commune de résidence, et de l'appartenance à la population cible de l'enquête.

Le calcul explicite des probabilités de sélection de première phase en fonction des probabilités d'inclusion finales souhaitées est alors possible. Notons p la probabilité d'inclusion de première

 ISSN: 2102-6238

phase de chacune des m personnes sélectionnables d'un ménage. Si la deuxième phase consiste en la sélection avec probabilités égales d'une seule personne parmi celles sélectionnées en première phase, on obtient que la probabilité d'inclusion finale π d'une de ces personnes est égale à

$$\pi = \frac{1}{m}\left[1-(1-p)^m\right]. \tag{1}$$

En effet, la probabilité de sélectionner au moins une personne dans le ménage est égale à $1-(1-p)^m$. Les probabilités conditionnelles de sélectionner l'un ou l'autre individu en deuxième phase sachant que le ménage contient des unités présélectionnées en première phase sont égales (les individus sont échangeables), et la somme de ces probabilités conditionnelles vaut 1 (on retient exactement une unité par ménage dans lequel des unités ont été présélectionnées). Chacune de ces probabilités conditionnelles vaut donc $1/m$. On obtient l'équation 1 en multipliant ces deux termes.

On peut déduire le paramètre p de la probabilité d'inclusion π souhaitée, en inversant l'équation 1, pour autant que $m\pi \leq 1$. La probabilité d'inclusion en première phase vaut alors

$$p = 1-(1-m\pi)^{\frac{1}{m}}. \tag{2}$$

Pour les très grands ménages dans les cantons qui ont choisi de financer une augmentation de leur échantillon pour l'enquête structurelle, la valeur de $m\pi$ peut dépasser 1. En effet, le taux de sondage de base des personnes π est, dans ces cantons, proche de 8%. Or, dans quelques ménages, m est supérieur ou égal à 13. Il n'y a alors pas de solution au problème. Nous avons simplement attribué une probabilité d'inclusion de première phase égale à 1 aux personnes de ces ménages.

Le plan de sondage obtenu est exactement identique à un plan de sondage en deux phases où la première phase serait une sélection de ménages à l'aide d'un plan de Poisson, avec probabilités d'inclusion proportionnelles au nombre de personnes sélectionnables du ménage, et la deuxième phase la sélection avec probabilités égales d'une personne par ménage retenu. La raison pour laquelle nous n'avons pas directement fait cela est que la constitution des ménages dans le cadre de sondage, et en particulier le suivi de ces ménages dans le temps, n'étaient pas assurées entre septembre 2010, date de création du premier registre d'échantillonnage SRPH, et décembre 2012. En effet, l'obligation légale, pour les cantons et communes, de fournir un identifiant de ménage pour chaque personne n'entrait en vigueur qu'au 31 décembre 2012. Il paraissait ainsi plus sage d'organiser la coordination des échantillons directement au niveau des personnes.

La méthode utilisée pour n'obtenir qu'une sélection par ménage conduit à une coordination sous-optimale entre les enquêtes. En effet, la coordination entre échantillons est réalisée entre les échantillons de première phase, dont les probabilités d'inclusion sont légèrement plus élevées que les probabilités d'inclusion finales. Pour la grande majorité des enquêtes, et des ménages cela ne fait pas une différence sensible. Pour les très grands ménages (plus de 13 personnes dans un canton qui aurait doublé son échantillon pour l'enquête structurelle), la coordination entre l'enquête structurelle et les autres enquêtes est inexistante.

3.3.2. Enquêtes thématiques

Le problème est plus complexe pour les enquêtes thématiques qui sont sélectionnés dans plusieurs cadres. Pour les ménages dont la composition reste identique lors des différents tirages, on peut

 ISSN: 2102-6238

assez simplement trouver une équation similaire à (1) pour relier les probabilités d'inclusion dans les différents blocs d'enquête et les probabilités d'inclusion finales. L'équation obtenue permet, lorsqu'une solution existe, de calculer explicitement les paramètres à utiliser lors de chaque tirage. Comme dans le cas des enquêtes structurelles et omnibus, il peut y avoir des problèmes pour certains très grands ménages liés à l'impossibilité de respecter les probabilités d'inclusion souhaitées tout en imposant au maximum une sélection par ménage sur l'ensemble des blocs d'enquête.

Dans les ménages dont la composition évolue, les probabilités d'inclusion des membres du ménage dans chaque bloc d'enquête ne sont pas nécessairement toutes égales. Par exemple, si le ménage contient lors du deuxième tirage une personne qui n'en faisait pas partie lors du premier tirage, il est possible que celle-ci ait eu une probabilité de sélection dans le premier bloc différente de celle des autres membres du ménage. Les calculs deviennent alors plus complexes. En effet, le polynôme (1) est alors remplacé, pour chaque personne, par un polynôme fonction des probabilités de première phase p_k de tous les membres du ménage à chaque bloc d'enquête passé. Le degré de ce polynôme est égal au nombre de personnes sélectionnables dans le ménage. Le problème est de calculer les racines de ces polynômes. On ne peut vraisemblablement pas obtenir une solution explicite dans les ménages où plus de quatre personnes sont sélectionnables. En effet, les polynômes de degré cinq ou plus ne sont en général pas résolubles par radicaux. Pour cette raison, nous avons développé une procédure de calcul des probabilités de première phase par approximations numériques successives. Cette procédure repose sur une version simplifiée de l'algorithme de Newton-Raphson dans lequel la différentielle de la fonction à inverser est remplacée par l'identité. En effet, cette différentielle est usuellement très proche de l'identité, et la perte d'efficacité de l'algorithme due à cette approximation est certainement largement compensée par l'économie de calculs réalisée à chaque étape.

Il faut noter que malgré tous nos efforts, la sélection d'échantillons qui respectent parfaitement ces exigences n'est en réalité pas possible car la composition des ménages est manquante pour 2 à 3% de la population dans le cadre d'échantillonnage. De plus, cette composition est susceptible de changer entre la date de constitution du fichier SRPH et la date d'enquête. Toutefois, l'objectif peut être considéré comme essentiellement rempli si l'on ne sélectionne qu'une personne par ménage selon l'état de nos connaissances.

4. Estimation et Pondération

Les estimateurs utilisés par l'OFS sont en principe des estimateurs par dilatation, c'est à dire des sommes pondérées des valeurs observées, ou bien des fonctions d'estimateurs par dilatation, par exemple des ratios de deux estimateurs. Un rôle central est joué par l'estimateur de Horvitz and Thompson (1952) d'un total

$$\widehat{Y}_{HT} = \sum_{k \in s} d_k y_k, \tag{3}$$

où les y_k sont les valeurs de la variable dont on veut estimer le total et les $d_k = 1/\pi_k$ sont les inverses des probabilités d'inclusion des individus k dans l'échantillon s. Cet estimateur n'est cependant pratiquement jamais directement utilisé, mais il sert de base pour construire l'estimateur pondéré final

$$\widehat{Y}_w = \sum_{k \in r} w_k y_k,$$

 ISSN: 2102-6238

où r est un sous-ensemble de répondants de s et $w_k \geq 0$.

Les poids d'extrapolation w_k sont calculés en trois, voire quatre grandes étapes. La première étape est le calcul des probabilités d'inclusion π_k, qui ne sont parfois pas connues avant l'enquête. La deuxième est la modélisation du mécanisme de non-réponse, et l'estimation de la probabilité de réponse de chaque unité observée. La troisième est le calage (Deville and Särndal, 1992). Suivant les enquêtes, une étape supplémentaire consiste à effectuer un partage des poids (Lavallée, 2002) lorsque les unités enquêtées ne sont pas directement les unités sélectionnées, ou bien une combinaison de poids lorsqu'un échantillon est obtenu par réunion de plusieurs autres échantillons. Chaque enquête nécessite le développement d'une pondération qui lui est propre dépendant de plusieurs facteurs : plan d'échantillonnage, procédé ou mode d'enquête, thématique, informations à disposition externes et internes à l'enquête.

4.1. Calcul des probabilités d'inclusion

Les tailles d'échantillons pour les enquêtes de l'OFS sont choisies en fonction de contraintes budgétaires et d'objectifs de précision. La répartition de ces échantillons dans les cantons est effectuée de manière équitable, c'est à dire proportionnellement à la population des cantons. Ces derniers ont la possibilité de financer des extensions de leur échantillon lorsque celui prévu par l'OFS est jugé insuffisant pour répondre à leurs besoins. Le plan de sondage naturel pour ces enquêtes est le sondage aléatoire simple stratifié (Tillé, 2006). Il n'est cependant en pratique quasiment jamais mis en œuvre : parfois le cadre de sondage ne contient pas l'information suffisante. Par exemple, pour sélectionner l'échantillon de l'Espa dans le répertoire Castem, on ne connaît pas les tailles des ménages à l'avance. Parfois l'échantillon est obtenu en agrégeant des échantillons sélectionnés dans des cadres différents. C'est le cas pour SILC, EBM, et les enquêtes thématiques. Enfin, le plan de sondage utilisé pour les tirages coordonnés dans le SRPH est le plan de Poisson. Les probabilités d'inclusion ne sont en conséquence pas toujours égales au rapport entre la taille de l'échantillon et la taille de la population. Elles doivent être calculées pour chaque enquête.

Ce calcul, point de départ de la construction des estimateurs peut être relativement compliqué en raison de la procédure d'enquête, ou du cadre de sondage utilisé pour atteindre la population cible. Dans les cas les plus simples, les probabilités d'inclusion n'ont pas à être calculées car elles sont spécifiées dans le plan de sondage. C'est par exemple le cas pour une enquête ponctuelle sélectionnée dans le SRPH, telle l'enquête Omnibus du système de recensement. C'est également le cas pour l'enquête EBM, dont l'échantillon de ménages est obtenu en sélectionnant un nombre fixé de numéros de téléphones dans le registre de numéros de téléphone Castem. Dans d'autres cas, les probabilités d'inclusion sont calculables à partir des cadres de sondage. Par exemple, pour les enquêtes thématiques qui sont constituées de vagues d'enquêtes sélectionnées dans trois ou quatre cadres d'échantillonnage SRPH successifs, la probabilité d'inclusion d'une unité est simplement la somme de ses probabilités d'inclusion aux trois ou quatre vagues de l'enquête. Ceci est dû à la coordination négative entre les vagues qui assure qu'elles ne se recouvrent pas.

Une spécificité de l'enquête structurelle est que le questionnaire comprend une clause qui permet aux personnes enquêtées de décider laquelle doit répondre dans le cas où, malgré les procédures mises en place pour l'éviter, plusieurs personnes d'un même ménage reçoivent un questionnaire. Cette procédure, déterministe, repose sur les dates de naissance des personnes du

 ISSN: 2102-6238

ménage. Elle conduit, dans certains cas à faire diminuer la probabilité de sélection d'une personne lorsque son ménage effectif contient des personnes qui n'y sont pas renseignées dans le cadre d'échantillonnage SRPH. Elles pourraient en effet avoir été sélectionnées en même temps, et, selon leur date de naissance, elles auraient alors été désignées par la procédure d'élimination pour répondre au questionnaire. La procédure d'élimination n'a pas toujours été comprise, mais les cas où elle a été appliquée et les cas où elle aurait dû l'être sont très peu nombreux (de l'ordre de quelques centaines). Les probabilités d'inclusion des répondants ont été recalculées en fonction de la composition de leur ménage déclarée à l'enquête et de sa comparaison avec la composition connue dans le cadre. Il faut toutefois noter que cela fait dépendre les probabilités de sélection, dans l'échantillon des personnes qui sont censées répondre au questionnaire, de la composition réelle des ménages déclarée à l'enquête et de leur compréhension ou bonne application des instructions qui accompagnent le questionnaire. En effet, ce sont les personnes contactées elles-même qui ont en charge d'effectuer l'ultime phase de filtrage de l'échantillon pour éliminer les cas de sélections multiples dans un ménage. La probabilité d'inclusion des ménages est elle aussi calculable en utilisant la composition déclarée des ménages observés, et en la rapprochant des probabilités de tirages connues pour toutes les unités du cadre de sondage.

