

ESTIMATION DE LA COUVERTURE DU RECENSEMENT DE LA POPULATION DE L'AN 2000 EN SUISSE

Anne Renaud¹

RÉSUMÉ

Les défauts de couverture sont estimés et analysés pour la première fois en Suisse dans le cadre du recensement de la population de l'an 2000. La composante de sous-couverture est estimée sur la base d'un échantillon stratifié à plusieurs niveaux et indépendant du recensement. La composante de sur-couverture est estimée sur la base d'un échantillon stratifié à deux niveaux tiré dans la liste 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.

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) et 2.2% au Royaume-Uni en 1991 (Brown et al, 1999). 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. On note une sous-couverture supérieure à 20% pour les jeunes hommes dans les centres urbains en 1991 au Royaume-Uni. Aux États-Unis en 2000, les Noirs ont une sous-couverture nette de 1.8% alors que les Blancs ont une sur-couverture de 1.1%.

Le recensement de la population de l'an 2000 donne une image de la population au 5 décembre 2000. La sous-couverture, la sur-couverture et la couverture nette résultante 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 un relevé (enquête de couverture), et du résultat d'un appariement avec les données du recensement (la personne de S_p a-t-elle 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, et du résultat d'une recherche de doubles et autres enregistrements erronés (l'enregistrement correspond-il à une vraie personne à recenser ?). 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 recombinaisonnés en suivant un modèle synthétique pour obtenir des résultats pour différents domaines de la population (Hogan, 2003).

La population d'intérêt pour les estimations de couverture est formée de la population au domicile économique et vivant dans un ménage privé. Un enregistrement est au domicile économique s'il se trouve à l'adresse où la personne habite en permanence, dans le cas où elle a un unique domicile, ou à l'adresse où elle passe le plus de temps, dans le cas où elle a deux domiciles (2.3% de la population). Les ménages collectifs tels les homes, les internats ou les prisons sont exclus pour des raisons pratiques de difficulté de relevé dans l'enquête complémentaire (enquête de couverture).

Cet article décrit les différentes étapes des estimations. La section 2 présente brièvement les échantillons S_p et S_E , ainsi que l'enquête de couverture. Les sections 3 et 4 exposent la méthode et les résultats des estimations des trois composantes de la couverture. La section 5 présente une conclusion générale sur le projet. Pour une information détaillée sur les méthodes et les résultats, voir Renaud (2004).

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

2. LES DEUX ÉCHANTILLONS ET L'ENQUÊTE

L'échantillon S_p , indépendant du recensement, est construit en deux parties : le canton du Tessin (TESSIN) et le reste de la Suisse (NORD) (Renaud, 2001 ; Renaud et Eichenberger, 2002). Le premier niveau consiste en la sélection de 303 unités primaires, communes pour le TESSIN et numéros postaux pour le NORD, selon un plan stratifié et un tirage proportionnel au nombre de bâtiments. Le deuxième niveau 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 niveau intermédiaire. L'échantillonnage est construit de manière à avantager les numéros postaux dans lesquels il est potentiellement plus facile de faire les listes de ménages et le relevé (p. ex. adresses postales complètes, peu de bâtiments codés comme inhabités) et à regrouper le travail sur le terrain, tout en limitant la variabilité des poids. Des listes exhaustives de ménages sont établies dans l'échantillon de environ 16'000 bâtiments, avec l'aide des employés postaux, avant de passer par un sous-échantillonnage de bâtiments permettant d'atteindre un total de environ 27'000 ménages. L'enquête de couverture consiste à contacter les ménages (téléphone ou face-à-face) et à relever les variables permettant un appariement avec le recensement et la définition de sous-groupes intéressants pour l'étude de la couverture (variables socio-démographiques, 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) avec une pondération dépendant de l'échantillonnage et d'un ajustement pour la non-réponse (Renaud et Poterat, 2002).

