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

Article

Estimation de la couverture du recensement de la population de l'an 2000 en Suisse : méthodes et résultats

par Anne Renaud

Décembre 2007

Statistique Canada Statistics Canada

Canada

Techniques d'enquête, décembre 2007
Vol. 33, n° 2, pp. 221-232
Statistique Canada, n° 12-001-X au catalogue

Estimation de la couverture du recensement de la population de l'an 2000 en Suisse : méthodes et résultats

Anne Renaud [1]

Résumé

Les défauts de couverture sont estimés et analysés pour le recensement de la population de l'an 2000 en Suisse. La composante de sous-couverture est estimée sur la base d'un échantillon indépendant du recensement et d'un appariement avec le recensement. La composante de sur-couverture est estimée sur la base d'un échantillon tiré dans la liste du recensement et d'un appariement avec le reste du recensement. Les composantes de sur- et sous-couverture sont ensuite combinées pour obtenir une estimation de la couverture nette résultante. Cette estimation est basée sur un modèle de capture-recapture, nommé système dual, combiné avec un modèle synthétique. Les estimateurs sont calculés pour la population entière et différents sous-groupes, avec une variance estimée par un jackknife stratifié. Les analyses de couverture sont complétées par une étude des appariements entre l'échantillon indépendant et le recensement afin de déterminer les erreurs de mesure et de localisation potentielles dans les données du recensement.

Mots clés : Recensement; erreurs de couverture; système dual; plan d'échantillonnage à plusieurs degrés; erreurs de mesure.

1. Introduction

Dans tout recensement, certaines personnes ne sont pas recensées alors qu'elles devraient l'être et d'autres sont comptées deux fois ou n'auraient pas dû être recensées. Il y a donc de la sous-couverture et de la sur-couverture, dont le bilan est très souvent une sous-couverture nette. La sous-couverture nette est par exemple estimée à 1,6 % aux États-Unis en 1990 (Hogan 1993), 2,2 % au Royaume-Uni en 1991 (Brown, Diamond, Chambers, Buckner et Teague 1999) et 3 % au Canada en 2001 (Statistique Canada 2004). Aux États-Unis en 2000, la couverture nette correspond par contre à une sur-couverture de 0,5 % (Hogan 2003). Les défauts de couverture peuvent fortement varier entre sous-groupes de la population. Aux États-Unis en 2000, on note par exemple que les Noirs avaient une sous-couverture nette de 1,8 % alors que les Blancs avaient une sur-couverture de 1,1 %. De plus, les valeurs varient souvent entre classes d'âge et régions par exemple. Ces défauts de couverture, et les autres erreurs telles les erreurs de mesure, conduisent à une image biaisée de la population. Elles sont donc étudiées afin d'avoir une information sur la qualité des données disponibles et trouver des pistes pour améliorer les relevés censitaires auprès de la population.

Le recensement de la population de l'an 2000 en Suisse donne une image de la population au 5 décembre 2000. Pour la première fois, on estime les défauts de couverture d'un recensement suisse. La sous-couverture, la sur-couverture et la couverture nette résultante du recensement 2000 sont toutes trois analysées. La sous-couverture est estimée sur la base d'un échantillon de personnes S_P, indépendant du recensement, sur lequel on organise une enquête de couverture quelques mois après le recensement (relevé entre avril et mai 2001). Les données de cette enquête sont alors appariées avec les données du recensement afin de déterminer si la personne de S_P a été recensée ou pas. La sur-couverture est estimée sur la base d'un échantillon de personnes S_E tiré parmi les enregistrements du recensement. Une recherche de doubles et d'autres enregistrements erronés permet alors de déterminer si l'enregistrement correspond à une vraie personne à recenser ou pas. La couverture nette est estimée sur la base d'un modèle de capture-recapture nommé système dual (Wolter 1986, Fienberg 1992). L'estimateur dual est appliqué dans des cellules homogènes et les résultats recombinés en suivant un modèle synthétique pour obtenir des résultats pour différents domaines de la population (Hogan 2003). Le but du projet n'est pas d'ajuster les chiffres du recensement mais d'obtenir des informations sur la qualité du recensement de l'an 2000 et les potentiels d'améliorations pour les recensements suivants.

Cet article présente les différentes étapes des estimations et les résultats. Les sections 2 et 3 décrivent les jeux de données et les estimateurs de couverture. La section 4 expose les détails de la construction des différents statuts utilisés dans les estimateurs. La section 5 décrit l'approche utilisée pour comparer les valeurs relevées dans le recensement et dans l'enquête pour les personnes appariées de S_P. Les sections 6 et 7 présentent les résultats numériques et la conclusion.

1. Anne Renaud, Service de méthodes statistiques, Office fédéral de la statistique, Espace de l'Europe 10, CH-2010 Neuchâtel, Suisse. Courriel : Anne.Renaud@bfs.admin.ch.

2. Les trois jeux de données

2.1 Recensement

Le recensement de l'an 2000 a été réalisé sous l'égide de l'Office fédéral de la statistique, avec la date de référence du 5 décembre 2000. Des informations ont été relevées pour 7,3 millions d'habitants, 3,1 millions de ménages, 3,8 millions de logements et 1,5 millions de bâtiments. Les différents niveaux ont ensuite été liés par des identificateurs communs lors du traitement des données.

Le relevé des personnes et ménages était sous la responsabilité des 2 896 communes politiques suisses. Ces dernières avaient le choix entre différentes formes de relevé :

- CLASSIC : agents recenseurs;
- SEMI-CLASSIC : préimpression de questionnaires sur la base du registre communal des habitants, envoi par la poste et collecte par des agents recenseurs;
- TRANSIT : préimpression de questionnaires, envoi et retour par la poste;
- FUTURE : idem à TRANSIT avec liens entre les ménages et les logements fournis par la commune;
- TICINO : similaire à TRANSIT mais restreint au canton du Tessin.

La majorité des communes SEMI-CLASSIC, TRANSIT, FUTURE et TICINO offraient également la possibilité de remplir les questionnaires sur Internet. Les 2 208 communes SEMI-CLASSIC, TRANSIT, FUTURE et TICINO qui ont utilisé la préimpression des questionnaires sur la base des registres communaux des habitants contiennent près de 96 % de la population. Les travaux d'envois et de contrôle des retours de la plupart de ces communes étaient organisés dans un centre national.

Le jeu de données des personnes comporte 7 452 075 entrées. Il a la particularité de comporter deux enregistrements pour la même personne si cette dernière a deux domiciles (2,3 % de la population, par exemple étudiant avec un domicile chez ses parents et un domicile proche de son école). Dans le cas de deux domiciles, l'un est codé comme *domicile économique* et l'autre comme *domicile civil*. Le domicile économique est l'endroit où la personne passe le plus de temps par semaine et le domicile civil est l'endroit où la personne a ses papiers officiels (acte de naissance pour les Suisses, permis de séjour pour les étrangers). Dans le cas d'un unique domicile, ce dernier est le domicile économique et le domicile civil tout à la fois. La *population résidante* en Suisse de 7 280 010 personnes est définie par l'ensemble des enregistrements au domicile économique.

Les ménages sont classifiés en *ménages privés*, *collectifs* et *administratifs*. Les ménages privés sont par exemple des familles, des couples ou encore des personnes vivant seules. Les ménages collectifs sont par exemple des groupes de locataires d'une résidence pour personnes âgées ou d'un internat, ou les détenus d'une prison. Les ménages administratifs rassemblent les personnes sans domicile fixe, les gens en voyage et les personnes - par bâtiment ou commune - n'ayant pas pu être assignées à des ménages privés ou collectifs (2,4 % de la population résidante).

