00

Bases statistiques et généralités

338-0078

Estimation de la couverture du nouveau recensement en Suisse en 2012

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Département fédéral de l'intérieur DFI
Office fédéral de la statistique OFS

Neuchâtel 2017

Rapports de méthodes de la section Méthodes statistiques

Les rapports de méthodes décrivent les méthodes mathématiques et statistiques à la base des résultats et des analyses de la statistique publique. Ils présentent également l'évaluation et le développement de nouvelles méthodes en vue d'une application future. Ces publications visent d'une part à documenter les méthodes utilisées ou envisagées dans un souci de transparence et de rigueur scientifique, et d'autre part à favoriser la collaboration avec le monde scientifique et universitaire.

Les résultats numériques présentés dans les rapports de méthodes illustrent les concepts mathématiques décrits, mais ne sont pas des résultats officiels des enquêtes concernées. De même, les méthodes réellement appliquées peuvent différer légèrement de celles décrites dans ces rapports.

Les rapports de méthodes sont disponibles sous forme électronique sur le site internet de l'OFS.

Estimation de la couverture du nouveau recensement en Suisse en 2012

Rédaction Anne Massiani OFS
Éditeur Office fédéral de la statistique (OFS)

Neuchâtel 2017

Éditeur: Office fédéral de la statistique (OFS)

Renseignements: Anne Massiani, OFS, tél. 058 463 62 65, anne.massiani@bfs.admin.ch

Rédaction: Anne Massiani, OFS

Contenu: Section Méthodes statistiques, OFS

Série: Statistique de la Suisse

Domaine: 00 Bases statistiques et généralités

Langue du texte original: Français

Mise en page: Section DIAM, Prepress/Print
Ce document a été produit automatiquement à partir de banques de données. Il ne répond donc pas aux normes typographiques des publications de l'OFS.

Page de titre: OFS; concept: Netthoevel & Gaberthüel, Bienne; photo: © NorthShoreSurfPhotos – Fotolia.com

Impression: en Suisse

Commandes d'imprimés: Office fédéral de la statistique, CH-2010 Neuchâtel, tél. 058 463 60 60, fax 058 463 60 61, order@bfs.admin.ch

Prix: gratuit

Téléchargement: www.statistique.ch (gratuit)

Numéro OFS: 338-0078

ISBN: 978-3-303-00570-5

Table des matières

Préambule

L'Office fédéral de la statistique a réalisé début 2013 une enquête visant à évaluer la qualité de la couverture de la statistique de la population et des ménages, ainsi que de la statistique des bâtiments et des logements. Les travaux méthodologiques ont été réalisés par Anne Massiani de la section METH, tandis que l'organisation de l'enquête était confiée à la division BB. L'auteure de ce rapport souhaite remercier la section METH pour son accompagnement durant ce projet. En particulier, l'auteure souhaite adresser ses plus sincères remerciements à Lionel Qualité, dont l'aide et les compétences ont été décisives pour l'élaboration du plan d'échantillonnage, ainsi qu'à Daniel Kilchmann et Philippe Eichenberger pour leur soutien sans faille et leurs précieux conseils durant ce projet ambitieux. Egalement un immense merci à Jean-Pierre Renfer et Daniel Assoulin pour leur patiente relecture. Leurs remarques pertinentes ont permis d'améliorer la qualité de ce rapport.
Enfin, l'auteure tient à remercier très chaleureusement la division BB, particulièrement Markus Schwyn et Rachel Fritschi, pour avoir su créer un climat de travail constructif et surtout très agréable, ainsi que pour tous les échanges enrichissants que nous avons eus.

Résumé

Depuis 2010, 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 couverture, ainsi que celle du registre des bâtiments et des logements. A cet effet, l'OFS a mené début 2013 une enquête de couverture sur la base d'un échantillon de zones géographiques au sein desquelles les bâtiments, logements et habitants sont ré-énumérés. Les données relevées lors de l'enquête sont alors comparées aux données de l'OFS afin de calculer des taux de sur- et de sous-couverture pour les résultats relatifs à l'année 2012. Pour des raisons discutées dans le document, l'estimation de la couverture des décomptes de personnes se base sur des méthodes de type capture-recapture, tandis que dans le cas des bâtiments et des logements, les estimations sont totalement "design-based".
Ce document décrit les aspects méthodologiques de cette enquête complexe. Les thèmes couverts concernent notamment le plan d'échantillonnage, la pondération, les méthodes d'estimation utilisées, ainsi que les calculs de variance que nous avons développés afin d'accompagner nos estimations de mesures de précision.
Les résultats de l'enquête de couverture ont montré que le nouveau système est de très bonne qualité et que les défauts de sous-couverture sont moins élevés qu'avec l'ancien recensement de la population.

