Méthodologie de l'enquête de couverture du nouveau recensement de la population en Suisse

Anne Massiani[1] & Lionel Qualité[2]

Office fédéral de la statistique, Espace de l'Europe 10, 2010 Neuchâtel, Suisse.
[1] *anne.massiani@bfs.admin.ch*
[2] *lionel.qualite@bfs.admin.ch*

Résumé. Le recensement suisse de la population a été remplacé par un système qui s'appuie sur des données administratives. L'Office fédéral de la statistique (OFS) souhaite en évaluer la qualité, ainsi que celle du registre des bâtiments et des logements. À cet effet, l'OFS a mené début 2013 une enquête de couverture basée sur des méthodes de type capture-recapture. L'objectif de cette présentation est d'en présenter les enjeux méthodologiques.

L'enquête porte sur un échantillon de zones géographiques. Les bâtiments, logements et habitants des zones sélectionnées sont énumérés. Les probabilités de sélection des zones dépendent de leur densité de population. De plus, l'échantillon est équilibré sur le nombre de bâtiments, de logements et de personnes se trouvant dans chaque zone selon les données de l'OFS.

Un des défis de la pondération est de corriger la non-réponse des personnes vivant dans les zones enquêtées alors qu'on ne sait pas combien d'unités contient réellement une zone donnée. Des méthodes de type capture-recapture sont également utilisées lors de cette étape. Les estimations obtenues à partir de l'enquête sont comparées aux données de l'OFS afin de calculer des taux de sur- et de sous-couverture.

L'exploitation des données n'est pas encore terminée et devrait s'achever fin juin 2014. Les premiers résultats, encore provisoires, laissent cependant penser que le nouveau système est de très bonne qualité et que les défauts de sous-couverture sont moins élevés qu'avec le recensement classique.

Mots-clés. Sous-couverture, sur-couverture, capture-recapture, plan équilibré, données administratives.

1 Introduction

L'Office fédéral de la statistique (OFS) est soucieux de la qualité des résultats qu'il publie, en particulier lorsqu'il s'agit de statistiques de base telles que les effectifs de la population. Certaines personnes peuvent ne pas être prises en compte dans la statistique alors qu'elles auraient dû l'être, ce qui engendre des défauts de sous-couverture. Inversement,

certaines personnes peuvent être comptées à tort ou à double, générant des défauts de sur-couverture. Une enquête de couverture (EC2000) avait été menée dans le cadre du dernier recensement classique réalisé en 2000 (cf. Renaud, 2007).

Depuis 2010, le recensement classique a été remplacé par un système qui s'appuie sur des données administratives et dont l'OFS souhaite évaluer la qualité. A cet effet, l'OFS a conduit au début de l'année 2013 une nouvelle enquête de couverture (EC2013) qui permet d'évaluer la qualité des résultats produits pour l'année 2012. Comme en 2000, la conception de l'EC2013 se base sur l'article fondateur de Wolter (1986) qui applique des méthodes de type capture-recapture à la mesure des erreurs de couverture. L'EC2013 a cependant un objectif supplémentaire par rapport à l'enquête EC2000 : elle doit aussi permettre de mesurer la couverture du registre des bâtiments et des logements. On dresse pour cela la liste de tous les bâtiments, logements et habitants d'une sélection de zones géographiques. Cette façon de procéder permet d'éviter un défaut de l'enquête menée en 2000 : les habitants des bâtiments qui ne faisaient alors pas partie de la base de sondage des bâtiments et des logements étaient exclus d'office de l'enquête de couverture. L'EC2013 est donc un peu plus ambitieuse et complexe que l'enquête EC2000.

Nous décrivons dans cet article la conception de l'EC2013 ainsi que les méthodes utilisées pour exploiter ses résultats. Les données permettant de calculer la sur-couverture n'étant pas encore analysées, nous n'abordons pour le moment que l'estimation de la sous-couverture. Il est prévu que l'exploitation des données soit complètement terminée fin juin 2014. Nous commençons par rappeler dans la section 2 quelques grands principes de l'article de Wolter (1986). Nous présentons dans la section 3 le plan d'échantillonnage de l'EC2013. Nous décrivons ensuite dans la section 4 le déroulement de l'enquête, puis exposons dans la section 5 les méthodes utilisées pour la pondération, avant de terminer par quelques mots de conclusion dans la section 6.

2 Fondements

Nous rappelons dans cette section l'estimateur que Wolter (1986) a proposé dans le cas simple du modèle de Petersen. On considère une population $\mathcal{U}$ de taille inconnue N. On dispose de deux énumérations incomplètes de $\mathcal{U}$:

- une liste A qui correspond au recensement, ou qui provient de données administratives, et dont on veut évaluer la qualité,
- une liste B établie lors de l'enquête de couverture. On suppose dans un premier temps que l'enquête de couverture est en fait une énumération menée sur tout le territoire.