Les données du recensement ne comportent aucune imputation au niveau des enregistrements car les communes ont envoyé les informations de base pour les non-répondants (unit nonresponse). Des valeurs sont cependant imputées dans les cas de données manquantes ou d'incohérence dans les questionnaires (item nonresponse).

La *population d'intérêt* pour les estimations de couverture est la population résidante (domicile économique) dans des ménages privés et administratifs. Les ménages collectifs, qui représentent 2,3 % de la population résidante recensée, sont exclus des estimations.

2.2 Échantillon S_P, enquête de couverture et appariement (sous-couverture)

L'objectif de taille de l'échantillon S_P est fixé à environ 50 000 personnes. Faute de bases existantes en Suisse, cette valeur a été déterminée de manière approximative sur la base des expériences faites à l'étranger. Les résultats australiens de 1996 ont notamment été utilisés car le plan d'échantillonnage de leur enquête de couverture était similaire à celui prévu en Suisse en 2000 (ABS 1997).

L'échantillon S_P, indépendant du recensement, est construit en deux parties : le canton du Tessin (TICINO) et le reste de la Suisse (NORD). Les deux parties utilisent un tirage à plusieurs degrés. Le premier degré consiste en la sélection de 303 unités primaires, communes politiques pour TICINO et numéros postaux pour NORD, selon un plan stratifié avec un tirage proportionnel au nombre de bâtiments. Le deuxième degré consiste en un tirage aléatoire simple d'un nombre fixe de 60 bâtiments par unité primaire. Dans le plan NORD, ces bâtiments sont répartis dans un maximum de 3 tournées de distribution du courrier grâce à un degré d'échantillonnage intermédiaire. L'échantillonnage est ainsi construit de manière à regrouper le travail sur le terrain tout en limitant la variabilité des poids. Pour des questions pratiques et de ressources, les numéros postaux comportant une grande proportion de bâtiments sans adresses postales complètes ou codés comme inhabités sont sélectionnés avec une probabilité plus faible que les autres numéros postaux. Il s'agit principalement de numéros postaux dans des régions rurales ou encore de zones industrielles, peu enclins à des défauts de couverture importants. Des listes exhaustives de ménages sont établies sur le terrain avec l'aide des employés postaux dans

l'échantillon d'environ 16 000 bâtiments avant de passer à un sous-échantillonnage de bâtiments permettant d'atteindre un total de environ 27 000 ménages. Pour plus d'information sur l'échantillonnage et la procédure d'enquête, voir Renaud (2001) et, plus détaillé, Renaud et Eichenberger (2002).

L'enquête de couverture consiste à contacter les 27 000 ménages; par téléphone si un numéro est trouvé et en face-à-face si cela n'est pas le cas. On relève les variables permettant un appariement avec le recensement et la définition de sous-groupes intéressants pour l'étude de la couverture (variables sociodémographiques, adresses). Le relevé porte sur tous les membres de tous les ménages des bâtiments sélectionnés.

L'échantillon final S_P contient $n_P = 49\,883$ personnes dans la population d'intérêt (domicile économique et ménage privé). 88 % des ménages ont été atteints par téléphone et 12 % en face-à-face. La pondération dépend de l'échantillonnage et d'un ajustement pour la non-réponse. L'ajustement pour la non-réponse repose sur un modèle d'homogénéité dans des cellules construites sur la base des strates d'échantillonnage et de la connaissance ou non de l'existence d'un numéro de téléphone (interviews prévu par téléphone ou en face-à-face). Il intègre également une estimation de la proportion de vrais ménages parmi les ménages à contacter. Une partie non négligeable des ménages à contacter étaient en effet formée de logements de vacances, de commerces ou encore d'entreprises. Aucun calage n'est appliqué car les données auxiliaires disponibles ne sont pas indépendantes du recensement. Il n'y a pas de non-réponse partielle. Les détails de la pondération sont documentés dans Renaud et Potterat (2004).

Sur la base des questions posées durant l'enquête et des divers contrôles de plausibilité, on fait l'hypothèse que les données de S_P sont correctes et utilisables pour l'appariement avec le recensement. Les critères de qualité utilisés sont les suivants :

- *complétude* : l'enregistrement permet d'identifier la personne;
- *pertinence* : la personne devait bien être recensée;
- *unicité* : la personne est listée une seule fois;
- *appartenance à la population d'intérêt*; la personne est listée au domicile économique et dans un ménage privé;
- *localisation* : la personne est listée à la bonne adresse au jour du recensement.

L'appariement entre l'échantillon S_P et le recensement permet de déterminer le *statut d'appariement* P_j de chaque élément j de S_P. Le statut P_j vaut 1 si l'élément est apparié dans le recensement (personne recensée) et 0 si cela n'est pas le cas (personne non recensée). Dans notre cas, les données relevées durant l'enquête de couverture, les données finales du recensement et les images des questionnaires du recensement sont utilisées pour l'appariement automatique, l'appariement manuel et les contrôles. Aucune interview complémentaire n'a lieu en plus de l'enquête de couverture. Les personnes ayant déménagé entre le jour du recensement et le jour de l'enquête sont échantillonnées à leur adresse au jour de l'enquête puis recherchées en priorité à l'adresse qu'elles ont indiqué pour le jour du recensement. Aucun cas n'est indéterminé à la fin du processus.

2.3 Échantillon S_E et recherche des enregistrements erronés (sur-couverture)

L'objectif de taille de l'échantillon S_E est fixé à environ 55 000 personnes. Cette valeur, un peu supérieure à celle de S_P, n'influe que peu sur le traitement des données car il n'y a pas de travail sur le terrain ni d'interview complémentaire au recensement.

L'échantillon S_E est sélectionné dans les données du recensement selon un tirage à deux degrés. Seuls les éléments faisant partie de la population d'intérêt sont éligibles (enregistrements au domicile économique sans les membres des ménages collectifs). Les unités primaires de S_E sont identiques aux unités primaires de S_P (numéros postaux et communes). La liste des numéros postaux du plan NORD utilisée pour S_P ne correspond cependant pas exactement à la liste des numéros postaux présents dans les données du recensement. Les enregistrements du recensement se trouvant dans des numéros postaux inexistants dans la liste utilisée pour S_P sont donc redistribués dans des numéros existants en tenant compte de la localisation géographique (assignation de numéros postaux fictifs pour l'échantillonnage). Au deuxième degré, on tire des enregistrements de la population d'intérêt selon un plan aléatoire simple, sans degrés intermédiaires. L'allocation est choisie de façon à obtenir des poids constants dans les strates d'échantillonnage des unités primaires. Au final, l'échantillon comporte $n_E = 55\,375$ enregistrements (Renaud 2003).

On fait l'hypothèse que les enregistrements de S_E permettent d'identifier les personnes (complétude) car il y a peu d'imputation dans les données du recensement et que la plupart des questionnaires étaient préimprimés à partir des registres des habitants. La pertinence et l'unicité sont déterminées lors d'un appariement entre S_E et le reste du recensement selon un procédé similaire à l'appariement entre S_P et le recensement. Dans notre cas, il s'agit d'une recherche de doublets ou de triplets des éléments de S_E, complétée par une analyse de cas suspects dans S_E. Un élément j est considéré comme pertinent s'il n'est pas considéré comme erroné dans l'analyse des suspects (par exemple note sur le questionnaire indiquant que la personne

est partie à l'étranger). Un élément j est considéré unique si aucun double/triple n'est détecté dans le recensement. Aucune interview complémentaire n'a lieu auprès du S_E. Il n'y a donc pas d'information complémentaire au recensement sur les personnes de S_E (localisation réelle ? type effectif de domicile et de ménage ?). La recherche des doublets/triplets et des cas suspects aboutit à un *statut d'énumération* E_i pour chaque élément j de S_E. Le statut E_i vaut 1 si l'élément devait bien être énuméré dans le recensement (valeur par défaut) et 0 s'il ne devait pas l'être. En pratique, il peut prendre des valeurs entre 0 et 1 si le cas n'est pas déterminé de manière précise. Ainsi, les doublets et triplets reçoivent respectivement les valeurs 1/2 et 1/3 s'il n'y a pas d'informations permettant de déterminer l'enregistrement correct parmi les enregistrements détectés. Ces cas, rares, correspondent à des personnes ayant remplis plusieurs questionnaires au recensement sans qu'un lien n'ait été fait entre eux durant le traitement des données.