1 Introduction

Il est important pour l'Office fédéral de la statistique (OFS) d'évaluer la qualité des résultats qu'il publie, en particulier lorsqu'il s'agit de statistiques de base telles que les effectifs de la population, qui servent de référence pour les autres enquêtes sur 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. La méthodologie ainsi que les résultats de cette enquête sont décrits dans (cf. Renaud, 2007), ou de façon plus détaillée dans Renaud and Eichenberger (2002), Renaud (2003), Renaud (2004) et Renaud and Potterat (2004).

Depuis 2010, le recensement classique a été remplacé par un système qui s'appuie sur des données administratives. Ce système sert de base pour produire la statistique de la population et des ménages (STATPOP), 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, la date de référence exacte étant le 31 décembre 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 de la statistique des bâtiments et des logements (StatBL). Corollairement, le registre des bâtiments et des logements (RegBL) n'est pas utilisé comme base de sondage pour l'enquête de couverture de la population. On évite ainsi un défaut de l'enquête menée en 2000, où les bâtiments étaient tirés dans le registre des adresses de bâtiments (REAB). Ceci avait pour conséquence que les habitants des bâtiments qui ne faisaient pas partie de cette base de sondage étaient exclus d'office de l'enquête de couverture. Puisque l'on souhaite s'affranchir des registres et bases de sondage de l'OFS, nous sélectionnons un échantillon de zones géographiques au sein desquelles nous ré-énumérons les bâtiments, logements et habitants. De ce point de vue, les objectifs et les exigences sont élargis par rapport à l'EC2000, ce qui augmente le niveau de complexité de l'enquête, d'une part du point de vue méthodologique, mais aussi du point de vue de l'organisation sur le terrain.

Nous décrivons dans ce rapport la conception de l'EC2013 ainsi que les méthodes utilisées pour estimer ses résultats. Nous commençons par rappeler dans la section 2 quelques grands principes de l'article fondateur de Wolter (1986). Nous présentons dans la section 3 le plan d'échantillonnage de l'EC2013, puis nous décrivons dans la section 4 le déroulement de l'enquête. Les méthodes utilisées pour calculer les taux de couverture du registre des bâtiments-logements sont présentées dans la sections 5. En ce qui concerne l'enquête auprès des personnes, nous présentons la pondération, relativement complexe, dans la section 6, puis nous décrivons dans la section 7 les méthodes utilisées pour calculer les taux de couverture de la population. Des formules d'estimation de variance sont proposées pour chaque estimateur considéré. Elles ont été développées notamment grâce à des techniques de linéarisation dues à Deville (1999). Pour les bâtiments et logements, les formules d'estimation de variance sont présentées dans la section 5. Pour les personnes, la complexité des calculs mérite que la section 8 soit consacrée aux formules d'estimations de variance. Finalement, nous présentons les résultats obtenus dans la section 9, avant de terminer le rapport par les conclusions dans la section 10.

2 Fondements

Nous rappelons dans cette section le principe de la méthode duale présentée dans l'article fondateur de Wolter (1986), et sur laquelle se sont déjà appuyés plusieurs autres offices nationaux pour réaliser leurs enquêtes de couverture. Cela est par exemple le cas du U.S. Census bureau (Sanchez et al., 2012) ou de l'Office for National Statistics au Royaume-Uni (Office for national statistics, 2013). Plus précisément, nous rappelons (voir page 8) l'estimateur que Wolter (1986) a proposé dans le cas simple du modèle de Petersen, et sur lequel nous nous basons dans la suite de ce document.
On considère une population $\mathcal{U}$ de taille inconnue N. On dispose de deux énumérations incomplètes de $\mathcal{U}$:

– une liste A (capture) qui correspond au recensement, ou qui provient de données administratives, et dont on veut évaluer la sous-couverture.

– une liste B établie lors de l'enquête de couverture (recapture). Comme Wolter (1986), nous supposons dans un premier temps que l'enquête de couverture est en fait une énumération menée sur tout le territoire. Dans un second temps, nous discuterons des adaptations à effectuer lorsque l'enquête de couverture n'est réalisée que sur un échantillon du territoire.

On suppose que les liste A et B ne comportent pas de sur-couverture. Si cette hypothèse est réaliste pour la liste B provenant de l'enquête de couverture, elle ne l'est pas nécessairement pour la liste A. La composante de sur-couverture de la liste A doit donc être préalablement estimée. Nous décrivons dans les sections 5 et 7 les procédures utilisées pour estimer la composante de sur-couverture respectivement des décomptes de bâtiments-logements, et personnes. La population $\mathcal{U}$ peut être décomposée en fonction de l'appartenance à la liste A et de l'appartenance à la liste B. Les notations utilisées pour les différents cas sont données dans le tableau 1.