L'échantillon S_E est construit de façon à avoir des unités primaires identiques à celles de S_p . Un traitement spécial est effectué pour redistribuer dans les unités primaires du cadre de S_p les enregistrements qui n'en font pas partie (assignation de numéros postaux fictifs pour l'échantillonnage). Au deuxième niveau, on tire des enregistrements de la population d'intérêt selon un plan aléatoire simple, sans niveaux intermédiaires. L'allocation est choisie de façon à obtenir des poids fixes dans les strates d'échantillonnage des unités primaires. Au final, l'échantillon comporte $n_E=55'375$ enregistrements (Renaud, 2003).

3. SOUS-COVERTURE ET SUR-COVERTURE

3.1 Préliminaires : choix des statuts de match et de pertinence

Un appariement entre le S_p et le recensement est appliqué afin de déterminer le statut de match $0 \leq P_{m,j} \leq 1$ de chaque élément j de S_p . Le statut $P_{m,j}$ vaut 1 si l'élément est correctement 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 et les contrôles. Aucune interview complémentaire n'a lieu en plus de l'enquête de couverture. Une autre procédure est appliquée afin de déterminer le statut de pertinence $0 \leq P_{ce,j} \leq 1$ de chaque élément j de S_E . Le statut $P_{ce,j}$ vaut 1 si l'élément devait bien être énuméré dans le recensement 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 exactement. Cette procédure peut être plus ou moins complexe et sophistiquée. Dans notre cas, il s'agit d'une recherche dans le jeu de données complet du recensement de doublets ou de triplets des éléments de S_E , complétée par une analyse de cas suspects dans S_E . Aucune interview complémentaire n'a eu 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 de domicile et de ménage ?).

La détermination des statuts de match $P_{m,j}$ et de pertinence $P_{ce,j}$ doit être faite avec soin. Il importe de définir précisément ce que l'on accepte sous l'appellation appariement correct dans S_p , respectivement enregistrement correct dans S_E , et ce que l'on refuse. Différents critères peuvent être appliqués (Hogan, 2003). Ils dépendent des informations disponibles et des procédures d'appariement et de recherche des enregistrements erronés. On définit 3 statuts de match : le statut de match simple $P_{m,j}^{(s)}$ qui vaut 1 si un appariement est trouvé, quelques soient ses caractéristiques dans le recensement, le statut de match dans la bonne population $P_{m,j}^{(pop)}$ qui vaut 1 uniquement si l'appariement est dans la population d'intérêt, ou encore le statut de match dans la bonne localisation $P_{m,j}^{(loc)}$ qui vaut 1 uniquement si l'appariement est trouvé proche de l'adresse au jour du recensement. Nous définissons également le statut de pertinence simple $P_{ce,j}^{(s)}$ valant 0 si l'enregistrement est clairement hors recensement, 1/2 s'il s'agit d'un double, 1/3 s'il s'agit d'un triplet et 1 dans les autres cas (p. ex. si l'enregistrement est apparié à un élément de S_p). Aucun critère lié à l'appartenance à la population ou à la

localisation ne peut être appliqué car il n'y a pas de relevé complémentaire dans S_E . Nous définissons cependant le statut de pertinence doubles et triples dans la population $P^{(pop)}_{ce,j}$ qui tient compte de l'appartenance des doublets et triplets à la population. Ainsi, pour les cas multiples, $P^{(pop)}_{ce,j}=1/d'$, avec d' = nombre de doubles/triples dans la population d'intérêt.