3. Estimateurs de la couverture

3.1 Sous-couverture et sur-couverture

Le *taux de sous-couverture* est estimé par $\hat{R}_{\text{sous}} = 1 - \hat{R}_m$, où $\hat{R}_m$ est l'estimation du *taux d'appariements corrects* basée sur l'échantillon S_P. De façon similaire, on définit le *taux de sur-couverture* $\hat{R}_{\text{sur}} = 1 - \hat{R}_c$, avec $\hat{R}_c$ l'estimation du *taux d'enregistrements corrects* basée sur l'échantillon S_E. Les taux d'appariements corrects et d'enregistrements corrects sont estimés par les moyennes pondérées des statuts d'appariement P_j et d'énumération E_j :

$$\hat{R}_m = \frac{\sum_{j \in S_P} w_{P,j} P_j}{\sum_{j \in S_P} w_{P,j}} \quad \text{et} \quad \hat{R}_c = \frac{\sum_{j \in S_E} w_{E,j} E_j}{\sum_{j \in S_E} w_{E,j}}, \quad (1)$$

avec $w_{P,j}$ le poids de l'élément j de l'échantillon S_P et $w_{E,j}$ le poids de l'élément j de l'échantillon S_E. Nous notons que le dénominateur de $\hat{R}_c$ est la somme des poids $w_{E,j}$ de S_E et non pas le nombre C d'enregistrements connus dans le recensement afin d'avoir un estimateur potentiellement moins biaisé.

L'estimation des taux de sous-couverture et de sur-couverture dans un domaine d est donnée par $\hat{R}_{\text{sous},d} = 1 - \hat{R}_{m,d}$ et $\hat{R}_{\text{sur},d} = 1 - \hat{R}_{c,d}$, avec

$$\hat{R}_{m,d} = \frac{\sum_{j \in S_P} w_{P,j} P_j I_{jd}}{\sum_{j \in S_P} w_{P,j} I_{jd}} \text{ et } \hat{R}_{c,d} = \frac{\sum_{j \in S_E} w_{E,j} E_j J_{jd}}{\sum_{j \in S_E} w_{E,j} J_{jd}}. \quad (2)$$

Les identificateurs I_{jd} et J_{jd} prennent la valeur 1 si l'élément j, respectivement de S_P et S_E, se trouve dans le domaine d et la valeur 0 sinon.

3.2 Couverture nette

Le *taux de sous-couverture nette* est estimé par $\hat{R}_{\text{sousnet}} = 1 - \hat{R}_{\text{net}}$ avec $\hat{R}_{\text{net}} = C / \hat{N}$ l'estimation du *taux de couverture nette*, C le nombre recensé dans la population d'intérêt et $\hat{N}$ l'estimation du vrai total dans la population d'intérêt. Si $\hat{R}_{\text{sousnet}}$ est négatif, nous sommes en présence d'une sur-couverture nette.

L'estimation du vrai total $\hat{N}$ est basée sur le modèle dual (Wolter 1986). Ce modèle repose sur un principe de capture (recensement) et recapture (enquête de couverture). Il est appliqué dans des cellules d'estimation $k = 1, ..., K$ afin de satisfaire au mieux les hypothèses du modèle; voir discussion plus bas. Ainsi, l'estimation du vrai total $\hat{N}$ est composée de la somme des vrais totaux estimés $\hat{N}_k$ dans des cellules d'estimations disjointes recouvrant l'ensemble de la population d'intérêt $k = 1, ..., K$:

$$\hat{N} = \sum_{k=1}^{K} \hat{N}_k. \quad (3)$$

Les totaux estimés $\hat{N}_k$ ont la forme donnée par le modèle dual :

$$\hat{N}_k = [N_{1+,k}] \left[\frac{N_{+1,k}}{N_{11,k}} \right], \quad (4)$$

avec $N_{1+,k}$ le total des enregistrements correctement comptés dans la cellule k durant la capture (recensement), $N_{+1,k}$ le total dans k durant la recapture (estimé sur la base de l'échantillon S_P) et $N_{11,k}$ le nombre d'enregistrements communs aux deux listes (estimé sur la base des appariements entre S_P et le recensement).

Les différents termes de l'équation (4) sont estimés sur la base des estimations de sous-couverture et de sur-couverture. Il s'agit d'une extension du modèle de Wolter (1986) proche de celle utilisée notamment par Hogan (2003). Ainsi, le total des enregistrements correctement comptés dans le recensement $N_{1+,k}$ est estimé par le produit du total recensé C_k par le taux d'enregistrements corrects $\hat{R}_{c,k}$ afin de tenir compte de la sur-couverture. De plus, le rapport entre le total dans la recapture $N_{+1,k}$ et le nombre d'enregistrements communs aux deux listes $N_{11,k}$ est estimé par l'inverse du taux d'appariement $\hat{R}_{m,k}$ entre l'enquête de couverture et le recensement afin de tenir compte de la sous-couverture. On obtient :

$$\hat{N}_k = [C_k \hat{R}_{c,k}][\hat{R}_{m,k}^{-1}] = C_k [\hat{R}_{c,k} \hat{R}_{m,k}^{-1}] = C_k \hat{F}_k, \quad (5)$$

avec $\hat{F}_k = \hat{R}_{c,k} \hat{R}_{m,k}^{-1}$ le *facteur de correction de la couverture* dans la cellule k. Le facteur $\hat{F}_k$ combine les effets de sur-couverture et de sous-couverture de la cellule k estimés sur la base des échantillons S_P et S_E. On note qu'une sous-couverture dans un domaine peut être compensée par une sur-couverture dans le même domaine.

Ainsi une sous-couverture nette nulle dans un domaine ne signifie pas qu'aucun défaut de couverture existe dans ce domaine.

Les estimations proposées reposent sur les hypothèses du modèle dual, le choix des cellules d'estimations, et le choix des statuts définissant les estimateurs $\hat{R}_{c,k}$ et $\hat{R}_{m,k}$. Le modèle dual est intéressant car il prend en compte le fait que certaines personnes ne sont atteintes ni par le recensement (capture) ni par l'enquête de couverture (recapture). Cependant, une série de contraintes doivent être respectées afin d'éviter des biais d'estimation. L'enquête de couverture et le recensement doivent être totalement indépendants. L'appariement doit être de très bonne qualité. Le modèle doit être appliqué dans des cellules avec des personnes ayant la même probabilité d'être énumérées dans le recensement, respectivement dans l'enquête; voir la section 3.3. Finalement, la population ne doit pas trop changer entre le jour du recensement et celui de l'enquête. De leur côté, les estimateurs $\hat{R}_{c,k}$ et $\hat{R}_{m,k}$ sont basés sur la qualité de l'appariement et de la recherche des enregistrements erronés. De plus, il s'agit de faire en sorte que la définition d'un appariement correct dans S_P et celle d'un enregistrement correct dans S_E soient identiques, *i.e.*, équilibre entre sur- et sous-couverture (« balancing »); voir la section 4. Tous ces éléments sont pris en compte dans la mesure du possible dans les présentes estimations.