Les différents panels qui constituent l'échantillon de l'enquête SILC sont des échantillons aléatoires simples de ménages, sélectionnés sans remise et stratifiés par régions géographiques. Les probabilités de tirage par strate sont proportionnelles à la taille relative de la population de chacune des régions. Tous les membres adultes des ménages contactés sont interrogés, soit directement, soit au travers d'une personne du ménage. L'unité d'enquête est donc tant le ménage que l'individu. SILC étant un panel rotatif, chaque année, le panel vieux de 4 ans est abandonné et un nouveau panel rejoint le relevé pour le remplacer. Plusieurs pondérations sont produites : une pondération pour extrapoler les informations au niveau des ménages une année donnée, une pondération pour extrapoler les informations au niveau des personnes une année donnée. Des poids dits "longitudinaux" sont aussi calculés pour les unités qui sont observées plusieurs années à la suite (Graf, 2008). Les différentes probabilités d'inclusion : dans l'un des panels qui constituent l'échantillon annuel, dans l'un des panels qui sont enquêtés deux années à la suite, et ainsi de suite, sont donc évaluées (voir figure 6).

L'échantillon de l'enquête Espa est un panel rotatif de personnes. Il est constitué chaque trimestre de quatre blocs de rotation, dont un est enquêté pour la première fois, et les trois autres ont été enquêtés pour la première fois respectivement le trimestre précédent, un an avant, et quinze mois auparavant (Renfer, 2009; OFS, 2012). Ces blocs sont eux-mêmes sélectionnés selon une procédure mixte. Une partie de l'échantillon est obtenue en tirant des numéros de téléphone dans le répertoire Castem. La liste des personnes de chaque ménage contacté est alors établie interactivement, et l'une des personnes âgées de 15 ans ou plus est sélectionnée aléatoirement pour répondre à l'enquête (Feusi Widmer, 2004). Les probabilités d'inclusion correspondantes sont donc fonction de la taille des ménages. L'autre partie de l'échantillon est sélectionnée directement dans le registre du système d'information central sur la migration. Les probabilités d'inclusion dans chacun des blocs, ainsi que dans leur union un trimestre donné, où sur une période donnée sont évaluées. Il ne peut s'agir là que d'approximations car l'OFS ne dispose pas de toute l'information nécessaire : la composition des ménages n'est connue qu'au moment de l'enquête. On ne sait donc par exemple pas précisément avec quelle probabilité un individu du nouveau bloc de rotation aurait pu être sélectionné dans le bloc de rotation qui répond pour la

 ISSN: 2102-6238

quatrième fois.

Ces quelques exemples montrent qu'il existe une diversité de situations et de besoins qui conduisent à des calculs différents. La création du cadre d'échantillonnage SRPH et la généralisation de son utilisation pour toutes les enquêtes de l'OFS devrait permettre d'harmoniser ces calculs, et de les simplifier grâce à la connaissance précise de la population et de sa probabilité de sélection dans chaque vague d'enquête ou bloc de rotation pour les enquêtes concernées.

4.2. Calcul de probabilités de réponse

Toutes les enquêtes auprès de la population sont entachées de non-réponse, et seule une partie $r \subseteq s$ de l'échantillon sélectionné s répond à une enquête donnée. Cette non-réponse correspond effectivement à un sondage, dont on ne connaît pas le plan, dans l'échantillon sélectionné à l'origine. Pour utiliser l'estimateur de Horvitz-Thompson (3) avec les données de l'échantillon observé r, on doit connaître les probabilités d'inclusion dans r. En effet, les valeurs π_k ne sont pas égales aux probabilités d'appartenance à r, et l'estimateur

$$\widehat{Y} = \sum_{k \in r} \frac{y_k}{\pi_k} \tag{4}$$

est généralement biaisé. Pour que l'estimateur de Horvitz-Thompson soit sans biais, il faut en outre que les probabilités d'inclusion dans r soient toutes strictement positives. On est donc amené à modéliser le processus de réponse, et à estimer les probabilités de réponse des unités sélectionnées pour construire un estimateur qui ressemble à (3).

On désigne par la suite $q(\cdot|s)$ le mécanisme de réponse dans l'échantillon s. C'est la loi de probabilité, inconnue, qui a généré l'échantillon de répondants r parmi les s unités sélectionnées. Le fait qu'une unité sélectionnée ait effectivement répondu est indiqué par une variable aléatoire binaire r_k, indicatrice de réponse (0=pas de réponse, 1=réponse), pour $k \in s$. Il s'agit d'une variable de Bernoulli. Son espérance a priori inconnue est notée θ_k.

Si l'on connaissait les θ_k, on pourrait utiliser l'estimateur de Horvitz-Thompson

$$\widehat{Y}_{HT} = \sum_{k \in r} \frac{y_k}{\pi_k \theta_k}.$$

Dans la pratique, les θ_k sont toujours inconnus, et on les estime en supposant que $q(\cdot|s)$ appartient à une famille de lois spécifiée. Une fois les estimations $\hat{\theta}_k$, $k \in r$ réalisées, on peut estimer le total Y par

$$\widehat{Y} = \sum_{k \in r} \frac{y_k}{\pi_k \hat{\theta}_k}, \tag{5}$$

qui n'est au mieux qu'approximativement sans biais. Dans ce cas le facteur $(\pi_k \hat{\theta}_k)^{-1}$ joue le rôle de poids pour l'unité observée k. On l'appelle poids de non-réponse et on parle aussi de repondération pour compenser la non-réponse.

La modélisation de $q(\cdot|s)$ est une étape déterminante de la pondération. On suppose souvent que le plan $q(\cdot|s)$ est un plan de Poisson, c'est à dire que la décision de répondre ou pas d'une unité est indépendante de la décision prise par les autres unités. Cela est pertinent dans une enquête auprès des personnes où la probabilité de sélectionner deux personnes dans le même ménage

 ISSN: 2102-6238

est extrêmement faible, voire nulle par conception. Dans une enquête auprès des ménages, la non-réponse est dans un premier temps modélisée au niveau du ménage. Cette hypothèse n'est pas suffisante pour pouvoir estimer les θ_k puisque l'on ne dispose que d'une observation de r_k par paramètre θ_k à estimer. On est donc amené à spécifier des contraintes sur les θ_k afin d'arriver à une situation où ceux-ci peuvent être estimés avec précision. Deux types de contraintes, de nature voisine, sont usuellement envisagées :

1. Les θ_k ont une certaine relation fonctionnelle avec des variables explicatives connues pour les répondants et les non-répondants à l'enquête. Ils sont alors généralement estimés en effectuant une régression logistique des valeurs observées des r_k sur les variables explicatives retenues. C'est la méthode utilisée pour les enquêtes EBM et Espa.
2. La population est partitionnée en groupes de personnes qui partagent la même valeur de θ_k : les Groupes Homogènes de Réponse (GHR). La répartition de l'échantillon dans ces groupes est estimée par une méthode de segmentation (Kass, 1980), en fonction de variables explicatives connues pour les répondants et les non répondants à l'enquête. Les θ_k sont alors estimés par les taux de réponse observés dans les GHR. Cette méthode est utilisée à l'OFS pour l'enquête SILC.

Selon la constitution du processus d'enquête, plusieurs modélisations de la non-réponse peuvent être requises. C'est par exemple le cas quand, dans une collecte téléphonique, une description du ménage est demandée avant de sélectionner une personne pour l'enquête. Il peut y avoir alors non-réponse avant l'énumération du ménage, ou non-réponse de la personne sélectionnée. L'information disponible pour modéliser la première non-réponse est moins riche que celle disponible pour modéliser la seconde.

La première étape de la modélisation de la non-réponse consiste à déterminer quelles sont les variables explicatives qui doivent être considérées dans le modèle afin de bien prédire la variable dichotomique de réponse (voir LarRoche, 2007). Le choix des variables est parfois assez restreint car elles doivent être connues tant pour les unités répondantes que pour les unités non-répondantes. Pour les enquêtes sélectionnées dans le répertoire Castem, le cadre de sondage contient très peu d'information utilisable. Il est donc essentiel de collecter le maximum d'informations même sur les ménages ou personnes qui décident de ne pas répondre au questionnaire. Pour les enquêtes sélectionnées dans le SRPH, on dispose au contraire d'un nombre important de variables explicatives possibles, mais leur pouvoir explicatif n'est lui pas nécessairement très bon. Il s'agit ensuite de dresser une liste des variables disponibles candidates à décrire le processus de non-réponse. La nature de cette liste varie selon les thèmes abordés par l'enquête, le mode de collecte et le plan d'échantillonnage. Tout facteur susceptible d'expliquer la non-réponse doit être pris en compte dans la mesure du possible. Les variables peuvent êtres des grandeurs mesurées dans l'enquête elle-même ou des informations auxiliaires obtenues par d'autres biais (par exemple en appariant l'échantillon avec des registres) sur les unités répondantes et non-répondantes. Dans les enquêtes répétées on a en principe un choix de variables beaucoup plus grand pour modéliser la non-réponse dès la deuxième vague car on peut alors se reposer sur des informations récoltées en première vague.

La régression logistique peut s'appuyer sur une procédure de sélection de variables pour obtenir la liste des variables explicatives les plus intéressantes afin de prédire la réponse des unités sélectionnées. Dans le cas de variables catégorielles, le croisement des modalités des variables

 ISSN: 2102-6238

retenues forme alors de facto des groupes homogènes de réponse.

4.2.1. Segmentation et régression logistique

Le choix entre une modélisation de la non-réponse par segmentation ou par régression logistique n'est pas simple. Dufour et al. (1998) ont montré sur deux études empiriques similaires à SILC que la méthode de segmentation était meilleure que la régression logistique pour créer des GHR. Effectivement, la méthode de segmentation est plus flexible que la régression logistique. La régression logistique force l'arbre de décision à être symétrique : une fois que les variables significatives sont déterminées, les GHR sont formés en prenant toutes les intersections possibles entre ces variables. Le modèle est donc nécessairement symétrique, et de petits GHR éventuellement non pertinents peuvent être formés. Cela augmente inutilement la variance des estimateurs. La modélisation par segmentation utilise quant à elle un processus itératif pour partitionner le fichier de données. Elle détermine à chaque embranchement de l'arbre de décision quelle variable est la plus significative, ce qui garantit que chaque GHR formé est pertinent. Ce procédé conduit habituellement à un modèle asymétrique. Si pour un nœud donné, aucune variable n'est significative pour expliquer la non-réponse, ou si l'un des critères d'arrêt de l'algorithme est rencontré, aucun nouvel embranchement n'est créé. La segmentation présente l'avantage de pouvoir utiliser un plus grand nombre de variables pour modéliser la non-réponse tout en arrivant à un nombre de GHR réduit. Plusieurs méthodes peuvent être employées pour déterminer quelle variable semble le plus influencer la réponse (Nakache and Confais, 2003).