La population $\mathcal{U}$ peut être décomposée selon l'appartenance à la liste A et la liste B. Les notations utilisées pour les différents groupes sont données dans le tableau 1.

Table 1: Effectifs des croisements entre la liste A et la liste B

		Liste B		
		appartient	n'appartient pas	total
Liste A	appartient	x_{11}	x_{12}	x_{1+}
	n'appartient pas	x_{21}	x_{22}	x_{2+}
	total	x_{+1}	x_{+2}	$x_{++} = N$

On suppose que les listes A et B ne comportent pas de sur-couverture, ou bien que celle-ci a déjà été traitée. On suppose de plus que pour tout individu i de la population $\mathcal{U}$, l'appartenance à la liste A et l'appartenance à la liste B sont deux variables aléatoires. On note p_{i1+} la probabilité que l'individu i appartienne à la liste A et p_{i+1} la probabilité qu'il appartienne à la liste B. Dans le modèle de Petersen, on fait les hypothèses supplémentaires suivantes :

- Indépendance. Pour tout individu i de la population $\mathcal{U}$, l'appartenance à la liste A est indépendant de l'appartenance à la liste B.
- Homogénéité. Tous les individus ont la même probabilité de faire partie de la liste A (respectivement B) et on note p_{1+} (respectivement p_{+1}) cette valeur commune, de sorte que pour tout i appartenant à $\mathcal{U}$, $p_{i1+} = p_{1+}$ et $p_{i+1} = p_{+1}$.

Wolter (1986) considère l'estimateur de N suivant :

$$\widehat{N} = x_{1+}\frac{x_{+1}}{x_{11}}, \tag{1}$$

qui est presque sans biais sous le modèle si la taille N de $\mathcal{U}$ est suffisamment grande.

En pratique, il n'est souvent pas possible de procéder à une énumération sur tout le territoire pour obtenir la liste B. On a donc sélectionné un échantillon de zones géographiques au sein desquelles on a mené l'enquête de couverture. Les quantités x_{+1} et x_{11} intervenant dans (1) sont remplacées par des estimations $\hat{x}_{+1}$ et $\hat{x}_{11}$. D'autre part, l'hypothèse d'homogénéité des probabilités de capture dans la liste A et de recapture lors de l'enquête sur tout le territoire est peu réaliste. On partitionne donc la population en K cellules d'estimation au sein desquelles l'hypothèse d'homogénéité est plus raisonnable et on considère l'estimateur :

$$\widetilde{N} = \sum_{k=1}^{K} x_{1+}^{k}\frac{\hat{x}_{+1}^{k}}{\hat{x}_{11}^{k}}, \tag{2}$$

où les x_{1+}^{k}, $\hat{x}_{+1}^{k}$, $\hat{x}_{11}^{k}$ sont les équivalents des x_{1+}, $\hat{x}_{+1}$, $\hat{x}_{11}$ au sein de la cellule k. Le choix des cellules d'estimation est délicat et doit se faire en respectant un équilibre entre deux exigences antagonistes : si les cellules sont trop grandes, on risque de s'écarter de l'hypothèse d'homogénéité. Si elles sont trop petites, la variabilité des estimateurs est importante.

3 Plan d'échantillonnage

Une base de sondage de zones géographiques a été constituée en réalisant un pavage du territoire suisse par des "carrés". Un échantillon de carrés est tiré au sein de cette base. Chaque carré constitue une grappe de bâtiments, logements et personnes, et correspond à une zone de travail pour un enquêteur. La charge de travail est différente selon la taille du carré, le nombre de personnes qui y habitent et le nombre de bâtiments qui s'y trouvent. Afin de limiter cette charge de travail, des carrés de taille différentes ont été définis en fonction de la densité de population et de bâti que l'on pense y trouver. Les carrés de base font 800 mètres de côté. Chaque carré qui contient plus de 250 personnes ou plus de 85 bâtiments selon les sources de l'OFS est découpé en quatre carrés de 400 mètres de côté. Puis si les limites sur le nombre de personnes ou de bâtiments sont encore dépassées, on continue à découper les carrés en quatre jusqu'à une taille minimale de 100 mètres de côté. Certains carrés de 800 mètres de côté sont complètement vides, dans le sens où ils ne comportent pas même un bâtiment d'après les données de l'OFS. Une partie de ces carrés vides est jugée inaccessible et est supprimée de la base de sondage. La base de sondage obtenue contient 102'851 carrés "classiques" et 9'411 carrés vides. Elle est divisée en deux strates : les carrés classiques et les carrés vides.