Tableau 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}	$x_{++} = N$

Les paramètres à estimer sont la taille N de la population, ainsi que le taux de sous-couverture x_{2+}/N de la liste A, qui peut aussi s'exprimer sous la forme :

$$\frac{x_{2+}}{N} = \frac{N - x_{1+}}{N} = 1 - F^{-} \tag{1}$$

où

$$F^{-} = \frac{x_{1+}}{N} \tag{2}$$

est un facteur de correction pour la sous-couverture de la liste A. On suppose 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 $p_{i,1+}$ la probabilité que l'individu i appartienne à la liste A et $p_{i,+1}$ la probabilité qu'il appartienne à la liste B. Le modèle de Petersen repose sur 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 $\mathcal{U}$, $p_{i,1+} = p_{1+}$ et $p_{i,+1} = p_{+1}$.

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

$$\widehat{N}_0 = x_{1+}\frac{x_{+1}}{x_{11}}. \tag{3}$$

L'estimateur (3) est presque sans biais sous le modèle si la taille N de $\mathcal{U}$ est suffisamment grande (cf. théorème 3.1 de Wolter, 1986) :

$$\mathbf{E}_m\left(\widehat{N}_0\right) \simeq N, \tag{4}$$

où $\mathbf{E}_m$ désigne l'espérance sous le modèle de Petersen. On estime alors le taux de sous-couverture de la liste A par $(1 - f_0^-)$, où :

$$f_0^- = \frac{x_{1+}}{\widehat{N}_0} = \frac{x_{11}}{x_{+1}}. \tag{5}$$

En pratique, il n'est souvent pas possible de procéder à une énumération sur tout le territoire pour obtenir la liste B. On sélectionne donc un échantillon de zones géographiques au sein desquelles on mène l'enquête de couverture. Wolter (1986) traite l'exemple d'un sondage aléatoire simple de zones géographiques, mais indique que la méthode se généralise facilement à des plans plus complexes. Les quantités x_{+1} et x_{11} intervenant dans (3) et (5) sont remplacées par des estimations $\hat{x}_{+1}$ et $\hat{x}_{11}$. D'autre part, l'hypothèse d'homogénéité sur tout le territoire des probabilités de capture dans la liste A et de recapture dans la liste B est peu réaliste. On partitionne donc la population en K cellules d'estimation qui résultent généralement du croisement de variables sociodémographiques et au sein desquelles l'hypothèse d'homogénéité est plus raisonnable. On considère alors l'estimateur :

$$\widehat{N} = \sum_{k=1}^{K} x_{1+}^k \frac{\hat{x}_{+1}^k}{\hat{x}_{11}^k}, \tag{6}$$

où l'exposant dans les notations x_{ab}^k et $\hat{x}_{ab}^k$, avec a, b appartenant à $\{1, 2, +\}$, est utilisé pour signaler qu'il s'agit des équivalents de x_{ab} et de $\hat{x}_{ab}$ au sein d'une sous-population particulière, ici la sous-population contenue dans la cellule d'estimation k. Si l'aléa du modèle de Pertersen et l'aléa du plan de sondage sont indépendants, comme cela est le cas si des zones géographiques correspondant à des grappes d'unités sont sélectionnées selon un plan aléatoire simple, l'estimateur $\widehat{N}$ est approximativement sans biais lorsque l'on combine l'aléa du modèle et l'aléa du plan de sondage. En effet, en désignant par $\mathbf{E}_p$ l'espérance sous le plan de sondage et par $\mathbf{E}$ l'espérance sous les deux sources d'aléa, on a :

$$\mathbf{E}\left(\widehat{N}\right) = \mathbf{E}_m\mathbf{E}_p\left(\widehat{N}\right) = \sum_{k=1}^{K} \mathbf{E}_m\mathbf{E}_p\left(x_{1+}^k \frac{\hat{x}_{+1}^k}{\hat{x}_{11}^k}\right) \tag{7}$$

$$\simeq \sum_{k=1}^{K} \mathbf{E}_m\left(x_{1+}^k \frac{x_{+1}^k}{x_{11}^k}\right), \tag{8}$$

la dernière approximation reposant sur le fait qu'au sein de chaque cellule d'estimation k, l'espérance d'un ratio peut raisonnablement être approximée par le ratio des espérances, si les cellules d'estimation ne sont pas trop petites. En notant N_k la taille de la population dans la cellule k, il découle de (3), (4) et (8) que :

$$E\left(\widehat{N}\right) \simeq \sum_{k=1}^{K} N_k = N. \tag{9}$$

On estime alors le taux de sous-couverture de la liste A par $(1 - f^-)$ où :