3.2 Estimateur

Les taux de sous-couverture R_{sous} et de sur-couverture R_{sur} sont estimés par une moyenne pondérée R:

$$\hat{R} = 1 - \frac{\sum_{j \in S} w_j P_j}{\sum_{j \in S} w_j} = \frac{\sum_{j \in S} w_j (1 - P_j)}{\sum_{j \in S} w_j} \quad (1)$$

avec w_j le poids et P_j le statut de l'élément j de l'échantillon S . Pour la sous-couverture, on a $\hat{R} = \hat{R}_{sous}$, $w_j = w_{P,j}$, $P_j = P_{m,j}$ et $S = S_P$. Pour la sur-couverture, on a $\hat{R} = \hat{R}_{sur}$, $w_j = w_{E,j}$, $P_j = P_{ce,j}$ et $S = S_E$. Nous notons que le dénominateur de \hat{R}_{sur} est la somme des poids $w_{E,j}$ de S_E , et non pas le total connu dans le recensement C , afin d'avoir un estimateur potentiellement moins biaisé et de plus petite variance. L'estimation dans le domaine d du taux R_d est obtenue en remplaçant P_j par $P_j I_{jd}$ et w_j par $w_j I_{jd}$ avec $I_{jd} = 1$ si $j \in d$ et 0 sinon. La variance des estimateurs est estimée par un jackknife stratifié appliqué sur le premier niveau de l'échantillonnage, sans correction pour la population finie. On observe que l'estimation de la variance est plutôt conservatrice.

3.3 Résultats

La sous-couverture est de 1.6% (écart-type=0.11%, statut $P^{(s)}_{m,j}$). Un ensemble de 1.6% de la population manque donc dans le recensement. Ce taux passe à 1.7% (écart-type=0.11%) si l'on se restreint à un appariement dans la population d'intérêt (statut $P^{(pop)}_{m,j}$). Il atteint 3.1% (écart-type=0.22%) pour un appariement proche de la bonne adresse et dans la population d'intérêt (combinaison des statuts $P^{(pop)}_{m,j}$ et $P^{(loc)}_{m,j}$). Le taux varie entre les sous-groupes. Si l'on choisit le statut $P^{(s)}_{m,j}$, les personnes entre 20 et 31 ans ont un taux de 3.5% (écart-type=0.34%) et les étrangers avec des permis d'établissement temporaires ont un taux de 8.0% (écart-type=0.85%). Les étrangers avec des permis permanents (1.8%, écart-type=0.29%) sont proches des Suisses (1.3%, écart-type=0.09%). Des différences sont aussi notées entre les régions mais par exemple pas entre les sexes.

La sur-couverture est de 0.4% (écart-type=0.03%, statut $P^{(s)}_{ce,j}$). Un ensemble de 0.4% des enregistrements sont donc en trop dans les données du recensement. La restriction des doubles/triples aux éléments de la population a peu d'effet (0.35%, écart-type=0.03%, statut $P^{(pop)}_{ce,j}$). Le maximum est atteint pour les personnes entre 20 et 31 ans (1%, écart-type=0.09%). Les différences sont faibles entre les catégories des autres variables. Il ne semble donc pas y avoir de groupes très touchés par la sur-couverture. Les résultats sont cependant probablement un peu sous-évalués car nous n'avons pas d'information, hors la recherche des doubles, pour étudier le bien-fondé de l'existence de l'enregistrement dans les données du recensement (pas de source de données complémentaire car pas d'interviews).

La comparaison des informations relevées dans le recensement et l'enquête de couverture, pour les personnes appariées, indique qu'il y a peu de différences pour le sexe, l'âge, l'état civil et le permis d'établissement (0.7-1.6%) mais que les erreurs potentielles de mesure sont non négligeables pour la vie active (actif, non actif, chômage), la position dans le ménage et la taille du ménage (8-13%). On note aussi que 55% des personnes ayant déménagé entre le jour du recensement et celui du relevé de l'enquête de couverture sont trouvés, dans le recensement, proche de leur adresse au jour de l'enquête et non pas proche de celle au jour du recensement. Ce décalage est dû aux difficultés de suivi des questionnaires préimprimés du recensement lors des déménagements.