Selon LarRoche (2007), un inconvénient de la modélisation par segmentation est directement relié à l'arbre de décision lui-même. L'arbre de décision formé peut devenir très grand, comporter plusieurs embranchements, et se montrer difficile à interpréter. De plus, les arbres de décisions sont aussi assez instables : l'ajout, le retrait ou le changement d'une variable ou d'une unité peut avoir un effet important sur l'arbre lui-même. La stabilité de l'arbre est d'autant plus précaire que le nombre de variables explicatives est grand ou que certaines variables sont fortement corrélées entre elles. Cependant, conformément à Rakotomalala (2005), nous avons pu constater que, sur les enquêtes que nous avons traitées à l'OFS, ces variations dans la structure des arbres n'ont que peu d'influence sur les poids et les estimations finales.

Dans les deux méthodes, on a la possibilité d'imposer des contraintes dans la formation des GHR, typiquement sur le nombre d'observations et le taux de réponse minimal dans chaque groupe. Ces choix ont une influence directe sur le nombre final de GHR et la taille de l'arbre de décision. Comme rapporté dans Graf (2009), l'une et l'autre méthode visent à trouver, pour chaque unité répondante, un facteur d'ajustement pour la non-réponse. Si les GHR sont définis de telle sorte que la non-réponse soit complètement uniforme à l'intérieur d'un groupe, alors le biais de (5) est négligeable (voir Tambay et al., 1998; Kalton and Kasprzyk, 1986).

4.2.2. Correction de la non-réponse et calage

L'estimation à partir de données d'enquêtes sujettes à de la non-réponse peut aussi être abordée par des méthodes de post-stratification et plus généralement de calage (Deville and Särndal, 1992). Dans certains cas, les résultats sont même strictement identiques. Le calage généralisé introduit dans Deville (2002) et la méthode "lune et étoile" de Särndal and Lundström (2005)

 ISSN: 2102-6238

(voir section 4.3.1) permettent d'utiliser des informations disponibles seulement sur l'échantillon et d'obtenir la pondération finale en une seule étape. La méthode "lune et étoile" est une méthode élégante, applicable dans le cas d'une pondération simple d'enquête où le nombre d'étapes de correction de la non-réponse n'est pas élevé, et où il n'y a pas de partage de poids ou de combinaison d'échantillons s'intercalant entre les stades de correction de la non-réponse et le calage final. Cette méthode a été appliquée pour pondérer l'Enquête Suisse sur la Santé (Graf, 2010).

Dans le cas de l'enquête structurelle, le calage exigé afin d'assurer la cohérence avec les totaux issus de l'exploitation exhaustive des registres est extrêmement complexe. Toutes les variables ou tous les croisements de modalités que l'on aurait envisagés pour tenter de modéliser la non-réponse ou constituer les GHR sont utilisés dans le calage. Il s'agit entre autres des croisements de lieu de résidence, sexe, de classe d'âge, d'état civil, de nationalité et permis de séjour. Dès lors on aurait pu considérer que la non-réponse est entièrement traitée par calage, suivant Särndal and Lundström (2005). Toutefois, on a décidé de ne pas complètement confondre la procédure de calage et la procédure de traitement de la non-réponse, mais simplement de simplifier cette dernière. Les coefficients $\widehat{\theta}_k$ sont estimés dans le cas du relevé structurel par les taux de réponse observés dans chaque canton. Cela correspond à un modèle de non-réponse homogène dans les cantons.

4.3. Calage

Les avantages des procédures de calage sont nombreux. Un calage peut permettre d'obtenir des estimateurs plus précis en profitant de la proximité entre une variable observée y_k et d'autres variables x_k dont on connaît les totaux sur la population. Il permet également d'obtenir des estimateurs qui reproduisent fidèlement des statistiques choisies, et ainsi de produire des séries de résultats cohérentes. On parle alors de calage cosmétique. Enfin, il peut être utilisé en complément ou à la place des méthodes présentées dans la section 4.2 pour calculer des poids qui prennent en compte la non-réponse à l'enquête. Le calage peut donc être utilisé pour réduire la variance, sans but statistique précis, ou même pour réduire le biais de non-réponse de l'estimateur (4). Dans la pratique, on est toujours confronté à un mélange de ces trois raisons.

Les méthodes de calage utilisées à l'OFS ont été développées et présentées dans l'article fondateur de Deville and Särndal (1992), et dans Deville et al. (1993); Le Guennec and Sautory (2002); Sautory (2003); Deville (2002); Kott (2006). Le principe du calage est, partant d'un jeu de poids d_k, $k \in r$, de trouver un nouveau jeu de poids w_k, le plus proche possible des poids de départ d_k pour une distance spécifiée, et qui vérifie certaines contraintes. Dans la méthode originelle de Deville and Särndal (1992), il s'agit, partant des poids d_k de l'estimateur de Horvitz-Thompson, d'obtenir que

$$\widehat{X}_w = \sum_{k \in r} w_k x_k^*$$

soit égal au total connu sur la population, pour certaines variables x_k^*, $k \in U$ choisis. L'estimateur de Horvitz-Thompson étant sans biais, la faible distance entre les d_k et les w_k assure que l'estimateur calé obtenu est peu biaisé. Le choix de la distance, autorisé dans le logiciel Calmar2 (Le Guennec and Sautory, 2002; Sautory, 2003) utilisé à l'OFS, n'a théoriquement pas une grande importance. Ce choix est effectué de manière à éviter d'obtenir quelques poids négatifs, trop

 ISSN: 2102-6238

grands ou trop éloignés des poids de départ. On cherche aussi souvent à éviter qu'il y ait des poids inférieurs à 1, ce qui est conceptuellement difficile à envisager, ou même inférieurs aux poids de départ quand la procédure est utilisée pour traiter de la non-réponse. En effet, on ne pourrait alors plus interpréter le rapport d_k/w_k comme une probabilité de réponse estimée.

4.3.1. Calage "lune et étoile"

La méthodologie développée dans Särndal and Lundström (2005) a été appliquée à l'OFS pour pondérer l'Enquête Suisse sur la Santé de 2007 (Graf, 2010). Elle distingue deux types d'information auxiliaire :

1. les variables x_k^* connues sur l'échantillon observé r et dont on connaît le total sur la population. Dans la terminologie de Särndal and Lundström (2005), il s'agit du cas "InfoU",
2. les variables x_k° connues sur l'échantillon sélectionné s, et dont on ne connaît pas, ou pas nécessairement le total sur la population. Il s'agit du cas "InfoS".

Lorsque les deux types d'informations sont disponibles on est dans le cas "InfoUS".

Le calage sur les totaux des variables "étoile" x_k^* correspond au calage classique de Deville and Särndal (1992). L'information apportée par les variables "lune" x_k° peut être utilisée pour adapter les poids à la situation de non-réponse. Le calage "lune" vise à trouver un jeu de poids w_k qui vérifie les contraintes :

$$\sum_{k\in r} w_k x_k^\circ = \sum_{k\in s} d_k x_k^\circ,$$

où $d_k = 1/\pi_k$ est l'inverse de la probabilité de sélection de k dans s. Un exemple tout à fait classique est celui du plan aléatoire simple de taille fixe n dans une population de taille N, avec pour seule variable de calage $x_k^\circ = 1$, et un échantillon observé de m répondants. Partant de $d_k = N/n$, la méthode conduit naturellement à utiliser le poids $w_k = N/m$. Le calage "lune et étoile" a l'avantage de permettre de calculer une pondération qui respecte simultanément les totaux connus au niveau de la population et au niveau de l'échantillon sélectionné.

Le choix des variables "lune" x_k° peut être effectué selon les principes décrits dans la section 4.2. En particulier, il peut s'agir d'indicatrices d'appartenance à des groupes homogènes de réponse obtenus par segmentation en utilisant des variables connues sur l'échantillon s, ou encore de variables retenues par une procédure de sélection lors d'une régression logistique des indicateurs de réponse r_k sur les différentes variables disponibles. Les variables "étoile" x_k^* devraient être "proches" des variables d'intérêt y_k. Leur choix peut être assisté par des procédures de sélection de variables adaptées à la régression linéaire, sur l'échantillon des répondants, des y_k sur les différentes variables x_k^* dont on connaît le total sur la population. Dans ce dernier cas, il semble utile d'utiliser une pondération temporaire qui tienne déjà compte de la non-réponse. Dans la pratique, comme dans le cas du relevé structurel, la liste des variables imposées pour le calage cosmétique peut déjà être assez étendue. Les procédures de sélection de variables peuvent alors servir à confirmer qu'aucun facteur significatif n'a été oublié, ou de guider le choix des interactions ou croisements de catégories utilisé. Enfin, il permet de repérer les variables qui ont le moins de justification statistique d'entrer dans le calage et d'en motiver l'élimination lorsque les demandes formulées par les unités de diffusion semblent vraiment excessives.

 ISSN: 2102-6238

4.3.2. Calage à deux niveaux

Dans les enquêtes auprès de la population, des résultats sont souvent produits au niveau des personnes et au niveau des ménages. Le calage simultané (Le Guennec and Sautory, 2002) permet de réaliser un calage à la fois sur des informations disponibles au niveau des personnes et sur des informations disponibles au niveau des ménages. En particulier, on peut exiger que les poids des personnes soient égaux au sein d'un même ménage. Des approches plus basiques ont été utilisées pour l'enquête SILC et pour le relevé structurel.

L'enquête SILC est une enquête ménage pour laquelle est produite à la fois une pondération ménages et une pondération personnes. Au moment de sa pondération, on ne disposait pas d'information précise sur la population des ménages, mais seulement sur la population des personnes.

Le relevé structurel comporte un questionnaire à remplir par tous les membres du ménage. Les répondants doivent en particulier fournir leur numéro de sécurité sociale (AVS), qui sert d'identifiant dans le cadre d'échantillonnage SRPH, indiquer leur langue principale et niveau de formation, et leurs liens familiaux avec les autres membres du ménage. Une pondération a donc été calculée pour pouvoir extrapoler ces informations au niveau des ménages et non plus des personnes. Cela a été rendu difficile par le fait que la formation des ménages dans le fichier SRPH est imparfaite, et que l'on n'a donc pas de totaux de calage à ce niveau-là.

Pour ces deux enquêtes, un calage a été réalisé au niveau des ménages sur des totaux connus seulement au niveau des personnes. Les caractéristiques des personnes ont dans un premier temps été sommées dans chaque ménage m pour constituer les variables de calage x_m^*.

4.4. Poids partagés et poids combinés

4.4.1. Poids partagés

La méthode du partage des poids (Lavallée, 2002) permet de pondérer un échantillon obtenu de manière indirecte. Elle nécessite l'existence et le comptage de liens entre la population sélectionnée et la population observée. Dans l'enquête SILC, les individus enquêtés n'ont pas tous été directement sélectionnés. En effet, on fait la distinction entre les individus présent dans les ménages sélectionnés lors de leur première vague d'observation, ce sont les individus longitudinaux ou OSM (Original Sample Member), et les individus, aussi enquêtés, qui font partie de leurs ménages seulement aux vagues d'enquête suivantes. Ces individus, qui se joignent à l'étude après la vague 1, n'ont été sélectionnés que de manière indirecte puisqu'ils ne faisaient pas partie de l'échantillon effectivement tiré pour la vague 1. A la différence des OSM, qui possèdent un poids résultant de leur probabilité de sélection et de réponse, les nouveaux individus, appelés cohabitants, ne possèdent pas de poids a priori.