$$f^- = \frac{x_{1+}}{\widehat{N}}. \tag{10}$$

Nous montrons dans l'annexe A que la formule (9) reste valable pour un plan plus complexe qu'un plan simple de zones géographiques, comme celui que nous utilisons (cf. section 3 pour une description du plan de sondage).
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 et l'approximation (8) n'est par ailleurs plus très fiable.
L'intérêt de la modélisation proposée par Wolter (1986) réside dans le fait qu'il n'est pas nécessaire de faire l'hypothèse que la liste B soit complète pour pouvoir estimer la sous-couverture de la liste A. Comme le montre le tableau 1, la modélisation proposée par Wolter (1986) tient compte du fait que l'enquête de couverture peut elle aussi être touchée par des défauts de sous-couverture, puisqu'il y a x_{22} individus qui ne font partie ni de la liste A ni de la liste B. Cependant, sous les hypothèses d'homogénéité et d'indépendance du modèle de Petersen, l'estimateur f^- défini par (10) est un estimateur approxitivement sans biais (sous l'aléa du modèle et du plan) de la sous-couverture de la liste A. Si on a de bonnes raisons de penser que la ré-énumération de l'enquête de couverture correspond parfaitement à la réalité, on peut se placer dans une optique totalement "design-based" dans laquelle les hypothèses du modèle de Petersen ne sont plus nécessaires. Dans ce cas, on peut utiliser l'estimateur $(1 - f_{db}^-)$, où :

$$f_{db}^- = \frac{\hat{x}_{1+}}{\hat{x}_{+1}} \tag{11}$$

est approximativement sans biais sous le plan, puisque :

$$\mathbf{E}_p(\hat{x}_{1+}) = x_{1+} \tag{12}$$

et

$$\mathbf{E}_p(\hat{x}_{+1}) = x_{+1} = x_{+1} + \underbrace{x_{+2}}_{0} = N, \tag{13}$$

ce qui conduit à

$$\mathbf{E}_p(f_{db}^-) \simeq F^-. \tag{14}$$

Observons que si la ré-énumération de l'enquête de couverture correspond parfaitement à la réalité, l'estimateur (11) donne la même valeur que l'estimateur (10) calculé sur la base d'une seule cellule d'estimation. On a en effet dans ce cas $\hat{x}_{12} = 0$, c'est-à-dire $\hat{x}_{1+} = \hat{x}_{11}$. Il en découle que :

$$f_{bd}^- = \frac{\hat{x}_{11}}{\hat{x}_{+1}} = f^-.$$

Pour des raisons qui seront explicitées à la section 4, nous utilisons pour estimer la couverture des personnes l'estimateur dual de Wolter, cf. formules (6) et (10), tandis que pour les bâtiments et logements, nous nous plaçons sous une optique design-based et utilisons l'estimateur (11).

3 Plan d'échantillonnage

Population cible

En ce qui concerne les personnes, la population cible est la population résidente permanente en ménage privé au domicile principal. Pour les bâtiments, il s'agit de l'ensemble des bâtiments d'habitation. Il n'y a pas de restriction pour les logements : tous les logements font partie de la population cible, qu'ils soient vides, occupés en tant que résidence principale ou en tant que résidence secondaire. Que l'on s'intéresse aux personnes, bâtiments ou logements, la date de référence est le 31 décembre 2012.

Base de sondage

Une base de sondage de zones géographiques a été constituée en réalisant un pavage du territoire suisse par des "carrés". 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 de bâtiments qui s'y trouvent. Afin que le travail de repérage sur le terrain (cf. section 4) pour un carré donné puisse être confié à un seul et même enquêteur, la charge de travail correspondante doit être limitée et raisonnable. Dans la mesure du possible, on souhaite également uniformiser la charge de travail d'un carré à un autre. A cet effet, des carrés de taille différentes ont été définis en fonction de la densité de population et de bâti que l'on pense y trouver d'après les données administratives disponibles lors du tirage de l'échantillon, c'est-à-dire en mai 2012. Les carrés de base font 800 mètres de côté. Chaque carré qui contient plus de 250 personnes [1] ou plus de 85 bâtiments [2] 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. L'intérêt d'imposer des bornes sur la densité des carrés est également de limiter l'éventuel effet de grappe, qui découle du fait que les carrés correspondent à des grappes de bâtiments, logements et personnes. Puisque cet effet négatif sur la précision dépend notamment de la densité des carrés, imposer une limite sur leur densité semble pertinent de ce point de vue également.

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, par exemple dans les Alpes mais pas uniquement, et est supprimée de la base de sondage. La base de sondage obtenue est divisée en deux strates : une strate de 102'851 carrés "classiques" et une strate de 9'411 carrés vides de 800 mètres de côté [3].

1. Toutes catégories de population confondues.
2. Tous types de bâtiments confondus.
3. Suite au découpage des carrés, certains carrés de moins de 800 mètres de côté peuvent ne contenir aucun bâtiment, mais sont malgré tout attribués à la strate des carrés classiques. Ce choix est essentiellement motivé par des raisons pratiques. Pour des questions d'organisation interne, le tirage au sein des carrés vides ne pouvait être réalisé aussi rapidement que celui des carrés classiques. Or il était nécessaire de livrer rapidement à l'institut chargé de l'enquête sur le terrain un échantillon le plus "complet" possible. Par ailleurs, les carrés vides de 800 mètres de côté se distinguent des autres carrés vides du point de vue du travail sur le terrain, et font l'objet d'un repérage spécifique. Ils sont en effet souvent relativement loin de toute zone habitée et difficiles d'accès, tandis que les autres carrés vides obtenus suite au découpage d'un carré de 800 mètres de côté trop dense sont par principe proche d'une zone habitée et dense.