Les variables d'intérêt principales de l'enquête sont le nombre de personnes, de logements et de bâtiments relevés dans chaque carré. On peut imaginer que ces nombres sont relativement proches de ceux issus des données administratives de l'OFS. On a de ce fait utilisé pour la strate des carrés classiques un plan de sondage à probabilités inégales, favorisant la sélection des zones les plus peuplées. Pour limiter l'amplitude des probabilités d'inclusion, qui fait courir un risque sur la précision des résultats, ces probabilités d'inclusion ne sont pas proportionnelles aux tailles de population mais à leur racine carrée. Elles sont calculées de sorte que l'échantillon brut contienne environ 57'000 personnes. On sélectionne de plus 10 carrés vides selon un plan aléatoire simple.

L'utilisation d'un plan équilibré (voir Deville & Tillé, 2004) sur des variables qui sont de bons prédicteurs des variables d'intérêt doit permettre de réaliser d'importants gains de précision. Une des particularités de l'EC2013 est que le risque d'observer de la non-réponse au niveau de l'échantillon de zones géographiques est a priori nul. Dès lors l'emploi d'un plan de sondage équilibré est particulièrement intéressant puisque l'équilibrage n'est pas brisé par la non-réponse. Les variables utilisées pour l'équilibrage sont calculées à partir des données de l'OFS. Il s'agit du nombre de bâtiments, logements et personnes qui se trouvent dans chaque carré. On note s_G l'échantillon de 488 carrés classiques et 10 carrés vides qui a été sélectionné selon cette procédure.

On dresse la liste de tous les bâtiments, logements et personnes présents dans s_G le jour de l'enquête. On ne conserve cependant dans l'échantillon de personnes que celles qui appartiennent à la population cible dont on veut mesurer la couverture, c'est-à-dire la population qui a son domicile principal en Suisse et qui vit en ménage privé. Les ménages collectifs sont donc exclus de la population cible. Pour les étrangers, le type et la durée du

permis de séjour sont pris en compte. Une personne cible peut avoir plusieurs résidences et pourrait être échantillonnée à plusieurs endroits. Sa probabilité d'inclusion devrait en tenir compte. Afin de contourner cette difficulté, nous avons choisi d'échantillonner les personnes cibles uniquement à leur résidence principale. Autrement dit, si une personne déclare lors de l'interview être atteinte à sa résidence secondaire, elle ne sera pas retenue dans l'échantillon.

4 Déroulement de l'enquête

L'enquête est réalisée en trois phases. Dans un premier temps, les enquêteurs se rendent dans les zones sélectionnées et tentent d'accéder aux bâtiments. Ils contrôlent, complètent et modifient la liste des bâtiments et des logements établie par l'OFS d'après son registre. Observons que pour estimer la couverture du registre des bâtiments et des logements, la procédure d'enquête devrait être indépendante de ce registre (cf. section 2). Or, le fait de procéder en contrôlant des listes établies par l'OFS est une entorse à cette hypothèse d'indépendance. Ce choix a cependant été fait pour des raisons pratiques. Il a en effet été jugé trop complexe de demander aux enquêteurs d'établir une liste de bâtiments et de logements de suffisamment bonne qualité en ne partant de rien. Or, une qualité élevée est indispensable pour pouvoir déterminer par la suite si une unité relevée lors de l'enquête est présente ou non dans le registre. Cette opération d'appariement est nécessaire pour le calcul des quantités $\hat{x}_{11}^k$ définies à la section 2.

Dans la mesure du possible, les enquêteurs cherchent également lors de leur premier passage à distinguer les logements occupés par des ménages privés des logements collectifs, vacants ou occupés par des entreprises, quitte à se renseigner auprès des voisins ou du concierge. Si un ménage est présent et disponible, l'enquêteur profite de son passage pour dresser la liste des membres du ménage et réaliser une interview en face à face durant laquelle certaines caractéristiques des individus telles que le nom, le prénom, l'âge etc. sont relevées. Lorqu'un ménage refuse de répondre, l'enquêteur tente de le convaincre de participer à l'enquête de non-réponse, au cours de laquelle on ne relève que le nombre de membres du ménage.

Dans une deuxième phase de l'enquête, d'autres moyens de participation sont proposés aux ménages qui n'ont ni été interrogés ni refusé de participer lors du premier passage de l'enquêteur. L'entretien peut avoir lieu par téléphone si un numéro est disponible. Les ménages ont également la possibilité de répondre par internet. Dans une troisième phase de l'enquête, l'enquêteur se rend à nouveau sur le terrain pour tenter d'atteindre les ménages restés injoignables et éventuellement clarifier le cas de certains logements dont on ne sait pas s'ils sont occupés ou non.