La méthode du partage des poids permet d'attribuer un poids aux nouveaux cohabitants en fonction des poids des OSM présents dans le ménage. Le lien entre la population sélectionnée et la population observée est la présence d'OSM dans les ménages observés. Plusieurs partages de poids ont été réalisés pour obtenir des pondérations individuelles et des pondérations ménages (Graf, 2008, 2009). La pondération ménage est caractérisée par le fait que tous les individus d'un même ménage ont un poids final identique. Pour la vague 2 ou plus, Lavallée (2002) montre que

 ISSN: 2102-6238

les poids calculés par la méthode du partage des poids permettent d'estimer sans biais des totaux et propose un estimateur de leur variance qui tient compte du partage. Massiani (2013) relève que cet estimateur de variance n'est généralement pas sans biais, même lorsque les probabilités de réponse sont connues, et apporte une correction qui annule ce biais.

4.4.2. Poids combinés

Pour produire des estimations à partir de plusieurs enquêtes, on peut calculer des combinaisons convexes des estimations obtenues aux différentes enquêtes. Cela revient à assembler les échantillons en multipliant les poids individuels par les coefficients de la combinaison convexe.

Plusieurs enquêtes de l'OFS ont des échantillons constitués de blocs de rotation ou vagues d'enquête. Il s'agit des enquêtes thématiques, de l'Espa, de SILC, de l'EBM. Pour les enquêtes thématiques, seule une estimation valable pour l'année d'enquête est produite. La pondération de ces enquêtes est réalisée sans modification spécifique liée à cette organisation de collecte en vagues : les probabilités d'inclusion à l'échantillon complet sont calculées puis les probabilités de réponse, etc.

L'Espa produit des résultats trimestriels, basés sur quatre blocs de rotation, et des résultats annuels, basés sur une partie des seize blocs d'enquête d'une année donnée. La pondération pour les résultats trimestriels est elle aussi calculée en évaluant la probabilité d'inclusion de chaque personne à l'union des quatre blocs correspondants, sa probabilité de réponse et en effectuant un calage.

Les informations utiles pour les estimations annuelles ne sont fournies que par une partie des blocs d'enquête : ceux qui sont en première et en troisième vague d'interrogation. En effet, selon le schéma de rotation décrit dans Renfer (2009), les individus enquêtés une année donnée en vague 2 et en vague 4 sont les mêmes que ceux enquêtés en vague 1 et en vague 3 cette même année, à l'exception du bloc en vague 4 au premier trimestre. Il y a ainsi une très forte corrélation entre les valeurs relevées en vagues 1 et 3 et celles relevées en vagues 2 et 4, ce qui peut justifier d'écarter ces dernières de l'estimation annuelle. Seuls huit blocs d'enquête constituent donc l'échantillon de l'Espa pour les estimations annuelles. Sa pondération est réalisée en deux temps :

1. une nouvelle pondération trimestrielle est réalisée en ne gardant que les blocs en vagues 1 et 3, sur le même principe que celle calculée pour produire les estimations trimestrielles de l'Espa,
2. les échantillons des quatre trimestres sont assemblés en divisant les poids des individus par 4. Cela revient à imposer que les estimations annuelles sont les moyennes simples d'estimations trimestrielles basées sur une partie des données.

L'enquête SILC fait l'objet de plusieurs pondérations : on calcule une pondération pour le panel qui a été enquêté quatre fois, une pondération pour les deux panels qui ont été enquêtés les trois dernières années, une pondération pour les trois panels qui ont été enquêtés les deux dernières années et une pondération pour les quatre panels enquêtés l'année courante (voir figure 6).

La méthode utilisée pour combiner les données des différents panels a été développée pour deux panels par Merkouris (voir Merkouris, 2001; Lévesque and Franklin, 2000). Elle consiste à attribuer un facteur p_i à chacun des panels, avec $\sum_i p_i = 1$. Par exemple, l'estimateur transversal

 ISSN: 2102-6238

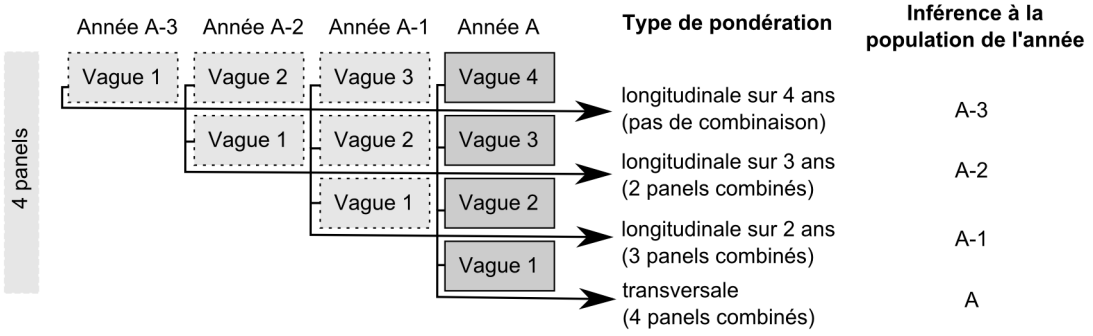

FIGURE 6. *Plan rotatif de l'enquête SILC.*

pour la dernière année est de la forme

$$\widehat{Y} = \sum_{i=1}^{4} p_i \widehat{Y}_i,$$

où $\widehat{Y}_i$ est l'estimation du total de la variable d'intérêt y_k obtenue en pondérant le panel i. La valeur optimale des coefficients p_i est calculée en minimisant la variance de $\widehat{Y}$. En pratique, un seul jeu de facteur est utilisé pour plusieurs variables d'intérêt y, et l'optimisation n'est faite que de manière approchée. Le facteur p_i est en pratique défini par

$$p_i = \frac{n_i d_i}{\sum_{i=1}^{4} n_i d_i}, \; i = 1, \ldots, 4$$

où d_i désigne l'effet de plan d'échantillonnage pour le panel i et n_i représente le nombre de personnes, ou de ménages, du panel i répondantes à la vague d'enquête considérée pour l'année en cours. Il s'agit de la vague 1 pour le nouvel échantillon entrant, de la vague 2 pour l'échantillon interrogé pour la deuxième année consécutive, de la vague 3 pour celui interrogé pour la troisième année consécutive et de la vague 4 pour l'échantillon interrogé pour la quatrième et dernière année consécutive. Comme les plans de sondage des différents panels sont identiques, on fait l'hypothèse que les effets de plan sont sensiblement les mêmes, et les p_i sont in fine proportionnels aux tailles d'échantillons.

5. Traitement des données et imputation

Les données d'enquête présentent quasiment systématiquement des défauts : les valeurs sont manquantes pour certaines variables et certaines unités, ou bien elles sont incohérentes, atypiques, suspectes. Le problème peut venir d'une mauvaise compréhension du questionnaire, d'une divergence de concepts, d'une incapacité à répondre à certaines questions, ou encore d'un refus de répondre. Il peut aussi être lié aux traitements effectués lors de la production du fichier de données à partir du résultat de la collecte. On peut citer par exemple les erreurs lors de la numérisation de questionnaires papiers, et les erreurs de traitement informatique. Ces défauts doivent être traités, dans la mesure du possible, et systématiquement documentés pour que les utilisateurs finaux

 ISSN: 2102-6238

soient informés des possibles lacunes de l'enquête, et que des indicateurs de qualité puissent être calculés. Cela fait l'objet du processus de préparation statistique des données.

5.1. Processus de préparation statistique des données

L'OFS suit les recommandations formulées dans le cadre des projets Euredit (E.U., 2003) et Edimbus (E.U., 2007) sur le traitement des données d'enquête et les imputations. Ces recommandations font l'objet d'une formation interne dispensée régulièrement aux collaborateurs de l'OFS (Graf and Kilchmann, 2010). Dans un schéma d'enquête, l'édition des données, l'imputation et la pondération participent au processus de préparation statistique des données (PPSD, voir Figure 7). Le PPSD a pour objet la description des données récoltées, l'amélioration de leur

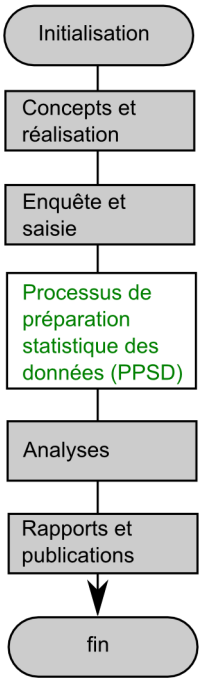

FIGURE 7. *Schéma d'enquête.*

qualité et de leur cohérence, et la préparation des données pour les analyses. À chacun de ces trois objets correspond une phase du PPSD, voir Figure 8. Chaque phase est constituée de plusieurs procédures, le tout pouvant être répété jusqu'à ce que les objectifs de qualité soient atteints. Les données initiales sont toujours conservées sans modifications, et des indicatrices sont créées pour repérer quelles valeurs des variables d'étude ont été générées par le PPSD.

La phase de préparation initiale des données permet de décrire la qualité des données, en créant des indicatrices de réponse pour chaque variable, des indicatrices de conformité à des règles devant être respectées par les valeurs observées. On peut aussi y trouver des procédures élémentaires d'imputation ou de modification automatique des données.

La phase de micro-préparation des données traite les observations au niveau de l'unité d'enquête,

 ISSN: 2102-6238

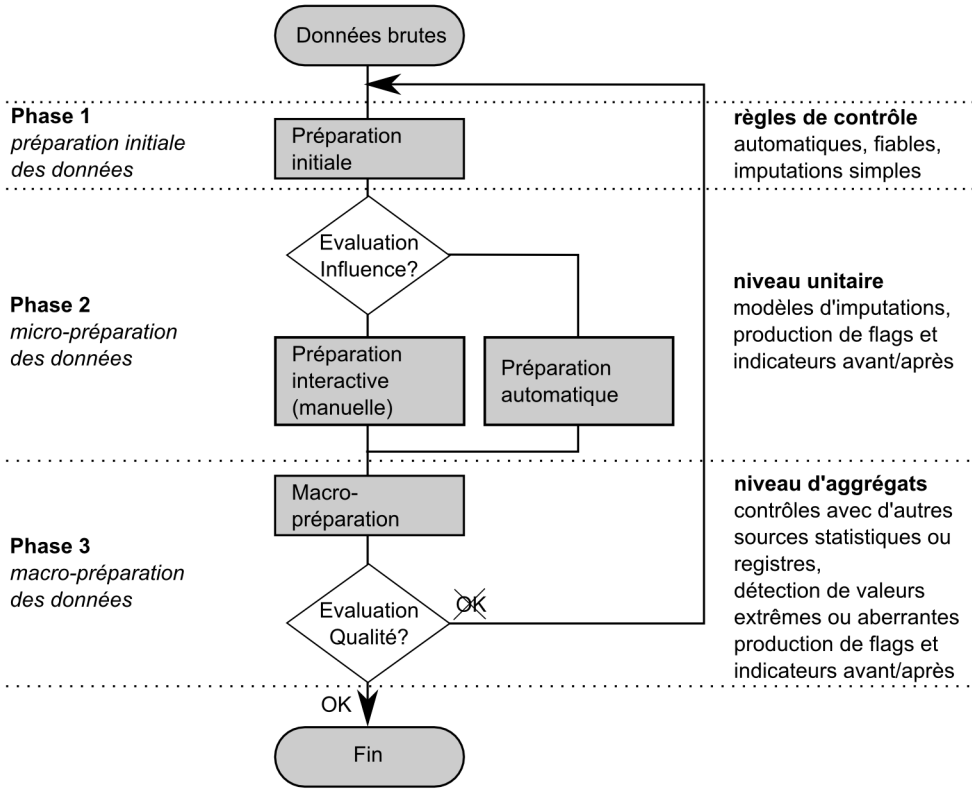

FIGURE 8. *Processus de préparation statistique des données.*

personne ou ménage. Elle s'attache à assurer la cohérence et la complétude des données pour chaque unité. Les traitements peuvent être automatisés ou bien interactifs, avec par exemple des comparaisons avec d'autres sources d'information, un retour vers l'unité enquêtée pour vérification des données, etc. Les traitements automatisés regroupent entre autre l'emploi de méthodes d'imputation. Les opérations réalisées dans cette phase ne doivent pas entrer en conflit avec les contrôles de qualité réalisés dans la phase de préparation initiale des données.