4. COUVERTURE NETTE

4.1 Méthode et estimateur

Le taux de sous-couverture nette est défini par $R_{sousnet} = 1 - R_{net}$, avec $R_{net} = C/N$ le taux de couverture nette, C le nombre recensé et N le vrai total dans la population. Le vrai total N est estimé par la somme des vrais totaux N_k dans des cellules d'estimations $k=1, \dots, K$ et chaque N_k est estimé sur la base du modèle dual (Wolter, 1986):

$$\hat{N}_k = [N_{1+,k}] \left[\frac{N_{+1,k}}{N_{11,k}} \right] = [C_k \hat{R}_{ce,k}] [\hat{R}_{m,k}^{-1}] = C_k [\hat{R}_{ce,k} \hat{R}_{m,k}^{-1}] = C_k \hat{F}_k \quad (2)$$

Le total dans le recensement $N_{1+,k}$ (capture) est estimé par le produit du total recensé C_k par le taux d'enregistrements corrects $\hat{R}_{ce,k} = 1 - \hat{R}_{sur,k}$ (estimé sur S_E) pour tenir compte de la sur-couverture. 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 entre l'enquête de couverture et le recensement $\hat{R}_{m,k} = 1 - \hat{R}_{sous,k}$ (estimé sur S_P) pour tenir compte de la sous-couverture. On définit également $\hat{F}_k = \hat{R}_{ce,k} \hat{R}_{m,k}^{-1}$ le facteur de correction de la couverture dans la cellule k . Le modèle dual a l'avantage de prendre en compte le fait que certaines personnes ne sont atteintes ni par le recensement (capture) ni par l'enquête de couverture (recapture). Pour éviter des biais, l'enquête de couverture et le recensement doivent être totalement indépendants. Les procédures d'appariement et de recherche des enregistrements erronés doivent être de bonne qualité. Le modèle doit être appliqué dans des cellules avec des personnes ayant la même probabilité d'être recensé dans le recensement, resp. dans l'enquête. La définition d'un appariement correct dans S_P et celle d'un enregistrement correct dans S_E doivent être identiques, i.e. équilibre entre sur- et sous-couverture (balancing). La population ne doit pas trop changer entre le jour du recensement et celui de l'enquête. 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 est donnée par $\hat{R}_{sousnet,d} = 1 - \hat{R}_{net,d} = 1 - C_d / \hat{N}_d$ avec C_d le nombre recensé dans le domaine et \hat{N}_d l'estimation du vrai total :

$$\hat{N}_d = \sum_{k=1}^K \hat{N}_{k,d} = \sum_{k=1}^K C_{k,d} \hat{F}_k \quad (3)$$

où $C_{k,d}$ est le nombre recensé 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'estimation du total \hat{N}_d est basée sur un modèle synthétique car on suppose que le facteur de correction est fixe dans chaque cellule $k=1, \dots, K$. Cette hypothèse est respectée si le comportement de tout sous-ensemble dans la cellule est identique à celui de la cellule entière. On reprend donc les cellules homogènes définies pour le modèle dual. La variance des estimateurs est estimée par un jackknife stratifié appliqué sur les unités primaires communes de S_P et S_E .

4.2 Choix des statuts et des cellules d'estimation

La sous-couverture nette est basée sur $\hat{R}_{ce,k}$ et $\hat{R}_{m,k}$, les moyennes pondérées des statuts $P_{ce,j}$ et $P_{m,j}$ dans les cellules. Les statuts doivent être choisis de manière à satisfaire l'équilibre entre la sur- et la sous-couverture. Par exemple, un appariement avec un élément hors population d'intérêt peut être refusé comme appariement correct ($P_{m,j}=0$) uniquement si la recherche des enregistrements corrects détecterait cet élément également comme incorrect car hors population ($P_{ce,j}=0$). Les informations sur S_E ne permettant pas de définir les enregistrements par erreur dans la population d'intérêt, il n'est pas non plus possible d'utiliser le critère de population pour le statut $P_{m,j}$. De même, le critère de localisation ne peut être utilisé car il n'y a pas d'information pour S_E . Nous choisissons donc le statut d'appariement simple $P(s)_{m,j}$ et le statut de pertinence $P^{(pop)}_{ce,j}$ qui tient compte de la population d'intérêt pour les enregistrements multiples uniquement pour éviter de considérer des doubles en trop.