Lors de toute la phase d'enquête au cours de laquelle on dresse la liste des personnes présentes dans les zones sélectionnées, les enquêteurs ne disposent d'aucune information, provenant de l'OFS, sur la population. Cette procédure d'enquête doit permettre de re-

specter au mieux l'hypothèse d'indépendance dans le cadre de l'estimation de la couverture du recensement de la population.

5 Pondération

La pondération des bâtiments et des logements est extrêmement simple car, sauf cas exceptionnel (logements dans des bâtiments auxquels l'enquêteur n'a pu accéder), il n'y a pas de non-réponse. Pour les personnes en revanche, la non-réponse est plus importante, estimée à 21.1%. Nous nous concentrons par conséquent dans cette section sur la pondération de l'échantillon des personnes. Notons s_P l'échantillon des personnes cibles répondantes, s_L l'échantillon de logements, s_L^{cont} le sous-ensemble de s_L que l'on a réussi à contacter et s_L^{int} ,le sous-ensemble de s_L^{cont} pour lequel on a réussi à réaliser une interview. Nous introduisons également la notion de logement cible : il s'agit d'un logement occupé le jour de l'enquête en tant que résidence principale par au moins une personne cible (cf. section 3). Nous définissons alors les variables suivantes :

$$O_\ell = \begin{cases} 1 \text{ si le logement } \ell \text{ est occupé par un ménage privé le jour de l'enquête,} \\ 0 \text{ sinon} \end{cases}$$

et

$$T_\ell = \begin{cases} 1 \text{ si le logement } \ell \text{ est un logement cible,} \\ 0 \text{ sinon.} \end{cases}$$

Dans la définition de O_ℓ, le logement est considéré comme occupé qu'il s'agisse d'une résidence principale ou secondaire. La variable T_ℓ n'est connue que si une interview a pu être réalisée dans le logement ℓ. En revanche, O_ℓ peut être connue même si le logement n'a pas pu être contacté. On considère en effet que le concierge ou les voisins peuvent nous renseigner de façon fiable sur le fait que le logement est occupé ou non, mais pas sur le fait qu'il s'agit d'une résidence principale ou d'une résidence secondaire.

On note d_g le poids de sondage de la zone géographique g dans s_G. Puisque l'on cherche à relever toutes les personnes cibles de s_G, les d_g sont également les poids de sondage des personnes de s_P. Ces poids doivent être corrigés pour la non-réponse. On désigne par p_i la probabilité que la personne i réponde, sachant qu'elle réside le jour de l'enquête dans une des zones de s_G, et par $\hat{p}_i$ une estimation de p_i. Le poids w_i d'une personne i de s_P est alors donné par $w_i = d_g/\hat{p}_i$. Afin d'estimer la probabilité p_i, on remarque que si la personne i occupe le logement ℓ, p_i peut se réécrire de la façon suivante :

$$\begin{aligned} p_i &= \mathrm{P}\left(\ell \in s_L^{int} \mid \ell \in s_L \text{ et } T_\ell = 1\right) && (3) \\ &= \underbrace{\mathrm{P}\left(\ell \in s_L^{int} \mid \ell \in s_L^{cont} \text{ et } T_\ell = 1\right)}_{p_\ell^1} \underbrace{P\left(\ell \in s_L^{cont} \mid T_\ell = 1\right)}_{p_\ell^2}. && (4) \end{aligned}$$

Nous décrivons ci-après la façon dont nous estimons p_ℓ^1 et p_ℓ^2.

Estimation de p_ℓ^1

Puisqu'en dehors des logements pour lesquels on dispose d'une interview, on ne sait pas quels sont les logements cibles parmi les logements contactés, on ne peut pas estimer p_ℓ^1 sans faire d'hypothèse supplémentaire. Pour estimer p_ℓ^1, on suppose que :

$$p_\ell^1 = \mathrm{P}\left(\ell \in s_L^{int} \mid \ell \in s_L^{cont} \text{ et } T_\ell = 1\right) = \mathrm{P}\left(\ell \in s_L^{int} \mid \ell \in s_L^{cont} \text{ et } O_\ell = 1\right). \tag{5}$$

On fait donc l'hypothèse qu'une fois le contact établi, la propension à répondre est la même que le logement soit une résidence principale ou secondaire. Nous utilisons alors l'ensemble des logements contactés pour modéliser la probabilité de réponse en fonction de variables telles que la grande région, la taille du carré qui est un proxy de la densité de population, la façon dont la personne a été contactée (internet, téléphone, face à face), ainsi que le nombre de personnes présentes dans le ménage, cette dernière information étant relevée lors de l'enquête de non-réponse.