L'estimation de la sous-couverture nette dans un domaine d a la forme $\hat{R}_{\text{sousnet},d} = 1 - \hat{R}_{\text{net},d} = 1 - C_d / \hat{N}_d$ avec C_d le nombre recensé dans le domaine et $\hat{N}_d$ l'estimation du vrai total. L'estimation du vrai total $\hat{N}_d$ est basée sur un modèle nommé *synthétique* qui suppose que le facteur de correction est fixe dans chaque cellule $k = 1, \ldots, K$:

$$\hat{N}_d = \sum_{k=1}^{K} \hat{N}_{k,d} = \sum_{k=1}^{K} C_{k,d}\, \hat{F}_k. \qquad (6)$$

$C_{k,d}$ est le nombre recensé dans la population d'intérêt dans l'intersection entre la cellule k et le domaine d et $\hat{F}_k$ est le facteur de correction de la couverture dans la cellule k. L'hypothèse du modèle synthétique est respectée si le comportement de tout sous-ensemble dans la cellule est identique à celui de la cellule entière. Cette homogénéité doit être contrôlée au mieux par le choix des cellules. On reprend ici les cellules homogènes définies pour le modèle dual.

3.3 Cellules d'estimation

Les cellules d'estimation $k = 1, \ldots, K$ sont construites de façon à grouper les éléments ayant des probabilités d'énumération homogènes dans le recensement, respectivement dans l'enquête, (hypothèse duale) et des taux de couverture nette homogènes (hypothèse synthétique). On désire un minimum de 100 personnes par cellule dans S_E et S_P afin de contrôler la variance et limiter le biais d'estimation. Les variables définissant les cellules sont sélectionnées à l'aide d'un modèle de régression logistique et d'une méthode de discrimination appliqués sur les données de S_P (variable binaire : P_j). Les trois variables les plus influentes sont croisées : nationalité en 2 catégories, état civil en 2 catégories et tailles de la commune en 3 catégories. Les autres variables sont ensuite intégrées successivement en faisant des regroupements lorsque les tailles de cellules sont trop petites (langue officielle de la commune en 2 catégories, classe d'âge en 7 catégories et sexe en 2 catégories). Au final, nous obtenons 121 cellules d'estimations; voir Renaud (2004) pour plus de détails.

3.4 Variance des estimateurs de couverture

La variance des estimateurs est estimée par un jackknife stratifié appliqué sur les unités primaires - identiques - de S_P et S_E. On note que la variance de la sous-couverture estimée $\hat{R}_{\text{sous}} = 1 - \hat{R}_m$ est égale à la variance du taux d'appariement estimé $\hat{R}_m$. De même la variance de la sur-couverture $\hat{R}_{\text{sur}} = 1 - \hat{R}_c$ est égale à celle du taux d'enregistrement correct $\hat{R}_c$ et la variance de la sous-couverture nette $\hat{R}_{\text{sousnet}} = 1 - \hat{R}_{\text{net}}$ est égale à celle du taux de couverture nette $\hat{R}_{\text{net}}$.

Soit θ le paramètre d'intérêt prenant la forme d'une moyenne pondérée de statuts dans le cas de la sous-couverture et de la sur-couverture, et la forme d'une fonction linéaire de quotients entre deux moyennes pondérées dans le cas de la sous-couverture nette. Son estimateur est $\hat{\theta}$.

Soient $h = 1, \ldots, H$ la strate utilisée au premier degré de l'échantillonnage, $i = 1, \ldots, m_h$ le numéro de l'unité primaire dans la strate h (numéro postal pour NORD ou commune pour TICINO), et $j = 1, \ldots, n_{hi}$ le numéro de la personne dans l'unité primaire i de h. Pour les besoins du jackknife, les échantillons S_P et S_E sont partitionnés, dans chaque strate h, en m_h sous-ensembles correspondant aux personnes dans les unités primaires $\alpha = 1, \ldots, m_h$.

Soit $\hat{\theta}_{(h\alpha)}$ l'estimateur ayant la même forme que $\hat{\theta}$ mais calculé sur l'échantillon auquel on a retiré l'unité primaire α de la strate h. On note que les estimateurs $\hat{R}_{m(h\alpha),k}$ et $\hat{R}_{c(h\alpha),k}$, $k = 1, \ldots, K$ sont combinés pour former $\hat{R}_{\text{net}(h\alpha)}$:

$$\hat{R}_{\text{net}(h\alpha)} = C\left[\sum_{k=1}^{K} C_k \frac{\hat{R}_{c(h\alpha),k}}{\hat{R}_{m(h\alpha),k}}\right]^{-1}. \qquad (7)$$

Les poids corrigés w'_{hij} utilisés pour le calcul des $\hat{R}_{m(h\alpha)}$ et $\hat{R}_{c(h\alpha)}$ ont la forme :

$$w'_{hij} = \begin{cases} 0 & \text{si} \quad i = \alpha \\ w_{hij}\dfrac{m_h}{m_h - 1} & \text{si} \quad \alpha \in h \text{ et } i \neq \alpha \\ w_{hij} & \text{si} \quad \alpha \notin h. \end{cases} \qquad (8)$$

Cette forme de correction est préférée au quotient entre la somme des poids des éléments dans la strate et la somme des poids sans l'unité primaire α car elle permet de prendre en compte la variabilité due au nombre inconnu d'éléments dans la strate.

L'estimateur du jackknife devient :

$$\hat{\theta}_{JK} = \frac{\sum_h \sum_{\alpha=1}^{m_h} \hat{\theta}_{h\alpha}}{\sum_h m_h}, \qquad (9)$$

avec les pseudo valeurs $\hat{\theta}_{h\alpha} = m_h \hat{\theta} - (m_h - 1)\hat{\theta}_{(h\alpha)}$. L'estimateur de sa variance peut prendre différentes formes; voir par exemple Shao and Tu (1995). Nous appliquons la forme suivante :

$$v(\hat{\theta}_{JK}) = \sum_h \frac{m_h - 1}{m_h} \sum_{\alpha=1}^{m_h} (\hat{\theta}_{(h\alpha)} - \hat{\theta}_{(h.)})^2, \qquad (10)$$

avec $\hat{\theta}_{(h.)} = \sum_{\alpha=1}^{m_h} \hat{\theta}_{(h\alpha)} / m_h$. Finalement, on utilise $v(\hat{\theta}_{JK})$ comme estimateur de la variance de $\hat{\theta}$. Les estimations dans des sous-groupes utilisent la même forme de l'estimateur avec l'intégration d'un indicateur de domaine dans la construction des $\hat{\theta}_{(h\alpha)}$. Aucune correction pour la population finie n'est appliquée dans les estimations. De plus on ne tient pas compte d'autres variabilités telle que celle induite par le modèle de pondération pour la non-réponse dans S_P.

Des problèmes, tels le manque de stabilité de l'estimation dans les strates avec peu d'unités primaires, sont apparus durant l'application. Les essais de partage de certaines unités primaires et une comparaison avec la linéarization de Taylor ou un jackknife simple permettent cependant de penser que les estimateurs de variance par le jackknife stratifié présentés dans ce document sont plutôt conservateurs.

4. Choix des statuts d'appariement et d'énumeration corrects

Un élément clé des estimations de couverture est la définition de *statut d'appariement correct* pour les éléments de S_P et du *statut d'énumération correct* pour les éléments de S_E. Ces statuts corrects sont définis à partir des statuts de bases P_j et E_j déterminés durant les appariements.