Echantillonnage

Un échantillon de carrés est tiré au sein de chaque strate. Dans la strate des carrés vides de 800 mètres de côté, on sélectionne 10 carrés selon un plan aléatoire simple. Pour la strate des carrés classiques, nous avons utilisé un plan de sondage à probabilités inégales, de façon à favoriser la sélection des zones les plus peuplées. Cette façon de procéder permet d'atteindre le nombre visé de personnes tout en limitant le nombre de carrés à enquêter. On limite ainsi les coûts liés aux déplacements des enquêteurs d'une zone à une autre, ce qui permet de respecter le budget disponible. Ce choix de probabilités d'inclusion est également motivé par le fait que sous certaines conditions, l'estimateur design-based défini par (11) et utilisé pour estimer le taux de sous-couverture des bâtiments et logements a une variance nulle sous le plan dans la strate des carrés classiques. Les hypothèses sont les suivantes :

(H1) il n'y a pas de non-réponse, ni au niveau des zones géographiques ni au niveau des unités (cas des bâtiments et des logements, cf. sections 5),

(H2) dans chaque zone géographique g, le nombre x^g_{+1} d'unités ré-énumérées par l'enquête de couverture dépend de la densité de population de la zone. Il en va de même pour le nombre x^g_{1+} d'unités présentes d'après nos données, dont on a préalablement éliminé ou nettoyé la composante de sur-couverture. Plus précisément, on suppose que x^g_{+1} et x^g_{1+} sont proportionnels au nombre de personnes cibles contenues dans la zone g selon nos données administratives.

(H3) La probabilité de sélection π_g de chaque zone g est proportionelle au nombre de personnes cibles contenues dans la zone g selon nos données administratives.

Sous l'hypothèse (H1), les estimateurs $\hat{x}_{+1}$ et $\hat{x}_{1+}$ apparaissant dans la formule (11) sont égaux à :

$$\hat{x}_{+1} = \sum_{g \in s_G} \frac{1}{\pi_g} x^g_{+1} \tag{15}$$

et

$$\hat{x}_{1+} = \sum_{g \in s_G} \frac{1}{\pi_g} x^g_{1+}, \tag{16}$$

où s_G désigne l'échantillon de carrés "classiques". Sous les hypothèses (H2) et (H3), on peut montrer que ces deux estimateurs ont une variance nulle sous le plan. Cela est donc également le cas de l'estimateur (11). Cependant, en pratique, un modèle n'est jamais totalement parfait et on peut plus ou moins s'éloigner des hypothèses correspondantes. Pour limiter l'amplitude des probabilités d'inclusion, qui fait courir un risque sur la précision des résultats lorsque l'on s'écarte des hypothèses, les probabilités d'inclusion ne sont pas proportionnelles aux nombres de personnes cibles contenues dans chaque zone mais à leurs racines carrées. Si un carré ne contient aucune personne cible selon nos données administratives (mais fait bien partie de la strate des carrés classiques), sa mesure de taille est fixée à 0.5 et non pas 0, de sorte que le carré ait une probabilité d'inclusion non nulle. Le nombre de zones sélectionnées est de 488, de façon à disposer d'un échantillon brut d'environ 57'000 personnes et d'ainsi respecter le budget disponible.
Le tirage est réalisé selon un plan équilibré (voir Deville and Tillé, 2004) sur des variables que l'on pense être corrélées aux quantités x^g_{1+} et x^g_{+1}, telles que le nombre de bâtiments, logements et personnes de la population cible qui se trouvent dans chaque carré selon les

données de l'OFS. Cela devrait permettre un gain de précision pour les estimateurs $\hat{x}_{+1}$ et $\hat{x}_{1+}$, cf. formules (15) et (16). 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 rompu par la non-réponse.
Quelques statistiques descriptives sur la dispersion des poids dans la strate des carrées classiques sont données dans le tableau 2, dans lequel Q25 et Q75 désignent respectivement le premier et le troisième quartile. La figure 1 représente par ailleurs les boxplots des poids en fonction de la taille des carrés, dans la strate des carrés classiques. On constate que bien que les probabilités d'inclusion soient proportionnelles à la racine carrée des nombres de personnes cibles contenues dans chaque zone plutôt qu'à ces nombres eux-mêmes, certains carrés très peu peuplés ont des poids relativement élevés.