La phase de macro-préparation des données permet d'évaluer les données collectées en les comparant avec d'autres statistiques ou sources de données, et de repérer des données aberrantes en comparant chaque donnée individuelle avec le reste de l'échantillon. Elle peut déboucher sur la création de nouvelles règles intégrées aux deux phases précédentes. La qualité des données est évaluée à la fin de cette phase, et celles-ci sont resoumises au PPSD si la qualité n'est pas suffisante.

5.2. Non-réponse

Un défaut commun aux données d'enquêtes est l'absence de réponse à certaines questions par certaines unités enquêtées. On parle de non-réponse partielle lorsqu'un questionnaire est en partie rempli mais ne l'est pas complètement. Il est possible de laisser ces observations en l'état, mais

 ISSN: 2102-6238

cela conduira à produire des tableaux de résultats avec des catégories “réponse manquante”. Typiquement le lecteur, déjà relativement averti, fera alors mentalement une règle de trois pour répartir cette catégorie manquante dans les catégories renseignées. On comprend donc que cela revient à laisser au lecteur le choix d'un modèle d'imputation ou de repondération pour arriver à des résultats sans valeurs manquantes. On préfère en général, lorsqu'on en a les moyens, prendre en charge cette modélisation à l'OFS et traiter les observations avec non-réponse partielle. Deux sortes de traitement sont possibles :

1. imputer les valeurs manquantes, c'est à dire prédire les réponses qui n'ont pas été observées. Il s'agit d'un processus complexe, avec une part d'arbitraire, qui demande une grande expertise, et dont l'influence sur les résultats et leur précision est difficile à évaluer. Cette approche, discutée dans la section 5.4, est cependant souvent indispensable pour pouvoir exploiter les données d'enquête, car le traitement alternatif (voir ci-dessous) conduirait à perdre une trop grande part de l'échantillon.
2. éliminer les questionnaires présentant de la non-réponse partielle, et pondérer le reste de l'échantillon en conséquence. Cette approche est généralement utilisée pour les questionnaires qui présentent beaucoup de non-réponse et peu d'information utile, ou bien lorsqu'une imputation semble irréalisable. Cela a par exemple été effectué massivement dans le cadre de l'enquête structurelle lors de l'exploitation des questionnaires ménages, et du calcul d'une pondération d'extrapolation pour les ménages.

   En effet, les ménages observés à l'enquête structurelle sont tendanciellement plus petits que ceux attendus d'après le fichier SRPH. Il y a en particulier un déficit assez important de cohabitants âgés entre 20 et 40 ans. Cela peut être rapproché de la pratique courante pour les jeunes adultes et étudiants de rester enregistrés au domicile des parents plutôt que de s'inscrire dans leur commune de résidence effective. Or, si notre connaissance des ménages n'est pas parfaite, on estime qu'elle est tout de même de très bonne qualité. En particulier, le nombre de ménages dans le cadre d'échantillonnage SRPH est très proche de toutes les estimations qui avaient pu en être faites jusque-là. On pense donc que l'on est confronté à un problème de non réponse. La non-réponse (totale) d'un individu membre du ménage est assimilable à une non-réponse partielle pour l'unité ménage concernée, dans la mesure où il manque toutes les informations de l'individu non-répondant sur le questionnaire ménage. Cette non-réponse peut être liée à un désir de ne pas répondre ou à une mauvaise compréhension de la notion de ménage décrite sur le questionnaire.

   Procéder à des imputations pour combler ces non-réponses (que l'on ne sait de plus pas identifier avec certitude puisque le fichier SRPH n'est pas absolument juste) semble hors de portée. Il faudrait en effet imputer toutes les relations familiales en respectant leur cohérence. On a donc choisi de traiter cette non-réponse par élimination et repondération. Tous les ménages dont la composition déclarée est différente de la composition relevée dans le fichier SRPH ont été écartés de l'échantillon de ménages. Ils représentent environ 20% de l'échantillon. De fait tous les ménages dont la composition a réellement changé entre la constitution du cadre d'échantillonnage et l'enquête sont exclus. Toutefois, le nombre de ces ménages qui ont évolué est très faible devant le nombre total de ménages écartés.

Que ce soit pour procéder par imputation ou par élimination, il est nécessaire d'avoir une assez bonne idée d'un mécanisme qui pourrait expliquer la non-réponse survenue. Pour ce faire, on

 ISSN: 2102-6238

considère la variable indicatrice de réponse, résultat du contrôle de l'existence d'une réponse : $r_{kj} = 1$, si la variable j de l'observation k est renseignée, $r_{kj} = 0$ sinon. Le taux de réponse pour une variable j est alors la moyenne $\bar{r}_j = n^{-1} \sum_{k \in s} r_{kj}$, où n est la taille de l'échantillon s. Le cas échéant, il faut distinguer les valeurs qui sont structurellement manquantes, par exemple parce qu'une unité est dispensée de réponse à une partie du questionnaire, des valeurs qui auraient dû être observées mais ne sont pas renseignées. Le vecteur colonne $\mathbf{r}_j = (r_{1j}, r_{2j}, ..., r_{nj})'$ est l'indicatrice de réponse d'une variable j. La matrice des r_{kj} est générée par un processus de réponse inconnu. La modélisation de ce processus de réponse détermine les méthodes d'imputations et de repondérations utilisées.

5.3. Valeurs erronées et valeurs aberrantes

Un autre défaut fréquent des données d'enquête est la présence de valeurs manifestement fausses car en dehors des domaines des variables, possiblement fausses, incohérentes avec d'autres valeurs relevées, ou bien simplement atypiques car très éloignées du reste des valeurs observées. Ces observations peuvent provenir d'erreurs systémiques, par exemple d'erreurs de formulation dans le questionnaire, d'erreurs de numérisation, d'erreurs de programmation. Elles peuvent être liées à une incompréhension sur l'unité des valeurs demandées et plus généralement à une incompréhension du questionnaire. Elles peuvent ne répondre à aucun mécanisme évident. Enfin, elles peuvent être tout à fait correctes lorsqu'il s'agit de valeurs atypiques, mais leur présence implique une forte variabilité des estimateurs pour les variables concernées.

La détection de ces valeurs est effectuée au moyen de règles de contrôle. Fellegi and Holt (1976) ont montré que celles-ci peuvent être décomposées en règles simples implémentables dans un programme qui permet de tester les données. Ces règles peuvent être de diverses natures :

1. il y a les règles absolues qui permettent de repérer une valeur anormale sans considération pour les autres variables de la même unité enquêtée ou des autres unités. Par exemple, une année de naissance trop éloignée, un loyer négatif, peuvent-être détectés de cette manière.
2. Il y a aussi les règles de cohérence internes au questionnaire, par exemple les règles qui vérifient que la somme des composants déclarés d'un total déclaré est bien égale à ce total, ou bien les implications logiques entre variables déclarées. Lorsque des incohérences impliquent plusieurs variables sans que l'on sache déterminer lesquelles sont correctes et lesquelles sont erronées, le principe de parcimonie veut, en accord avec Fellegi and Holt (1976), que l'on essaye de traiter le plus petit nombre de variables pour rendre les observations cohérentes.
3. Enfin, il y a les règles qui permettent de détecter les valeurs influentes ou atypiques sans qu'il n'y ait nécessairement une incohérence ou une erreur. Celles-ci peuvent être une simple comparaison entre la valeur observée pour une unité et les valeurs observées sur le reste de l'échantillon. Un outil utile est alors le MAD (Median Absolute Deviation), défini comme la médiane des écarts absolus à la médiane de l'échantillon,

$$\mathrm{MAD}_y = \mathrm{med}_{k \in s}(|y_k - m_y|),$$

où $m_y = \mathrm{med}_{k \in s}(y_k)$ est la médiane des valeurs y_k observées. Cette mesure a l'avantage de ne pas être perturbée par un faible nombre de valeurs extrêmes. Les observations telles

 ISSN: 2102-6238

que l'écart $|y_k - m_y|$ excède une certaine fois MAD_y sont considérées comme extrêmes ou aberrantes. Hulliger (1999) propose de multiplier ces observations par un facteur $u_k < 1$ fonction de $\text{MAD}_y/|y_k - m_y|$. D'autres règles reposent implicitement sur une modélisation des relations entre les variables observées. On calcule fréquemment certains quotients y_{kj}/y_{km} entre variables relevées, avec l'idée que ces variables sont quasiment proportionnelles et que le quotient doit rester dans une plage de valeurs restreintes. Dans la cas de panels ou d'enquêtes répétées, l'évolution relative y_t/y_{t-1}, où y_t et y_{t-1} sont les valeurs pour l'enquête courante et la précédente est aussi étudiée (Hidiroglou and Berthelot, 1986). Plus généralement, on peut régresser certaines variables observées sur d'autres variables de l'enquête, de préférence de manière robuste, puis comparer l'écart entre la valeur observée et la valeur prédite à une mesure de dispersion elle aussi, si possible, robuste.

Dans les enquêtes de l'OFS auprès de la population et des ménages, les deux premiers types de règles de contrôle sont toujours employés. Le troisième type de règle peut aussi être utilisé lorsque le questionnaire porte sur les salaires, loyers et dépenses. Le non-respect d'une ou plusieurs règles conduit à l'élimination des données de l'unité concernée dans les cas les plus sévères, et généralement à l'imputation d'une valeur admissible dans les autres cas.

5.4. Imputations

L'OFS utilise plusieurs stratégies pour imputer de nouvelles valeurs à la place des valeurs manquantes, et de certaines valeurs aberrantes. En général, plusieurs stratégies sont utilisées pour chaque enquête. Les méthodes les plus complexes ne sont, par manque de temps ou d'expertise, pas systématiquement utilisées. De plus, la plupart du temps les imputations ne peuvent pas tenir compte de toutes les règles de contrôle. Un contrôle après imputation doit donc être mené. La méthode de Fellegi and Holt (1976) permet de tenir compte des règles de contrôle pour des variables qualitatives, mais cela nécessite qu'elles soient définies de manière stricte et complète, ce qui n'est pas toujours le cas.