Un appariement avec un élément du recensement faisant partie d'un ménage collectif est-t-il accepté comme appariement correct pour un élément de S_P ou s'agit-il d'une sous-couverture de la population d'intérêt ? Un double hors de la population d'intérêt pour un élément de S_E est-il vraiment considéré comme un double, et donc une sur-couverture, ou devrait-il être exclu? Une définition claire s'impose. De plus, les statuts utilisés dans les estimations de la sous-couverture nette doivent être choisis de manière à satisfaire l'équilibre entre la sur- et la sous-couverture; voir la notion d'équilibrage ou de « balancing » par exemple dans Hogan (2003). Un appariement ($P_j = 1$) avec un élément hors population d'intérêt peut par exemple être refusé comme appariement correct (statut d'appariement correct = 0, pas de sous-couverture) uniquement si la recherche des enregistrements corrects détecterait cet élément également comme incorrect car hors population (statut d'énumération correct = 0, pas de sur-couverture).

Les critères de définition des statuts corrects sont construits à partir des informations disponibles pour les éléments de S_P et S_E. Pour ce qui concerne S_P, nous partons de l'hypothèse que les enregistrements du recensement qui ont été appariés à des éléments de S_P permettent d'identifier les personnes (complétude) et que ces personnes devaient bien être recensées (pertinence). On estime aussi qu'ils sont uniques car l'unicité, bien que pas contrôlée pour les appariements, est réalisée dans la grande majorité des cas contrôlés dans S_E. Les critères d'appartenance à la population et de localisation sont contrôlés par comparaison avec les informations relevées dans l'enquête de couverture; considérées comme référence. Aucun relevé complémentaire n'était organisé pour régler les cas peu clairs. Pour ce qui concerne S_E, nous avons à disposition le critère de complétude considéré comme respecté dans les données du recensement et les résultats concernant l'unicité et la pertinence obtenus dans l'appariement avec le reste du recensement. Pour les doubles et triples, on définit $E_j = 1/d'$, avec d' = nombre de doubles/triples dans la population d'intérêt selon le recensement. Les critères d'appartenance à la population d'intérêt et de localisation des éléments de S_E. ne peuvent être contrôlés car nous n'avons pas de données de référence complémentaires au recensement.

Pour les estimations de sous-couverture nette, il est important de satisfaire la contrainte d'équilibrage. Les critères utilisés dans la définition des statuts corrects sont donc la complétude, la pertinence et l'unicité. Les critères de l'appartenance à la population d'intérêt et de la localisation ne peuvent pas être considérés car ils ne sont pas utilisables dans la définition du statut d'énumération correct. Les critères de complétude, pertinence et unicité sont déjà intégrés dans la construction des statuts de base P_j et E_j. On fait donc les estimations avec les statuts de base P_j et E_j.

Pour les estimations n'utilisant pas le système dual et le besoin d'équilibrage, il est possible d'utiliser d'autres critères pour définir les statuts corrects. D'autres types de statuts d'appariements corrects sont notamment utilisés dans l'analyse des erreurs potentielles de mesure de la section 5 et les analyses plus détaillées des appariements et des énumérations présentées dans Renaud (2004).

5. Comparaison des appariements

5.1 Erreurs potentielles de mesure

Les erreurs de mesures ou erreurs de classification sont liées aux erreurs de couverture. En effet, une personne classifiée dans le domaine d selon le recensement (par exemple personne entre 10 et 19 ans) alors qu'elle est en réalité hors du domaine (par exemple personne de 60 ans) aboutirait à une sur-couverture dans le domaine d et une sous-couverture hors de ce domaine. Cette erreur de classification (« misclassification ») ne provoque pas d'erreur de couverture au niveau global mais une erreur au niveau de sous-groupes de la population.

Les raisons des différences entre les valeurs relevés dans deux enquêtes telles que le recensement et l'enquête de couverture peuvent être très diverses et complexes à dissocier. Il faut en effet compter sur des erreurs d'appariement, des différences résultant des méthodes de relevé (questionnaire papier ou téléphone/face-à-face) et du traitement des données, ou encore sur des différences réelles dues au décalage temporel entre les relevés (décembre 2000 ou avril-mai 2001). De plus, il est difficile de déterminer la réponse correcte si on a deux valeurs différentes : le recensement ? l'enquête ? une autre valeur pas relevée ?

Les erreurs de mesures potentielles des données du recensement sont analysées sur la base de l'ensemble des appariements entre l'échantillon indépendant S_P et le recensement. On choisit de déterminer quelles sont les variables qui montrent respectivement peu ou beaucoup de problèmes potentiels de classification, sans faire un jugement sur la qualité de l'un ou l'autre des relevés. Ces informations sont notamment utiles pour évaluer le choix des cellules d'estimations pour le système dual et choisir les sous-groupes pour lesquels les estimations des défauts de couverture sont les plus fondées.

On définit, pour la variable catégorielle X, le taux d'appariement dans le bon domaine $\hat{R}_X$ comme suit :

$$\hat{R}_X = \frac{\sum_{j \in S_P \text{ match}} w_{P,j} P_{X,j}}{\sum_{j \in S_P \text{ match}} w_{P,j}}, \qquad (11)$$

avec $w_{P,j}$ le poids de l'élément j de l'échantillon S_P apparié (S_P match) et le *statut de classification* $P_{X,j}$ qui vaut 1 si l'élément j se trouve dans la même classe dans le recensement et dans l'enquête, et 0 sinon. La valeur de $\hat{R}_X$ est estimée avec l'ensemble des éléments appariés et avec le sous-groupe des éléments sans imputation dans le recensement.

On définit également une mesure d'asymétrie $\varphi_X(d,\ d')$ pour les classes d et d' de la variable X :

$$\varphi_X(d,\ d') = \frac{\sum_{j \in S_P \text{ match}} w_{P,j} I_j(d,\ d')}{\sum_{j \in S_P \text{ match}} w_{P,j} I_j(d',\ d)}, \qquad (12)$$

avec $I_j(d,\ d') = 1$ si l'élément j se trouve dans le domaine d selon l'enquête et dans le domaine d' selon le recensement, et 0 sinon. Le facteur $\varphi_X(d,\ d')$ vaut 1 s'il y a équilibre dans les erreurs de classification, c'est-à-dire si le nombre d'éléments dans d selon l'enquête et dans d' selon le recensement est égal au nombre dans d' selon l'enquête et dans d selon le recensement. Plus il s'éloigne de 1, plus l'équilibre est rompu.

5.2 Erreurs potentielles de localisation

Les comparaisons entre le recensement et l'enquête permettent également d'étudier la localisation géographique des personnes. Dans les données du recensement, nous avons une unique adresse si la personne a un unique domicile et deux adresses - principale et secondaire - si la personne a deux domiciles. Dans les données de l'enquête, nous avons une ou deux adresses au jour du recensement, une ou deux adresses au jour de l'enquête et l'information sur un éventuel déménagement entre les deux dates. Si une personne a un unique domicile et n'a pas déménagé, son adresse principale au jour du recensement et son adresse principale au jour de l'enquête sont identiques. Elle n'a pas d'adresses secondaires.

Différentes mesures de distance sont envisagées pour déterminer les erreurs potentielles de localisation dans le recensement. Pour des raisons pratiques, notamment de données à disposition, nous définissons des aires géographiques autour de l'adresse principale de la personne telle que relevée durant l'enquête pour le jour du recensement (*adresse de référence*). Les aires sont des ensembles de communes politiques. Elles sont définies à partir des numéros postaux relevés durant l'enquête. L'*aire de base* de la personne est définie par l'ensemble des communes qui possèdent des bâtiments dans le numéro postal de son adresse de référence. La définition de cette aire utilise les données du registre suisse des bâtiments car ce dernier possède une information sur l'adresse postale et la commune des bâtiments. L'*aire étendue* comporte les communes de l'aire de base et l'ensemble des communes qui leurs sont adjacentes; voir Renaud (2004) pour des exemples.