Tableau 2 Dispersion des poids dans la strate des carrés classiques

Min	Q25	Médiane	Q75	Max
76.4	124.1	154.0	205.4	3155.0

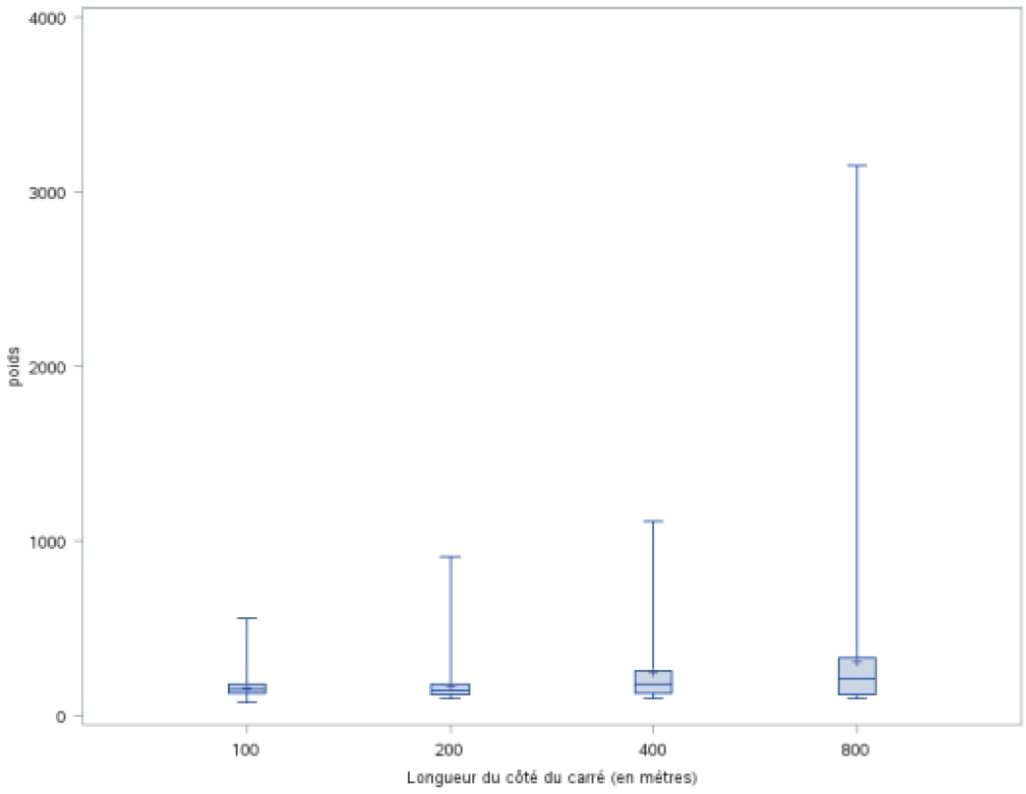

FIGURE 1 Boxplots des poids en fonction de la taille des carrés dans la strate des carrés classiques

On dresse la liste de tous les bâtiments, logements et personnes présents dans l'échantillon de zones géographiques s_G le jour de l'enquête. Cela implique qu'en cas de déménagement entre la date de référence et le jour de l'enquête, la personne est attribuée au carré dans lequel elle réside le jour de l'enquête. On ne conserve ensuite dans l'échantillon que les unités qui appartiennent, le jour de référence, à la population cible dont on veut mesurer la couverture. Par exemple, un bébé né après le 31 décembre 2012 sera relevé par l'enquêteur mais ne sera pas

retenu dans l'échantillon. 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 se trouver dans sa résidence secondaire, elle ne sera pas retenue dans l'échantillon [4].

4 Déroulement de l'enquête

Une lettre est tout d'abord envoyée aux adresses des zones sélectionnées dans la strate des carrés classiques afin d'informer la population concernée sur la tenue de l'enquête. Un code d'accès valabe tout au long de l'enquête est également fourni dans la lettre afin que les ménages puissent prendre part à l'enquête par internet. Sur le terrain, l'enquête est réalisée en trois phases. Dans une première phase, 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. Si un ménage est présent et disponible, et s'il n'a pas déjà participé par internet, 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. Les enquêteurs cherchent également à repérer lors de leur premier passage les logements occupés par des ménages collectifs, vacants ou occupés par des entreprises, et pour lesquels il est inutile de chercher à réaliser une interview.
Dans une deuxième phase de l'enquête, lorsqu'un numéro de téléphone fixe est disponible, les enquêteurs tentent d'interroger par téléphone les ménages qui n'ont ni été interrogés ni refusé de participer lors du premier passage de l'enquêteur.
Enfin, dans la 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 par un ménage privé.
En ce qui concerne les dix zones vides de 800 mètres de côté, les enquêteurs sillonnent le terrain à l'aide de cartes et de GPS.
Dans le cas des bâtiments et des logements, le fait de procéder en contrôlant des listes établies par l'OFS est une entorse à l'hypothèse d'indépendance sous-jacente au modèle de Pertersen (cf. section 2). 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 $\hat{x}_{11}$ définies à la section 2. En revanche, lorsque les enquêteurs dressent la liste des personnes présentes dans un logement, ils ne disposent d'aucune information provenant de l'OFS sur la population. Cette procédure d'enquête doit permettre de respecter au mieux l'hypothèse d'indépendance dans le cadre de l'estimation de la couverture du recensement de la population. Comme nous l'avons explicité dans la section 2, un des intérêts du modèle de Petersen est

4. Dans moins d'une dizaine de cas, la variable qui permet de savoir si la personne a été atteinte à son domicile principal était manquante. Nous avons conservé ces personnes dans l'échantillon et avons supposé que :
   – chaque personne concernée a exactement deux résidences en Suisse,
   – si la personne avait été atteinte à son autre résidence, la variable aurait également été manquante.