Les stratégies d'imputation utilisées comprennent :

1. les imputations interactives, dont la reprise de résultats de rappels téléphoniques et les traitements manuels assistés par ordinateur. Ces opérations nécessitent une intervention humaine pour chaque observation. Leur volume est donc très limité. De plus, les imputations manuelles requièrent une grande expertise et ont le défaut de ne pas être reproductibles.
2. Les imputations automatiques basées sur des règles. Ce sont les résultats de déduction logique des valeurs d'imputations à l'aide de sources internes à l'enquête, comme les valeurs relevées d'autres variables, ou de sources externes, par exemple un registre qui serait apparié à l'échantillon. Ces imputations déterministes peuvent être implémentées de façon automatique. On leur accorde souvent un grand niveau de confiance, mais il faut prendre garde au fait qu'il y a parfois plusieurs valeurs imputées possibles, par exemple lorsque l'on a plusieurs sources d'information externes. Et dans les cas de reprise d'information externe, on prend le risque de dégrader les corrélations entre variables d'enquête.
3. Les imputations assistées par un modèle. On peut citer l'imputation par la moyenne, par la moyenne dans un groupe, par le quotient, par une prédiction de régression. Ces imputations reposent sur une modélisation des relations entre les variables observées et les variables

 ISSN: 2102-6238

inobservées. La modélisation est guidée par le type de non-réponse supposé. On peut distinguer trois classes de processus de réponse (Little and Rubin, 2002), les processus MCAR (Missing Completely at Random), MAR (Missing At Random) et NMAR (Not Missing At Random). La distinction entre ces trois classes réside dans la relation entre le processus de réponse et les valeurs observées et inobservées des variables y_j. Si la loi du processus ne dépend pas du tout de ces valeurs, on dit que le processus est MCAR. Si elle dépend de valeurs observées, mais pas des valeurs inobservées, on dit que le processus est MAR. Enfin, si elle dépend des valeurs observées et inobservées, on dit que le processus est NMAR. Ce dernier cas est typique des variables portant sur les revenus. En effet, la propension a répondre peut dépendre du montant du revenu, sans que cette dépendance puisse être contrôlée par l'utilisation des autres variables à disposition. Les paramètres du modèle sont estimés sur le sous-échantillon de répondants pour lesquels les questionnaires sont complets. La procédure d'imputation a un impact sur la distribution produite des variables d'enquête. Un problème classique de l'imputation par la moyenne ou par la prédiction est qu'elle tend à réduire la dispersion de la variable d'enquête. Ceci est parfois compensé par l'ajout aux valeurs prédites d'une erreur générée aléatoirement. La cohérence entre toutes les variables d'enquête est difficile à maintenir.

4. Les imputations par donneur. Les valeurs déclarées par une ou plusieurs autres unités enquêtées sont utilisées pour l'imputation. Ces donneurs sont choisis à l'aide d'une distance définie entre les répondants. On peut citer par exemple la méthode des plus proches voisins (Sande, 1981). L'avantage de ce genre de méthode est que, grâce à un même donneur, il est possible de remplacer simultanément plusieurs valeurs manquantes ou erronées. On préserve ainsi la cohérence entre les valeurs imputées. La difficulté réside dans le choix de la fonction de distance entre les observations.

Elles sont en général utilisées à l'OFS dans cet ordre chronologique lors du traitement des données d'enquête.

L'expérience d'imputation la plus complexe à l'OFS a été menée dans le cadre de l'enquête SILC. On a essayé d'y utiliser une méthode d'imputation multiple assistée par des modèles de régression. Plusieurs variables ont été imputées à l'aide du programme IVEware (Raghunathan et al., 2001). IVEware permet d'effectuer des imputations individuelles ou multiples au moyen d'une méthode de régression séquentielle. Le type de régression utilisé est adapté à la nature des variables à imputer : quantitative, dichotomique, catégorielle, ou bien variable de comptage. Les variables sont traitées par proportion croissante de valeurs manquantes. Soit $\mathbf{X}$ la matrice regroupant le sous-ensemble des variables explicatives sans valeurs manquantes. Soient aussi $y_1, y_2, ..., y_p$ les p variable à imputer, où y_1 présente le moins de valeurs à imputer et y_p en présente le plus. Ces variables peuvent être de nature différentes. La fonction de densité commune de $y_1, y_2, ..., y_p$ sachant $\mathbf{X}$ peut être factorisée comme suit (Raghunathan et al., 2001) :

$$f(Y_1, Y_2, ..., Y_p|\mathbf{X}) = f_1(Y_1|\mathbf{X}) f_2(Y_2|\mathbf{X}, Y_1) ... f_p(Y_p|\mathbf{X}, Y_1, Y_2, ..., Y_{p-1})$$

où les f_j, $j = 1, 2, ..., p$ sont des fonctions de densité conditionnelles. On modélise chaque densité conditionnelle à l'aide d'un modèle de régression paramétrique approprié. Les densités conditionnelles sont alors estimées séquentiellement.

Le premier cycle d'une imputation est effectué comme suit : la densité conditionnelle f_1 est estimée à partir des observations pour lesquelles on dispose des valeurs de y_1. Des valeurs

 ISSN: 2102-6238

imputées sont alors générées selon cette loi conditionnelle pour compléter la variable y_1. La loi conditionnelle f_2 est estimée en utilisant à partir des observations pour lesquelles on dispose des valeurs de y_2, et en utilisant les valeurs imputées pour y_1 à l'étape précédente. Des valeurs imputées pour y_2 sont alors générées. Le processus est poursuivi jusqu'à disposer de valeurs imputées pour toutes les variables.

Le deuxième cycle d'une imputation est réalisé en estimant la densité de la variable y_1 conditionnellement à $\mathbf{X}$ et aux $p-1$ autres variables. De nouvelles valeurs imputées sont générées selon cette densité estimée et remplacent celles générées au premier cycle. Il est fait de même successivement pour les $p-1$ autres variables à imputer. Le nombre de tels cycles effectués peut être choisi, et on arrête le processus lorsque les résultats ne changent plus vraiment.

L'ensemble de la procédure, première imputation et cycles successifs compris, peut être répété m fois pour obtenir m jeux de valeurs imputées différents pour chacune des p variables. L'existence de plusieurs jeux de valeurs imputées permet alors d'avoir une idée de la distribution des valeurs imputées, et en particulier d'estimer simplement la variance due aux imputations. C'est là, avec la simplicité d'utilisation, l'un des avantages certains du logiciel IVEware. D'autres aspects peuvent être vus comme des désavantages : le tout fonctionne comme une boite noire, avec des cycles, des répétitions, on ne saurait calculer explicitement une densité même en connaissant les densités réelles des observations en entrée. Le code source du logiciel n'est pas disponible. On ne peut spécifier de pondérations. De plus, les régressions employées sont sensibles aux valeurs extrêmes, et mal adaptées aux variables monétaires. Enfin, ce qui est pour l'estimation de variance un avantage, la production de plusieurs valeurs imputées, est pour l'estimation directe source d'insatisfaction. En effet, il a été constaté que la variabilité des valeurs imputées était trop grande pour maintenir une cohérence acceptable avec d'autres informations du fichier de données si l'on choisit au hasard une des valeurs imputées. Actuellement, la valeur imputée correspond à la moyenne des valeurs obtenues sur 50 imputations. Un système d'imputation qui ne présente pas ces inconvénients est à l'étude (Graf and Tillé, 2012).

6. Estimation de variance

La variance des estimateurs utilisés dépend du plan de sondage, de la non-réponse et de son traitement, des éventuels partages ou combinaisons de poids, et des calages et imputations effectués. En toute rigueur, une estimation de cette variance devrait aussi prendre en compte tous ces éléments. En premier lieu, il faudrait pouvoir disposer d'une méthode de calcul spécifique à chaque plan de sondage. Or aucun outil n'est disponible à l'heure actuelle qui soit capable de répondre à ces besoins. Il est même douteux qu'il puisse en exister un tant les traitements peuvent être variés, et parfois arbitraires. La technique de Bootstrap (Efron, 1979; Davison and Hinkley, 1997) peut sembler être une manière de contourner ces difficultés. Mais elle demande également des développements spécifiques à chaque situation (Chauvet, 2007; Antal and Tillé, 2011).

6.1. Pour le relevé structurel et les autres enquêtes sélectionnées de manière coordonnée dans le fichier SRPH

Les méthodes d'estimation de variance s'appuient le plus possible sur les capacités offertes par les procédures spécifiques aux données d'enquêtes du logiciel SAS utilisé à l'OFS. Ces

 ISSN: 2102-6238

procédures n'offrent pas la souplesse nécessaire pour prendre en compte tous les aspects des plans d'échantillonnage et des méthodes de pondération effectivement employées. En tout état de cause, ces plans d'échantillonnage excluent la possibilité de sélectionner simultanément deux personnes dont on sait qu'elles sont dans le même ménage. Il n'est donc pas possible d'estimer sans biais la variance des estimateurs de totaux (Särndal et al., 1992).

Toutefois, pour l'ensemble des enquêtes appartenant au nouveau système de recensement, les plans de sondages sont relativement proches de plans simples stratifiés par cantons, conditionnellement à la taille des échantillons sélectionnés dans les cantons. En modélisant le processus de non-réponse par un processus bernoullien de sélection des répondants dans les cantons, et en conditionnant par le nombre de répondants, on obtient encore un plan de sondage pour les répondants proche d'un plan simple stratifié. Cette approximation justifie l'utilisation d'estimateurs de variance classiques pour les plans stratifiés.

Les effets du calage réalisé pour pondérer les données sont pris en compte en utilisant une technique de linéarisation (Deville, 1999). En pratique, une macro procédure SAS a été programmée pour permettre un calcul facile d'estimateurs de variance pour des estimateurs de totaux et de quotients. Cette procédure calcule dans un premier temps les résidus de régression des variables d'intérêts sur les variables de calages de l'enquête, puis elle fait appel aux procédures standard de SAS pour obtenir les estimations de variances. Les sorties sont ensuite mises en forme pour ressembler aux sorties standard des procédures SAS.

6.2. Pour l'enquête SILC

Lors du lancement du projet SILC en Suisse, trois options pour estimer les variances des estimateurs utilisés ont été envisagées : la technique de Bootstrap, le jackknife (Quenouille, 1949) et les techniques de linéarisation (Deville, 1999). Des poids bootstrap ont été produits avec l'aide de Statistique Canada (Cauchon, 2006) utilisés avec la macro BOOTVAR (Statcan, 2005). L'OFS dispose de macros SAS utilisant la méthode jackknife pour estimer la variance de certaines statistiques (Canty and Davison, 1998).

Des comparaisons entre les variances obtenues sur la base de 1'000 jeux de poids bootstrap, par jackknife, ainsi que par linéarisation ont été menées pour les enquêtes pilotes de SILC ainsi que pour des vagues spécifiques du Panel Suisse de Ménages. Les estimations de variance obtenues pour les statistiques calculées, totaux, ratios, différences entre ratios, percentiles, régressions linéaires et logistique, tests du Chi-deux, sont assez proches les unes des autres pour les trois méthodes considérées. Ces essais nous ont conduit à retenir en production la méthode de linéarisation couplée à l'utilisation des procédures de SAS spécifiques pour le traitement des données d'enquête.

Récemment, une procédure qui tient compte de toutes les étapes de la construction de poids pour l'enquête SILC a été détaillée et programmée (Massiani, 2013). Il y est tenu compte du plan d'échantillonnage, de la modélisation de la non-réponse, du partage des poids, de la combinaison des panels et du calage, par la technique de linéarisation. Des macros SAS adaptées à SILC permettant des estimations de variances et tenant compte des spécificités de l'enquête ont été produites.

Des statistiques complexes, non linéaires, sont produites à partir de l'enquête SILC. Il s'agit des indices de pauvreté et d'exclusion sociale. Plusieurs grands projets européens se sont concentrés

 ISSN: 2102-6238

sur l'étude des indices de pauvretés estimés à partir d'enquêtes par échantillonnage. On peut citer Dacseis (E.U., 2004), KEI E.U. (2008), Sample (E.U., 2011b) et Ameli (E.U., 2011a). Eurostat a également produit les rapports Eurostat (2005a, 2004c,a) et Eurostat (2013). Les programmes de linéarisation développés par Osier (2009) et basées sur Deville (1999) et Demnati and Rao (2004) sont utilisées dans Massiani (2013) pour estimer les variances de ces estimateurs.