De façon similaire aux erreurs de classification, les erreurs de localisation ne provoquent pas d'erreurs de couverture au niveau global mais des erreurs au niveau de sous-groupes tels des régions ou des types de communes. Différents taux peuvent être définis. On retiendra le taux de localisation de base et le taux de localisation étendue; tous deux pondérés par w_{Pj} le poids de l'élément j de l'échantillon S_P apparié. Le statut de localisation prend la valeur 1 si l'élément est trouvé dans l'aire de base ou respectivement étendue, et 0 sinon. On étudiera en particulier la localisation des personnes qui ont déménagé, afin

de détecter d'éventuels problèmes de décalage temporel entre le jour du recensement et le jour du relevé effectif des données du recensement.

6. Résultats

6.1 Estimations des défauts de couverture

Le taux global de sous-couverture nette est estimé à 1,41 % avec un écart-type de 0,12 %. Le taux de sur-couverture est de 0,35 % (écart-type = 0,03 %) et le taux de sous-couverture est de 1,64 % (écart-type = 0,11 %). Ces résultats sont dans l'ordre de grandeur de ceux d'autres pays, bien que plutôt dans les valeurs les plus basses; voir Tableau 1.

La sur-couverture est peu importante dans la grande majorité des domaines étudiés. Le plus haut taux est observé pour les personnes entre 20 et 31 ans (0,93 % avec un écart-type de 0,09 %); voir Tableau 2. La sous-couverture est par contre élevée dans plusieurs domaines. On note par exemple un taux de 8,03 % (écart-type = 0,85 %) pour les étrangers avec des permis d'établissement temporaires (« autres permis ») et un taux de 3,50 % (écart-type = 0,50 %) pour les 20-31 ans. On note aussi un taux de sous-couverture de 2,4 % dans la région italophone du pays (langue de la commune : italien, région NUTS : Ticino, et relevé : TICINO). Les résultats sont cependant liés à une relativement grande variabilité (écart-type de env. 0,5 %) car les échantillons S_P et S_E ne comportent respectivement que 1 500 et 1 700 personnes dans cette région.

Tableau 1
Comparaison internationale des résultats globaux. Taux estimés de sur-couverture $\hat{R}_{sur}$, sous-couverture $\hat{R}_{sous}$ et sous-couverture nette $\hat{R}_{sousnet}$, avec écarts-type estimés correspondants. Références : Statistique Canada (1999, 2004), Hogan (1993, 2003), McLennan (1997) et Trewin (2003)

		Sur-couverture [%]	Sous-couverture [%]	Sous-couverture nette [%]
Suisse	2000	0,3 (0,03)	1,64 (0,11)	1,41 (0,12)
Canada	1996	0,74 (0,04)	3,18 (0,09)	2,45 (0,10)
	2001	0,96 (0,05)	3,95 (0,13)	2,99 (0,14)
États-Unis	1990	3,1	4,7	1,6 (0,10)
	2000	-	-	-0,5 (0,21)
Australie	1996	0,2	1,8	1,6 (0,10)
	2001	0,9	2,7	1,8 (0,10)

La sous-couverture nette est positive dans tous les domaines étudiés. Il n'y a donc pas de sur-couverture nette. Les plus grandes valeurs sont observées pour les étrangers avec permis permanent ou temporaire (2,89 % et 3,48 %, écarts-type = 0,32 % et 0,39 %) ainsi que chez les 20-31 ans (2,84 %, écart-type = 0,36 %). Aucune différence significative n'est observée entre hommes et femmes, entre langues et entre régions NUTS. La faible taille de l'échantillon avec la variante de relevé TICINO ne permet pas de différencier cette méthode des autres utilisées dans le pays. Des différences sont par contre significatives entre états civils, ainsi qu'entre les types et tailles de communes.

On note que le taux de sous-couverture nette est supérieur au taux de sous-couverture dans le cas des permis d'établissement permanent. Cet effet, irréaliste, est dû au choix des cellules d'estimations et au lissage qui s'en suit. La construction des cellules a en effet nécessité un regroupement des étrangers avec permis permanents et temporaires en une seule catégorie lors des agrégats permettant d'atteindre la taille minimale de 100 personnes par cellule. Par ce regroupement nous considérons les étrangers comme un groupe homogène alors qu'il ne l'est pas. Ceci montre les limites de la méthode et la difficulté de satisfaire aux hypothèses des modèles utilisés lors de l'application. Dans le cas des étrangers, on note cependant que les intervalles de confiance des taux de sous-couverture nette et de sous-couverture se recoupent. Les conséquences des faiblesses de l'application sont donc restreintes.

Notons encore que les résultats sont présentés dans des domaines définis par des variables pour lesquelles on a observé de faibles erreurs potentielles de mesure. Des résultats pour des groupes tels que définis par les caractéristiques de ménage ou de vie active ne seraient en effet que peu fiables; voir section 6.2.

La précision des résultats obtenus est en général meilleure que l'objectif fixé au début du projet. Ce dernier était en effet d'avoir un écart-type de 0,3 % pour des sous-groupes de 10 000 individus dans S_P. Dans le cas par exemple des classes d'âge 32-44 et 45-59 qui comportent entre 10 000 et 12 000 personnes les écart-types sont 0,19 et 0,14 %.

6.2 Erreurs potentielles de mesure et de classification

Parmi les 49 107 éléments appariés entre l'enquête de couverture et le recensement, 96 % ne présentent aucune différence dans le sexe, les 7 classes d'âge, les 3 classes d'état civil et les 3 classes de permis d'établissement (suisse, permanent, temporaire). Le taux d'appariement dans le bon domaine $\hat{R}_X$ vaut 99,3 % pour le sexe (avec et sans imputations), 98,3 % pour l'état civil (98,4 % parmi les valeurs non imputées) et 98,7 % pour le permis d'établissement (98,8 % parmi les valeurs non imputées). Le taux $\hat{R}_X$ vaut 99,5 % pour les classes d'âge (avec ou sans imputations). Il faut cependant noter que la date de naissance était, avec le nom et le prénom, une des principales variables dans l'appariement. Les différences d'âge sont donc possibles uniquement lors d'un appariement non automatique (assisté par ordinateur ou manuel). Trois variables montrent un taux d'appariement dans le bon domaine nettement plus

faible que celui observé pour le sexe, l'âge, le permis et l'état civil. Il s'agit de la variable sur la vie active (actif, sans emploi, non actif), la position dans le ménage (seul/seule, époux/épouse, union libre, une personne avec enfant(s), autre chef de ménage, apparenté au chef de ménage, autres; résultats restreints aux ménages privés) et la taille du ménage de la personne (selon domicile économique et dans les ménages privés). Le taux $\hat{R}_X$ vaut 90,4 % pour la vie active (91,1 % parmi les valeurs non imputées), 91,4 % pour la position dans le ménage (94,9 % parmi les valeurs non imputées) et 88,3 % pour la taille du ménage.