Afin de tenir compte du fait que ces personnes peuvent être échantillonnées à deux endroits, nous avons ensuite réalisé un partage des poids qui consiste à diviser le poids de ces personnes par 2.

que, sous les hypothèses d'indépendance et d'homogénéité, il n'est pas nécessaire de supposer que l'enquête de couverture est "parfaite", dans le sens où elle peut être elle-même touchée par des défauts de sous-couverture. Une estimation "fiable" de la sous-couverture du recensement peut malgré tout être obtenue si les hypothèses du modèle sont respectées. Dans le cas de l'estimation de la sous-couverture des personnes, il nous semble crucial de respecter sur le terrain l'hypothèse d'indépendance aussi bien que possible, parce qu'il n'est pas réaliste de supposer que personne n'est "oublié" par l'enquête de couverture. Dans le cas des bâtiments et des logements, en revanche, il semble raisonnable d'admettre qu'aucune unité n'est "oubliée" par l'enquête. En effet, les bâtiments et les logements ne sont pas mobiles comme le sont les personnes. Ils sont facilement repérables par un enquêteur sérieusement formé et ayant un grand nombre d'outils à sa disposition. On peut donc supposer que la liste des bâtiments et des logements fournie par les enquêteurs correspond parfaitement à la réalité, et utiliser l'estimateur design-based défini par (11) pour estimer le taux de sous-couverture.

5 Procédure d'estimation pour les bâtiments et les logements

Définition des quantités à estimer pour les bâtiments

Contrairement à ce qui avait été supposé lors de la conception de l'enquête, la majorité des bâtiments et des logements absents de nos données ont été découverts dans la strate des zones vides de 800 mètres de côté et non pas dans la strate des carrés classiques. En effet, parmi les dix zones sélectionnées dans la strate des zones vides de 800 mètres de côté, six se sont avérées effectivement vides de tout bâtiment d'habitation, mais dans quatre d'entre elles, 23 bâtiments ont été découverts, contre 15 bâtiments en sous-couverture sur l'ensemble des 488 zones sélectionnées dans la strate des carrés classiques. Les 23 bâtiments découverts dans la strate des zones vides de 800 mètres de côté contiennent 20 logements [5] mais aucune personne cible (pas de domicile principal). La conjonction de la faible taille de l'échantillon de zones vides de 800 mètres de côté et de ces observations inattendues pose un problème de précision des estimations et ne permet pas de conclusions claires sur la qualité de nos données administratives dans la strate des zones vides de 800 mètres de côté. En raison du manque de précision des estimations, il a été décidé de ne publier les estimations de la couverture des décomptes de bâtiments et de logements que pour la strate des carrés classiques. En accord avec les notations introduites à la section 2 :

N désigne ici le "vrai" nombre de bâtiments dans la strate des carrés classiques.

x_{1+} désigne ici le nombre de bâtiments dans la strate des carrés classiques d'après nos données administratives, mais nettoyé de la sur-couverture.

Nous désignons de plus par C le nombre de bâtiments dans la strate des carrés classiques d'après nos données administratives, y compris la sur-couverture. Les quantités à estimer sont les suivantes.

– Le taux de sous-couverture $(1 - F^{-})$, où :

$$F^{-} = \frac{x_{1+}}{N}. \tag{17}$$

5. Il peut y avoir moins de logements que de bâtiments d'habitation car un bâtiment d'habitation peut ne contenir aucun logement, par exemple dans le cas d'un refuge de montagne qui ne contient qu'un dortoir.

– Le taux de sur-couverture $(1 - F^+)$, où :

$$F^+ = \frac{x_{1+}}{C}. \tag{18}$$

– Le taux de sous-couverture nette $(1-F)$, qui reflète le bilan entre la sous- et la sur-couverture, et est défini par :

$$F = \frac{F^-}{F^+} = \frac{C}{N}. \tag{19}$$

Estimateurs considérés pour les bâtiments

Nous rappelons que pour les raisons discutées à la section 4, nous nous plaçons dans le cas des bâtiments et des logements sous une optique totalement design-based et utilisons l'estimateur f_{db}^- défini par (11) pour estimer la sous-couverture. Il n'y a pas eu de non-réponse au niveau des bâtiments. Par conséquent, les quantités $\hat{x}_{1+}$ et $\hat{x}_{+1}$ intervenant dans (11) sont respectivement définies par les équations (15) et (16) rappellées ci-dessous :