7. Anticipations

7.1. Sur le traitement de la non-réponse NMAR : le calage généralisé

Le calage généralisé de Deville (2002), voir aussi Le Guennec and Sautory (2002); Sautory (2003); Kott (2006), autorise l'utilisation de variables connues uniquement sur l'échantillon des répondants, en plus des variables auxiliaires disponibles pour le calage classique. Tout comme ce dernier, il permet de calculer un jeu de poids w_k, proche des poids initiaux, qui satisfait des contraintes de calage

$$\sum_{k \in r} w_k x_k = X,$$

où les totaux X sur la population des variables auxiliaires x_k sont connus. La différence entre le calage généralisé et le calage classique réside dans la manière d'obtenir ce nouveau jeu de poids. Les fonctions utilisées dans le calage généralisé dépendent des variables x_k et de variables z_k que l'on peut librement choisir, y compris parmi les variables qui ne sont connues que sur l'échantillon de répondants $r \subseteq s$. Ces variables z_k sont appelées instruments de calage, car cette méthode présente une analogie certaine avec la régression instrumentale (Theil, 1953). En particulier, la variance de l'estimateur pondéré du total obtenu peut être estimée par une technique de linéarisation. Elle demande de calculer les résidus de la régression instrumentale des variables d'intérêt y_k sur les variables de calage x_k avec les instruments z_k.

Cette méthode permet de traiter un problème de non-réponse NMAR lorsque certaines variables connues uniquement sur l'échantillon observé sont déterminantes de la propension à répondre (Deville, 2002; Graf and Tillé, 2012; Osier, 2013). Ces variables doivent alors être utilisées comme instruments de calage. L'utilité de la méthode du calage généralisé est conditionnée par un choix pertinent des variables z_k. De même que les techniques de sélection de modèles pour la régression linéaire peuvent être utilisées pour choisir les variables de calage dans le cas d'un calage classique, les résultats théoriques sur la régression instrumentale devraient pouvoir être transposés au calage généralisé. Nous étudions la possibilité de mettre au point l'analogue de tests de Durbin-Wu-Hausman (voir Durbin, 1954; Wu, 1973; Hausman, 1978; Ruud, 1984) dans le cadre d'enquêtes par échantillonnage afin de disposer d'un outil pour identifier les bonnes variables instrumentales.

7.2. Sur l'estimation de la variance due à l'imputation

Comme mentionné dans la section 5.4, l'OFS utilise actuellement le logiciel IVEware pour procéder à des imputations multiples dans l'enquête SILC dans les cas de non-réponse partielle. Ce logiciel utilise des modèles de régression mal adaptés aux variables de revenu. Nous développons une méthode d'imputation paramétrique reposant sur l'ajustement d'une loi Bêta Généralisée de

 ISSN: 2102-6238

Seconde Espèce GB2 (McDonald, 1984; E.U., 2011a). En effet, plusieurs études empiriques (voir par exemple Kleiber and Kotz, 2003; Jenkins, 2007; Dastrup et al., 2007; Sepanski and Kong, 2008; Jones et al., 2011; McDonald et al., 2013) montrent qu'une distribution GB2 s'ajuste bien à de telles données et qu'elle est souvent plus adaptée que d'autres distributions à quatre paramètres. Cet ajustement est réalisé en s'appuyant sur des poids obtenus par calage généralisé qui corrigent au mieux la non-réponse NMAR dont est affectée la variable d'intérêt, voir Graf and Tillé (2012).

7.3. Sur la sélection des échantillons pour les enquêtes de l'OFS dans un avenir proche

Au cours de l'année 2014, les dernières enquêtes dont les échantillons sont tirés dans le cadre Castem, à savoir l'enquête SILC, l'Espa et l'EBM seront intégrées au système de coordination d'enquêtes et sélectionnées dans le fichier SRPH. Cela permettra de coordonner proprement ces enquêtes avec les enquêtes du système de recensement, et de ne plus avoir recours à la méthode "d'historisation". SILC et l'EBM sont des enquêtes qui touchent toutes les personnes des ménages sélectionnés. Or le système actuel ne permet pas de sélectionner directement et systématiquement toutes les personnes d'un même ménage dans les échantillons. Nous avons donc réfléchi à la meilleure manière d'organiser ces tirages.

Il est en outre apparu à l'usage que la coordination des sélections des personnes ne suffisait pas à répondre aux attentes des équipes de l'OFS qui commandent les échantillons pour leurs enquêtes. Celles-ci désirent en effet qu'il y ait une coordination au niveau des ménages, de sorte que deux personnes d'un même ménage ne puissent être sélectionnées de manière trop rapprochée.

Ces deux points, et l'entrée en vigueur de l'obligation légale de constitution des ménages dans les registres de population des cantons et communes nous à conduit à repenser toute notre procédure de tirage, et en même temps à proposer une méthode plus simple pour obtenir des échantillons avec une seule sélection par ménage. D'ici à fin 2013, l'OFS prévoit de constituer un cadre de sondage de ménages à partir du fichier SRPH et de sélectionner ses échantillons en deux phases. Une première phase sera un tirage coordonné de ménages dans le cadre de ménages, et, pour les enquêtes auprès des personnes, la deuxième phase sera la sélection d'un individu par ménage sélectionné en première phase. On obtiendra de cette manière un système qui répondra bien aux besoins de l'OFS quant à la répartition de la charge d'enquête sur les ménages et qui permet d'éviter les complications évoquées dans la section 3.3. En outre, les plans transversaux des enquêtes ne seront pas modifiés par ce changement. La coordination au niveau des personnes sera par contre dégradée, puisque rien n'est prévu pour éviter que la même personne d'un ménage de plusieurs personnes ne soit re-sélectionnée lorsque, après un certain temps, le ménage lui-même est à nouveau re-sélectionné.

La transition vers ce nouveau processus de tirage est actuellement à l'étude afin de permettre une assez bonne coordination négative entre les enquêtes tirées selon la nouvelle procédure et celles tirées selon l'ancienne procédure. Plus précisément, les numéros aléatoires permanents des ménages seront générés en fonction des sélections et probabilités d'inclusion de leurs membres aux enquêtes passées.

 ISSN: 2102-6238

8. Conclusions

Nous avons effectué un tour d'horizon des enquêtes par échantillonnage auprès des personnes et des ménages à l'OFS. La généralisation à toutes les enquêtes de la coordination des échantillons permettra, dans un avenir proche, de répartir au mieux la charge d'enquête sur les ménages et les personnes. Outre le confort supplémentaire que cela offre à la population, l'OFS espère que cette répartition équitable de la charge permettra d'améliorer les taux de réponse à ses enquêtes. Dans le domaine des pondérations, les quelques exemples traités montrent qu'il existe une diversité de situations et de besoins qui conduisent à des calculs et l'implémentation de méthodologies différents. La création du cadre d'échantillonnage SRPH et la généralisation de son utilisation pour toutes les enquêtes de l'OFS devraient permettre d'harmoniser davantage ces calculs. La simplification de ces calculs sera possible grâce à la connaissance précise de la population et de sa probabilité de sélection dans chaque vague d'enquête ou bloc de rotation pour les enquêtes de l'OFS. L'Office tente d'instaurer dans tous ses services l'utilisation de bonnes pratiques pour le traitement des données. L'imputation est un sujet complexe et le potentiel d'amélioration dans ce domaine est grand. Les progrès sont conditionnés par un investissement conséquent en temps et en moyens humains.

Par ailleurs, dans la mesure de ses capacités, l'OFS étudie et met en pratique les avancées récentes dans le domaine en constante évolution de la méthodologie d'enquêtes. Ainsi, les méthodes de calage "lune et étoile" et de calage généralisé ont déjà été mises à l'essai avec un certain succès. L'OFS participe également à des projets internationaux en collaboration avec d'autres instituts nationaux de statistique et universités. Il poursuit également sa collaboration étroite avec l'Institut de Statistique de l'Université de Neuchâtel. Ces collaborations couvrent des domaines variés comme l'estimation de variance, le traitement des données d'enquête, les méthodes d'estimation sur des petits domaines, la mesure des inégalités. L'OFS s'attache à mettre en pratique les recommandations formulées par les grands projets européens sur la méthodologie d'enquête. La qualité des données et la probité des méthodes employées sont au cœur des préoccupations de l'Office.

Cet article est partiellement soutenu par une convention de collaboration entre l'Office fédéral de la statistique et l'Université de Neuchâtel. Son contenu n'engage que les auteurs et en aucun cas l'Office fédéral de la statistique. Pour obtenir des informations sur les enquêtes de l'OFS, le lecteur peut consulter le site internet de l'office (http ://www.bfs.admin.ch/bfs/portal/fr/index.html). Les auteurs tiennent à remercier les arbitres, éditeurs invités et éditeurs qui ont, par leurs remarques pertinentes, permis d'améliorer la qualité de cet article.

Références

Antal, E. and Tillé, Y. (2011). A direct bootstrap method for complex sampling designs from a finite population. *Journal of the American Statistical Association*, 106 :534–543.

Brewer, K. R. W., Early, L. J., and Hanif, M. (1984). Poisson, modified Poisson and collocated sampling. *Journal of Statistical Planning and Inference*, 10 :15–30.

Brewer, K. R. W., Early, L. J., and Joyce, S. F. (1972). Selecting several samples from a single population. *Australian Journal of Statistics*, 3 :231–239.

Canty, A. J. and Davison, A. C. (1998). Variance estimation for two complex surveys in switzerland. document interne, Office Fédéral de la Statistique.