Tableau 2 Nombre recensé *C* et taux estimés de sur-couverture $\hat{R}_{sur}$, sous-couverture $\hat{R}_{sous}$ et sous-couverture nette $\hat{R}_{sousnet}$ pour différents domaines [%], avec les écart-types estimés (*E-T*) correspondants

Variable	Catégories	*C*	$\hat{R}_{sur}$	E-T	$\hat{R}_{sous}$	E-T	$\hat{R}_{sousnet}$	E-T
global		7 121 626	0,35	0,03	1,64	0,11	1,41	0,12
sexe	homme	3 497 940	0,37	0,04	1,74	0,13	1,46	0,13
	femme	3 623 686	0,33	0,03	1,55	0,10	1,37	0,13
classe âge	≤ 9	810 373	0,26	0,05	1,46	0,21	1,34	0,26
	10-19	833 185	0,27	0,05	1,30	0,19	1,04	0,22
	20-31	1 115 804	0,93	0,09	3,50	0,34	2,84	0,36
	32-44	1 544 721	0,33	0,05	1,65	0,16	1,43	0,19
	45-59	1 431 771	0,22	0,04	1,18	0,14	1,04	0,14
	60-79	1 146 709	0,10	0,03	0,91	0,13	0,82	0,12
	≥ 80	239 063	0,11	0,06	1,20	0,31	1,03	0,27
permis d'établissement	Suisse	5 674 266	0,33	0,03	1,28	0,09	0,98	0,10
	étranger permanent	1 020 242	0,33	0,06	1,85	0,29	2,89	0,32
	étranger temporaire	427 118	0,56	0,11	8,03	0,85	3,48	0,39
état civil	célibataire	2 975 643	0,50	0,05	2,07	0,18	1,72	0,19
	marié/e	3 377 223	0,23	0,04	1,27	0,11	1,25	0,12
	veuf/veuve	369 339	0,25	0,08	1,23	0,26	0,79	0,13
	divorcé/e	399 421	0,24	0,08	1,95	0,35	1,02	0,10
langue commune	allemand + romanche	5 128 353	0,33	0,04	1,50	0,11	1,28	0,12
	français	1 680 062	0,35	0,06	1,89	0,25	1,79	0,27
	italien	313 211	0,53	0,12	2,35	0,49	1,56	0,19
région NUTS	région lémanique	1 296 464	0,37	0,07	2,19	0,38	1,84	0,28
	espace Mittelland	1 640 489	0,35	0,09	1,39	0,15	1,25	0,10
	Nordwestschweiz	976 699	0,18	0,04	1,50	0,27	1,32	0,12
	Zurich	1 221 014	0,31	0,05	1,58	0,19	1,46	0,13
	Ostschweiz	1 020 897	0,40	0,07	1,29	0,23	1,24	0,12
	Zentralschweiz	665 904	0,36	0,06	1,57	0,25	1,19	0,12
	Ticino	300 159	0,54	0,12	2,38	0,52	1,57	0,19
taille commune	petite	1 372 958	0,34	0,05	1,50	0,15	1,12	0,14
	moyenne	2 398 256	0,41	0,07	1,32	0,16	1,07	0,19
	grande	3 350 412	0,31	0,03	2,01	0,19	1,77	0,19
type	ville	2 078 780	0,35	0,04	1,96	0,17	1,82	0,20
	agglomération	3 145 541	0,36	0,06	1,49	0,19	1,34	0,12
	rural	1 897 305	0,32	0,04	1,56	0,17	1,07	0,12
relevé	CLASSIC	265 607	0,39	0,05	1,91	0,28	1,07	0,12
	SEMI-CLASSIC	174 501	0,37	0,08	1,07	0,24	1,16	0,13
	TRANSIT + FUTURE	6 381 359	0,33	0,03	1,62	0,11	1,42	0,12
	TICINO	300 159	0,54	0,12	2,38	0,52	1,57	0,19

La mesure de l'asymétrie $\varphi_X(d, d')$ prend la valeur 1,33 pour le sexe (d = homme et d' = femme). Il y a plus de personnes codées hommes selon l'enquête qui sont codées femmes dans le recensement que de femmes selon l'enquête qui sont des hommes selon le recensement. La proportion d'hommes est légèrement supérieure dans l'enquête. Ces résultats doivent cependant être relativisés car ils sont basés sur très peu de cas; voir le Tableau 3. Un test de McNemar est juste significatif au seuil de 5 % sans tenir compte du plan, mais il ne le reste pas si le plan est pris en compte. On observe des asymétries par contre très nettes dans l'état civil. Il y a moins de célibataires dans l'enquête qui sont mariés dans le recensement que l'inverse (facteur 0,33 pour d = célibataire et d' = marié). De même, il y a moins de mariés dans l'enquête qui sont veuf/veuve ou divorcé/e dans le recensement que l'inverse (facteur 0,42 pour d = marié et d' = autres). L'asymétrie est également visible sur la variable du permis d'établissement. La tendance est d'avoir plus de Suisses dans l'enquête qui sont notés étrangers dans le recensement que l'inverse; de même plus de permis permanents dans l'enquête et permis temporaires dans le recensement que l'inverse (facteurs 5,22 pour d = suisse et d' = étranger avec permis permanent et 3,83 pour d = étranger avec permis permanent et d' = étranger avec permis temporaire). Les facteurs calculés se basent sur peu de cas. Ils donnent cependant un aperçu des différences potentielles entre un relevé par le biais du questionnaire du recensement et une enquête réalisée principalement par téléphone. La variable de la vie active comporte plus de cas divergents; voir le Tableau 4. Nous avons donc par exemple moins de personnes occupées dans l'enquête et non actives dans le recensement que l'inverse (facteur de 0,46 pour d = actif et d' = non actif). Il y a également moins de personnes sans emploi dans l'enquête et non actives dans le recensement que l'inverse (facteur 0,26 pour d = sans emploi et d' = non actif). La variable de la position dans le ménage montre également des asymétries, mais ces dernières sont basées sur peu d'éléments car la dispersion des éléments dans les cases (d, d') est importante. Les variables du recensement au niveau ménage (position dans le ménage et taille de ménage) sont influencées par le processus complexe de la formation des ménages. Elles sont moins fiables que celles qui concernent les personnes. Les valeurs au niveau ménage sont plus fiables dans l'enquête.

Tableau 3 Comparaison des valeurs relevées dans l'enquête et le recensement pour la variable sexe

sexe			Enquête		
			homme	femme	total
Recensement	non apparié	total	393	383	776
	apparié	total	24 171	24 936	49 107
	apparié	homme	23 967	166	24 133
		femme	204	24 770	24 974
	apparié	homme	6	0	6
	(variable imputée)	femme	0	13	13
	total		24 564	25 319	49 883

Tableau 4 Comparaison des valeurs relevées dans l'enquête et le recensement pour la variable vie active

vie active			Enquête				
			occupé	sans emploi	non actif	≤ 15 ans	total
Recensement	non apparié	total	424	23	217	112	776
	apparié	total	25 163	498	14 501	8 945	49 107
	apparié	occupé	23 953	188	2 007	13	26 161
		sans emploi	300	221	323	1	845
		non actif	901	89	12 143	18	13 151
		≤ 15 ans	9	0	28	8 913	8 950
	apparié	occupé	564	22	312	6	904
	(variable imputée)	sans emploi	14	8	26	1	49
		non actif	92	15	881	5	993
		≤ 15 ans	0	0	0	0	0
	total		25 587	521	14 718	9 057	49 883

6.3 Erreurs potentielles de localisation et décalage temporel

Parmi les 49 107 éléments appariés entre l'enquête de couverture et le recensement, 97,7 % sont trouvés dans l'aire de base autour de l'adresse de référence relevée durant l'enquête. Cette valeur vaut 98,1 % pour les personnes n'ayant indiqué aucun déménagement entre le jour du recensement et le jour de l'enquête. Elle vaut 83,9 % pour ceux qui ont indiqué un déménagement (1 512 personnes); voir nombres absolus dans le Tableau 5.