Les cellules d'estimation 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 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 $y : P^{(s)}_{m,j}$). Les trois variables les plus influentes sont croisées (nationalité, état civil, tailles de commune) puis les autres variables sont intégrées successivement en faisant des regroupements lorsque les tailles de cellules sont trop petites (langue officielle de la commune, classe d'âge et sexe). Au final, nous obtenons 121 cellules.

4.3 Résultats

Le taux global de sous-couverture nette est égal à 1.4% (écart-type=0.12%). La sous-couverture de 1.6% est donc partiellement compensée par la sur-couverture de 0.4%. Le taux de sous-couverture nette des 20-31 ans (2.8%, écart-type=0.36%) et bien supérieur à celui des 60-79 ans (0.8%, écart-type=0.12%). De même les étrangers (2.9-3.5%, écart-type=0.32-0.39%) ont un taux supérieur aux Suisses (1.0%, écart-type=0.10%). On observe également des différences entre états civils, régions urbaines et rurales ou encore entre petites et grandes communes. Des effets de lissage sont observés dans les résultats. Les cellules d'estimations ne différencient par exemple pas les types de permis des étrangers (permanent et temporaire) car les tailles sont trop petites dans les échantillons S_P et S_E . Ces deux groupes ont cependant des comportements différents et la non prise en compte de ces spécificités dans les cellules lisse les résultats. Il est donc important de bien observer le lien entre cellules d'estimations et domaines d'estimations afin de rester critique sur les résultats détaillés.

5. CONCLUSION

Les défauts de couverture globaux du recensement de la population de l'an 2000 sont dans l'ordre de grandeur des recensements dans d'autres pays. Des spécificités apparaissent cependant au niveau des sous-groupes (par ex. 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 . 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'étudier l'approche modèle tel qu'appliqué au Royaume-Uni au lieu des cellules d'estimations utilisées traditionnellement aux États-Unis. 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 pour la préparation des prochains recensements, qu'ils se déroulent de façon similaire à 2000 ou basés principalement sur les registres. L'expérience acquise lors de ce projet permettra également d'améliorer les estimations futures.

RÉFÉRENCES

- Brown, J. J., Diamond, I. D., Chambers, R. L., Buckner, L. J. , Teague, A. D. (1999), « A methodological strategy for a one-number census in the UK », *J. R. Statist. Soc. A*, 162(2), p. 247-267.
- Fienberg, S. E. (1992), « Bibliography on capture-recapture modelling with application to census undercount adjustment », *Survey methodology*, 18(1), p. 143-154.
- Hogan, H. (1993), «The post enumeration survey : operation and results», *Journal of the American Statistical Association*, 88(423), p. 1047-1060.
- Hogan, H. (2003), «The Accuracy and Coverage Evaluation : Theory and design », *Survey Methodology*, 29(2) : p. 129-138.
- 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. 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. (2003), « Estimation de la couverture du recensement de la population de l'an 2000. Echantillon pour l'estimation de la sur-couverture (E-sample) », rapport de méthodes 338-0019, 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.
- 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.
- Wolter, K. M. (1986), «Some coverage error models for census data », *Journal of the American Statistical Association*, 81(394), p. 338-346.

Remerciements : je tiens à remercier Philippe Eichenberger du Service de méthodes statistiques de l'Office fédéral de la statistique, ainsi que Rajendra Singh et ses collègues du Decennial Statistical Studies Division du U.S. Census Bureau pour les nombreuses discussions méthodologiques durant le projet.