 ISSN: 2102-6238

Cauchon, C. (2006). Swiss household panel and silc. bootstrap samples and variances. Technical report, Statistics Canada, Ottawa.
Chauvet, G. (2007). *Méthodes de Bootstrap en Population Finie*. Thèse de doctorat, Université Rennes 2, Rennes, France.
Cotton, F. and Hesse, C. (1992a). Co-ordinated selection of stratified samples. In *Proceedings of Statistics Canada Symposium*.
Cotton, F. and Hesse, C. (1992b). Tirages coordonnés d'échantillons. Document de travail de la Direction des Statistiques Économiques E9206. Technical report, INSEE, Paris.
Dastrup, S. R., Hartshorn, R., and McDonald, J. B. (2007). The impact of taxes and transfer payments on the distribution of income : A parametric comparison. *Journal of Economic Inequality*, 5 :353–369.
Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and Their Application*. Cambridge University Press, Cambridge.
De Ree, S. J. M. (1983). A system of co-ordinated sampling to spread response burden of enterprises. In *44th Session of the ISI Madrid*, pages 673–676.
Demnati, A. and Rao, J. N. K. (2004). Estimateurs de variance par linéarisation pour des données d'enquête. *Techniques d'enquête*, 30(1) :17–27.
Deville, J.-C. (1999). Estimation de variance pour des statistiques et des estimateurs complexes : techniques de résidus et de linéarisation. *Techniques d'enquête*, 25(2) :219–230.
Deville, J.-C. (2002). La correction de la non-réponse par calage généralisé. In *Journées de Méthodologie Statistique*, Paris. INSEE.
Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87 :376–382.
Deville, J.-C., Särndal, C.-E., and Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88 :1013–1020.
Dufour, J., Gagnon, F., Morin, Y., Renaud, M., and Särndal, C.-E. (1998). Measuring the impact of alternative weighting schemes for longitudinal data. In *Survey Research Methods Section*, pages 552–557. American Statistical Association.
Durbin, J. (1954). Errors in variables. *Review of the International Statistical Institute*, 22 :23–32.
Efron, B. (1979). Bootstrap methods : Another look at the jackknife. *Annals of Statistics*, 7 :1–26.
Ernst, L. R. (1996). Maximizing the overlap of sample units for two designs with simultaneous selection. *Journal of Official Statistics*, 12 :33–45.
E.U. (2003). Project euredit : The development and evaluation of new methods for editing and imputation.
E.U. (2004). Project dacseis : Data quality in complex surveys within the new european information society. FP5.
E.U. (2007). Project edimbus : Recommended practices for editing and imputation in cross-sectional business surveys. Luzi, O. et al.
E.U. (2008). Project kei : Knowledge economy indicators. FP6.
E.U. (2011a). Project ameli : Advanced methodology for european laeken indicators. FP7.
E.U. (2011b). Project sample : Small area methods for poverty and living conditions estimates. FP7.
Eurostat (2004a). Common cross-sectional eu indicators based on eu-silc ; the gender pay gap. Working papers and studies, Office for Official Publications of the European Communities, Luxembourg. EU-SILC 131-rev/04.
Eurostat (2004b). Cross sectional weighting : first year of each sub-sample. Working papers and studies, Office for Official Publications of the European Communities, Luxembourg. EU-SILC 134/04.
Eurostat (2004c). Theoretical study of the gini index. Working papers and studies, Office for Official Publications of the European Communities, Luxembourg. EU-SILC 131-A/04.
Eurostat (2005a). The continuity of indicators during the transition between ECHP and EU-SILC. Working papers and studies, Office for Official Publications of the European Communities, Luxembourg.
Eurostat (2005b). Cross-sectional weighting : from second year of the survey onwards. Working papers and studies, Office for Official Publications of the European Communities, Luxembourg. EU-SILC 157/05.
Eurostat (2013). Handbook on precision requirements and variance estimation for ess household surveys.
Fellegi, P. and Holt, D. (1976). A systematic approach to automatic edit and imputation. *Journal of the American Statistical Association*, 71 :17–35.
Feusi Widmer, R. (2004). L'enquête suisse sur la population active (espa). Technical report, Office Fédéral de la Statistique, Neuchâtel. Statistique de la Suisse.
Graf, E. (2006). Silc, sample sizes requirements for switzerland according to the rules defined by eurostat for the

 ISSN: 2102-6238

european union and rotational design. Internal document, Swiss Federal Statistical Office, Neuchâtel.
Graf, E. (2008). Pondérations du silc pilote silc_i vague 2, silc_ii vague 1, silc_i et silc_ii combinés. Rapport de méthodes, Office Fédéral de la Statistique, Neuchâtel.
Graf, E. (2009). Pondérations du panel suisse de ménages. Rapport de méthodes, Office Fédéral de la Statistique, Neuchâtel, Suisse.
Graf, E. (2010). Enquête suisse sur la santé 2007. plan d'échantillonnage pondérations et analyses pondérées des données. Rapport de méthodes, Office Fédéral de la Statistique, Neuchâtel.
Graf, E. and Kilchmann, D. (2010). La formation spécialisée en statistique : Données manquantes et données éronnées. document interne, Office Fédéral de la Statistique. Cours B4.2.
Graf, E. and Tillé, Y. (2012). Imputations de données de revenu à l'aide de calage généralisé et de lois GB2 : illustration sur les données silc suisses 2009. In *Colloque Francophone sur les Sondages*, Rennes.
Hájek, J. (1981). *Sampling from a Finite Population*. Marcel Dekker, New York.
Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica*, 46(6) :1251–1271.
Hidiroglou, M. A. and Berthelot, J. M. (1986). Contrôle statistique et imputation dans les enquêtes-entreprises périodiques. *Techniques d'enquête*, 12(1) :79–89.
Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47 :663–685.
Hulliger, B. (1999). Simple and robust estimators for sampling. In *Survey Research Methods Section*, pages 54—-63. American Statistical Association.
Jenkins, S. P. (2007). Inequality and the gb2 income distribution. *Discussion Paper IZA No. 2831*.
Jones, A., Lomas, J., and Rice, N. (2011). Applying beta-type size distributions to healthcare cost regressions. Technical report, University of York. Health, Econometrics and Data Group.
Kalton, G. and Kasprzyk, D. (1986). Le traitement des données d'enquête manquantes. *Techniques d'enquête*, 12(1) :1–17.
Kass, G. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, 29 :119–127.
Keyfitz, N. (1951). Sampling with probabilities proportional to size : adjustmentfor changes in the probabilities. *Journal of the American Statistical Association*, 46 :105–109.
Kish, L. and Scott, A. (1971). Retaining units after changing strata and probabilities. *Journal of the American Statistical Association*, 66 :461–470.
Kleiber, C. and Kotz, S. (2003). *Statistical Size Distributions in Economics and Actuarial Sciences*. Wiley, New York.
Kott, P. S. (2006). Utilisation de la pondération par calage pour la correction de la non-réponse et des erreurs de couverture. *Techniques d'enquête*, 32(2) :149–160.
Kröger, H., Särndal, C.-E., and Teikari, I. (1999). L'échantillonnage mixte de poisson : une famille de plans permettant une sélection coordonnée à l'aide de nombres aléatoires permanents. *Techniques d'enquête*, 25(1) :3–12.
LarRoche, S. (2007). Pondérations longitudinale et transversale de l'enquête sur la dynamique du travail et du revenu. année de référence 2003. Technical report, Statistique Canada, Ottawa.
Lavallée, P. (2002). *Le sondage indirect ou la méthode généralisée du partage des poids*. Ellipses, Paris.
Le Guennec, J. and Sautory, O. (2002). Calmar 2 : Une nouvelle version de la macro calmar de redressement d'échantillon par calage. In *Journées de Méthodologie Statistique*, Paris. INSEE.
Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis With Missing Data*. Wiley, New York, 2nd edition.
Lévesque, I. and Franklin, S. (2000). Pondérations longitudinale et transversale de l'enquête sur la dynamique du travail et du revenu. année de référence 1997. Technical report, Statistique Canada, Ottawa.
Massiani, A. (2013). Estimation de la variance d'indicateurs transversaux pour l'enquête SILC en Suisse. *Techniques d'enquête*, 39(1) :139–167.
McDonald, J. B. (1984). Some generalized functions for the size distribution of income. *Econometrica*, 52 :647–663.
McDonald, J. B., Sorensen, J., and Turley, P. A. (2013). Skewness and kurtosis properties of income distribution models. *Review of Income and Wealth*, 2 :360–374.
Merkouris, T. (2001). Estimation transversale dans le cas des enquêtes auprès des ménages à panels multiples. *Techniques d'enquête*, 27(2) :189–200.
Nakache, J.-P. and Confais, J. (2003). *Statistique explicative appliquée : analyse discriminante, modèle logistique, segmentation par arbre*. Editions Technip, Paris.
OFS (2012). L'enquête suisse sur la population active dès 2010. concepts - bases méthodologiques - considérations pratiques. Technical Report do-f-03-sake-2012-03, Office Fédéral de la Statistique, Neuchâtel. Encyclopédie

 ISSN: 2102-6238

statistique de la Suisse.
Ohlsson, E. (1995). Coordination of samples using permanent random numbers. In Cox, B. G., Binder, D. A., Chinnappa, B. N., Christianson, A., Colledge, M. J., and Kott, P. S., editors, *Business Survey Methods*, pages 153–169, New York. Wiley.
Osier, G. (2009). Variance estimation for complex indicators of poverty and inequality using linearization techniques. *Survey Research Methods*, 3 :167–195.
Osier, G. (2013). Dealing with non-ignorable non-response using generalised calibration : Simulation study based on the luxemburgish household budget survey. In *Economie et Statistiques : Working papers du STATEC*, volume 65. STATEC, Luxembourg.
Patterson, H. D. (1950). Sampling on successive occasions with partial replacement of units. *Journal of the Royal Statistical Society*, B12 :241–255.
Qualité, L. (2009). *Unequal probability sampling and repeated surveys*. Thèse de doctorat, Université de Neuchâtel, Neuchâtel, Suisse.
Quenouille, M. H. (1949). Approximation tests of correlation in time series. *Journal of the Royal Statistical Society*, B11 :18–84.
Raghunathan, T. E., Lepkowski, J. M., van Hoewyk, J., and Solenberger, P. (2001). Une technique multidimensionnelle d'imputation multiple des valeurs manquantes à l'aide d'une séquence de modèles de régression. *Techniques d'enquête*, 27(1) :91–103.
Rakotomalala, R. (2005). Arbres de décision. *Revue MODULAD*, 33 :163–187.
Renfer, J.-P. (2009). Etude de la précision des estimateurs de l'enquête continue suisse sur la population active selon le choix du schéma de rotation des échantillons. In *Journées de Méthodologie Statistique*, Paris. INSEE.
Rivière, P. (1998). Description of the chosen method : Deliverable 2 of supcom 1996 project (part "co-ordination of samples"). Report, EUROSTAT.
Rivière, P. (1999). Coordination of samples : the microstrata methodology. In *13th International Roundtable on Business Survey Frames*, Paris. INSEE.
Rivière, P. (2001a). Coordinating samples using the microstrata methodology. In *Proceedings of Statistics Canada Symposium*.
Rivière, P. (2001b). Random permutations of random vectors as a way of co-ordinating samples. Technical report, University of Southampton, Southampton.
Rosén, B. (1997a). Asymptotic theory for order sampling. *Journal of Statistical Planning and Inference*, 62 :135–158.
Rosén, B. (1997b). On sampling with probability proportional to size. *Journal of Statistical Planning and Inference*, 62 :159–191.
Ruud, P. A. (1984). Tests of specification in econometrics. *Econometric Reviews*, 3(2) :211–242.
Sande, I. G. (1981). Imputations in surveys : Coping with reality. *Survey Methodology*, 7(1) :21–43.
Särndal, C.-E. and Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. Wiley, New York.
Särndal, C.-E., Swensson, B., and Wretman, J. H. (1992). *Model Assisted Survey Sampling*. Springer, New York.
Sautory, O. (2003). Calmar 2 : A new version of the calmar calibration adjustment program. In *Proceedings of Statistics Canada Symposium*.
Sepanski, J. H. and Kong, J. (2008). A family of generalized beta distributions for income. *Advances and Applications in Statistics*, 10 :75–84.
Statcan (2005). *BOOTVAR, Guide de l'usager*. Statistique Canada. Bootvar 3.1, version SAS.
Strand, M. M. (1979). Estimation of a population total under a "Bernoulli sampling" procedure. *The American Statistician*, 33 :81–84.
Tambay, J. L., Schiopu-Kratina, I., Mayda, J., Stukel, D., and Nadon, S. (1998). Traitement de la non-réponse du cycle de deux enquêtes sur la santé de la population. *Techniques d'enquête*, 24(2) :159–169.
Theil, H. (1953). Repeated least square applied to complete equation systems. Technical report, Central Planning Bureau, The Hague.
Tillé, Y. (2006). *Sampling Algorithms*. Springer, New York.
Van Huis, L. T., Koeijers, C. A. J., and de Ree, S. J. M. (1994a). EDS, sampling system for the central business register at statistics netherlands. Technical report, Statistics Netherlands, The Hague.
Van Huis, L. T., Koeijers, C. A. J., and de Ree S. J. M. (1994b). Response burden and co-ordinated sampling for economic surveys. Technical report, Statistics Netherlands, The Hague.
Wu, D.-M. (1973). Alternative tests of independence between stochastic regressors and disturbances : Finite sample results. *Econometrica*, 42(2) :529–546.

 ISSN: 2102-6238