Il est intéressant de noter que 9,4 % des personnes du NORD n'ayant pas déménagé sont retrouvées proche de leur adresse de référence mais pas exactement dans le même bâtiment. Ces problèmes de localisation fine ont une influence négligeable sur les données du recensement. Ils montrent cependant la difficulté à identifier les bâtiments échantillonnés lors de la construction des listes des ménages sur le terrain durant l'enquête, tout comme la difficulté de l'assignation des personnes aux bâtiments durant le traitement des données du recensement. Un relevé complémentaire serait cependant nécessaire pour évaluer l'effet respectif des deux difficultés.

La localisation des personnes qui ont déménagé indique que $151 = 145 + 6$ personnes sont trouvées autour de leur adresse au jour de l'enquête et non pas autour de leur adresse au jour du recensement (9 %, pondéré). De plus, un ensemble de 688 personnes du NORD, parmi les 922 se trouvant dans les deux aires de base, sont en fait trouvées dans le bâtiment au jour de l'enquête. Un soin spécial ayant été mis durant l'enquête de couverture sur les questions concernant les adresses au jour du recensement et au jour de l'enquête, on considère ici que les adresses des personnes ayant déménagé sont de meilleure qualité dans les données de l'enquête que dans celles du recensement. Sur cette base, on déduit qu'au moins $151 + 688 = 839$ personnes ayant déménagé sur 1 512 sont énumérés dans le recensement à une adresse qu'ils n'avaient pas au jour officiel du relevé mais quelques temps après ce jour. Le décalage exact n'est pas connu car la date du déménagement n'a pas été relevée dans l'enquête.

7. Conclusion

Les défauts de couverture globaux du recensement de la population de l'an 2000 en Suisse sont dans l'ordre de grandeur des recensements dans d'autres pays. Des spécificités apparaissent cependant au niveau des sous-groupes (par exemple régions). Parmi les trois composantes, celle de la sous-couverture est fort intéressante car elle détecte non seulement des groupes de personnes plus ou moins bien recensés mais permet également d'analyser les erreurs de localisation et de mesures. Les estimations de sur-couverture sont de leur côté limitées par le manque d'informations complémentaires au recensement pour S_E. Elles pourraient être améliorées dans le futur par la collecte d'informations complémentaire sur les caractéristiques au jour du recensement lors d'une enquête auprès des personnes de cet échantillon (par exemple localisation et type de ménage). Les estimations de la sous-couverture nette sont basées sur plusieurs hypothèses. Les résultats dans de grands domaines semblent fiables mais certains risques, liés notamment au choix des cellules d'estimations, existent lorsque les domaines sont plus petits. Pour de futures estimations, on propose d'évaluer l'approche modèle tel qu'appliqué au Royaume-Uni au lieu des cellules d'estimations utilisées traditionnellement aux États-Unis.

Tableau 5
Comparaison de la localisation des personnes appariées. Les aires sont définies pour l'adresse au jour du recensement (selon information relevée dans l'enquête) et pour l'adresse au jour de l'enquête (également selon information relevée dans l'enquête). Présence dans l'aire de base, l'aire étendue (hors aire de base) ou hors de l'aire étendue pour les personnes n'ayant pas déménagé (fixes) et les personnes ayant déménagé (déménagements) entre le recensement et l'enquête

		Jour enquête				
		fixes	déménagements			
		base	base	étendue	hors étend.	total
Jour	base	46 689	922	69	277	1 268
recensement	étendue	258	42	4	3	49
	hors étendue	648	145	6	28	179
	manquant	0	15	1	0	16
	total	47 595	1 124	80	308	1 512

Un élément important à revoir pour de futures estimations est le choix de la population d'intérêt. Le choix de se limiter aux personnes dans des ménages privés et au domicile économique a provoqué quelques difficultés dans les estimations car une délimitation précise de cette population était difficile. Une future estimation pourrait exclure les ménages collectifs pour éviter les problèmes pratiques de relevé mais conserver tous les types de domiciles. L'ensemble des enregistrements au domicile économique serait alors traité comme un domaine.

L'estimation des défauts de couverture d'un recensement est un projet ambitieux qui a montré son intérêt. Les résultats donnent des informations sur la qualité des données du recensement 2000 et les différents problèmes de couverture. Les prochains recensements seront essentiellement basés sur des registres. Les estimations de couverture se baseront sur l'expérience acquise lors des estimations de 2000 avec de probables adaptations pour tenir compte du nouveau système de relevé.

Remerciements

Je tiens à remercier Philippe Eichenberger du Service de méthodes statistiques de l'Office fédéral de la statistique pour les discussions fructueuses durant tout le projet. Un grand merci également à Dr. Rajendra Singh et ses collègues du Decennial Statistical Studies Division du U.S. Census Bureau pour leur assistance lors du développement des méthodes et des estimations. Merci également à toutes les personnes du recensement qui ont effectués des travaux et fourni des informations pour le bon déroulement du projet, et à Paul-André Salamin du Service de méthodes statistiques pour la relecture attentive du papier.

Bibliographie

ABS (1997). The 1996 census of population and housing. Rapport annuel 1996-97, Australian Bureau of Statistics.

Brown, J.J., Diamond, I.D., Chambers, R.L., Buckner, L.J. et Teague, A.D. (1999). A methodological strategy for a one-number census in the UK, *Journal of the Royal Statistical Society*, Série A, 162(2), 247-267.

Fienberg, S.E. (1992). Bibliographie sur la modélisation à l'aide de la saisie-resaisie avec application au redressement des chiffres du recensement pour éliminer le sous-dénombrement. *Techniques d'enquête*, 18, 1, 157-169.

Hogan, H. (1993). The post enumeration survey: Operation and results. *Journal of the American Statistical Association*, 88(423), 1047-1060.

Hogan, H. (2003). L'évaluation de l'exactitude et de la couverture : théorie et conception. *Techniques d'enquête*, 29, 2, 145-156.

McLennan, W. (1997). Census of Population and Housing, Data Quality - Undercount. Autralia 1996. Article d'information, 2940.0. Australian Bureau of Statistics.

Renaud, A. (2001). Methodology of the Swiss Census 2000 Coverage Survey. *Proceedings of the Survey Research Methods Section* [CD-ROM], American Statistical Association.

Renaud, A. (2003). Estimation de la couverture du recensement de la population de l'an 2000. Échantillon pour l'estimation de la surcouverture (E-sample). Rapport de méthodes, 338-0019, Office fédéral de la statistique.

Renaud, A. (2004). Coverage estimation for the Swiss population census 2000. Estimation methodology and results. Rapport de méthodes, 338-0027, Office fédéral de la statistique.

Renaud, A., et Eichenberger, P. (2002). Estimation de la couverture du recensement de la population de l'an 2000. Procédure d'enquête et plan d'échantillonnage de l'enquête de couverture. Rapport de méthodes, 338-0009, Office fédéral de la statistique.

Renaud, A., et Potterat, J. (2004). Estimation de la couverture du recensement de la population de l'an 2000. Echantillon pour l'estimation de la sous-couverture (P-sample) et qualité du cadre de sondage des bâtiments. Rapport de méthodes, 338-0023, Office fédéral de la statistique.

Shao, J., et Tu, D. (1995). *The Jackknife and Bootstrap*. Springer Series in Statistics.

Statistique Canada (1999). Couverture. Rapport technique du recensement de 1996, 92-370-XIF, Statistique Canada.

Statistique Canada (2004). Couverture. Rapport technique du recensement de 2001, 92-394-XIF, Statistique Canada.

Trewin, D. (2003). Census of Population and Housing, Data Quality - Undercount. Autralia 2001. Article d'information, 2940.0. Australian Bureau of Statistics.

Wolter, K.M. (1986). Some coverage error models for census data. *Journal of the American Statistical Association*, 81(394), 338-346.