Estimation de p_ℓ^2

On suppose que les probabilités de contact p_ℓ^2 sont homogènes au sein des grandes régions, indicées par gr, et on note p_{gr}^2 la valeur commune des p_ℓ^2 au sein de la grande région gr. On note n_{LCi}^{gr} le nombre de logements cibles dans $s_L \cap gr$, $n_{LCi,0}^{gr}$ le nombre de logements cibles que l'on n'a pas réussi à contacter, et $n_{LCi,1}^{gr}$ le nombre de ceux que l'on a réussi à contacter, de sorte que :

$$n_{LCi}^{gr} = n_{LCi,0}^{gr} + n_{LCi,1}^{gr} \quad \text{et} \quad p_{gr}^2 \approx \frac{n_{LCi,1}^{gr}}{n_{LCi,0}^{gr} + n_{LCi,1}^{gr}}.$$

Les quantités $n_{LCi,0}^{gr}$ et $n_{LCi,1}^{gr}$ sont inconnues puisque T_ℓ n'est connu que lorsqu'une interview a pu être effectuée. Nous devons les estimer. Or, p_ℓ^1 est la probabilité d'obtenir une réponse dans l'échantillon des logements cibles contactés. On peut donc estimer $n_{LCi,1}^{gr}$ par :

$$\hat{n}_{LCi,1}^{gr} = \sum_{\ell \in s_L^{int} \cap gr} \frac{1}{\hat{p}_\ell^1}. \tag{6}$$

Pour estimer $n_{LCi,0}^{gr}$, nous utilisons des méthodes de capture-recapture. Nous dressons la liste des logements qui sont, d'après nos données administratives, des logements cibles situés dans s_G. Cette liste joue le rôle de la liste A mentionnée dans la section 2. Pour chaque grande région, on peut estimer un facteur de correction en comparant cette liste à la liste des logements cibles relevés dans l'enquête, qui est elle aussi imparfaite, notamment à cause de la non-réponse. On applique alors le facteur de correction calculé pour la grande région gr au nombre de logements de la liste A que l'on n'a pas réussi à contacter

dans la grande région gr. On note $\hat{n}^{gr}_{LCi,0}$ les estimations qui résultent, puis on estime les probabilités de contact p^2_{gr} par :

$$\hat{p}^2_{gr} = \frac{\hat{n}^{gr}_{LCi,1}}{\hat{n}^{gr}_{LCi,0} + \hat{n}^{gr}_{LCi,1}}. \tag{7}$$

Les cellules d'estimation ont été créées grâce à un algorithme de segmentation. Les poids finaux $w_i = d_g/\hat{p}^1_\ell \hat{p}^2_{gr}$ permettent de calculer, pour chaque cellule k, les quantités $\hat{x}^k_{+1}$ et $\hat{x}^k_{11}$ qui interviennent dans la formule (2).

6 Conclusion

En 2000, le taux de sous-couverture du recensement était de 1.64% au niveau national (cf. Renaud, 2007). L'exploitation des données de l'EC2013 est encore en cours. Sur la base de données provisoires, à considérer avec précaution, nous nous attendons à ce que le nouveau système basé sur des données administratives soit de très bonne qualité et que ses défauts de sous-couverture soient moins élevés qu'en 2000.

L'exploitation des données doit être poursuivie et devrait être terminée fin juin 2014. Il reste notamment à traiter l'estimation de la sur-couverture, ainsi qu'à analyser les données relatives aux bâtiments et logements. Une autre grande étape est de calculer la variance des estimations. Contrairement à ce qui est souvent pratiqué pour les enquêtes de couverture, nous ne prévoyons pas de recourir à un bootstrap, mais de développer une formule basée sur des techniques de linéarisation ainsi que sur l'article de Deville & Tillé (2005), qui permet de calculer la variance pour des plan équilibrés.

Bibliographie

Deville, J.-C. & Tillé, Y. (2004). Efficient balanced sampling: The cube method. *Biometrika* **91**, 893–912.

Deville, J.-C. & Tillé, Y. (2005). Variance approximation under balanced sampling. *Journal of Statistical Planning and Inference* **128(2)**, 569–591.

Renaud, A. (2007). Estimation de la couverture du recensement de la population de l'an 2000 en Suisse : méthodes et résultats. *Techniques d'enquête* **33(2)**, 221–232.

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