$$\hat{x}_{1+} = \sum_{g \in s_G} \frac{1}{\pi_g} x_{1+}^g, \tag{20}$$

$$\hat{x}_{+1} = \sum_{g \in s_G} \frac{1}{\pi_g} x_{+1}^g. \tag{21}$$

On pose de plus :

$$\widehat{C} = \sum_{g \in s_G} \frac{1}{\pi_g} C_g, \tag{22}$$

où C_g désigne le nombre de bâtiments présents dans la zone g d'après nos données administratives, y compris la sur-couverture. Le taux de sur-couverture est estimé par $(1 - f_{db}^+)$, où :

$$f_{db}^+ = \frac{\hat{x}_{1+}}{\widehat{C}}. \tag{23}$$

Le taux de sous-couverture nette est estimé par $(1 - f_{db})$, où :

$$f_{db} = \frac{f_{db}^-}{f_{db}^+} = \frac{\widehat{C}}{\hat{x}_{+1}}. \tag{24}$$

Variance des estimateurs utilisés pour les bâtiments

Pour estimer la variance sous le plan de l'estimateur f_{db}^- défini par (11) et utilisé pour estimer le taux de sous-couverture des bâtiments, nous faisons tout d'abord appel à des techniques de linéarisation (cf. Deville, 1999). Dans le cas de l'estimateur d'un ratio comme l'estimateur f_{db}^-, cela permet d'obtenir l'approximation suivante de la variance (cf. exemple 3 de Deville, 1999) :

$$\mathbf{Var}_p\left(f_{db}^-\right) \simeq \mathbf{Var}_p\left(\sum_{g \in s_G} \frac{1}{\pi_g} \ell_g^-\right), \tag{25}$$

où $\mathbf{Var}_p$ désigne la variance sous le plan et où

$$\ell_g^- = \frac{1}{N}\, x_{1+}^g - \frac{x_{1+}}{N^2}\, x_{+1}^g. \tag{26}$$

La formule (26) fait intervenir les quantités x_{1+} et N qui, en pratique, sont inconnues. Lors de l'estimation de variance, ces quantités sont remplacées par leurs estimations respectives :

$$\hat{\ell}_g^- = \frac{1}{\hat{x}_{+1}}\, x_{1+}^g - \frac{\hat{x}_{1+}}{\hat{x}_{+1}^2}\, x_{+1}^g. \tag{27}$$

La formule (25) montre que le problème est ramené à l'estimation de la variance de l'estimateur de Horvitz-Thompson d'un total sous le plan de sondage de l'échantillon de zones géographiques. Deville and Tillé (2005) ont proposé plusieurs estimateurs de la variance de l'estimateur de Horvitz-Thompson d'un total sous un plan équilibré. En appliquant la formule la plus simple correspondant selon leurs notations au choix $c_g = (1 - \pi_g)$, nous obtenons l'estimateur suivant :

$$\widehat{\mathbf{Var}}_p\left(f_{db}^-\right) = \sum_{g \in s_G} c_g e_g^2, \tag{28}$$

où e_g est le résidu de la régression pondérée des $\hat{\ell}_g^-/\pi_g$ sur les variables d'équilibrage divisées par π_g, le facteur de pondération utilisé étant égal à c_g. Pour estimer la variance des estimateurs (23) et (24) utilisés pour évaluer la sur-couverture et la sous-couverture nette, nous nous basons sur les mêmes principes que pour la sous-couverture. La linéarisée de f_{db}^+ est donnée par :

$$\ell_g^+ = \frac{1}{C}\, x_{1+}^g - \frac{x_{1+}}{C^2}\, C_g \tag{29}$$

et est remplacée en pratique par :

$$\hat{\ell}_g^+ = \frac{1}{\widehat{C}}\, x_{1+}^g - \frac{\hat{x}_{1+}}{\widehat{C}^2}\, C_g. \tag{30}$$

La linéarisée de f_{db} est donnée par :

$$\ell_g = \frac{1}{N}\, C_g - \frac{C}{N^2}\, x_{+1}^g \tag{31}$$

et est remplacée en pratique par :

$$\hat{\ell}_g = \frac{1}{\hat{x}_{+1}}\, C_g - \frac{\widehat{C}}{\hat{x}_{+1}^2}\, x_{+1}^g. \tag{32}$$

Procédure d'estimation pour les logements

La définition des estimateurs de la couverture, ainsi que le calcul de leur variance, se basent sur les mêmes principes que ceux utilisés pour les bâtiments. La seule différence provient du fait qu'il y a pour les logements des cas très exceptionnels de non-réponse. Les logements de certains bâtiments dans lesquels l'enquêteur n'a pas pu accéder n'ont en effet pas pu être contrôlés en intégralité. Cela concerne seulement 25 bâtiments parmi environ 12'000. Pour l'estimation de la couverture des logements, tous les logements des bâtiments qui n'ont pas pu être entièrement controlés sont exclus du calcul des quantités x_{1+}^g et x_{+1}^g intervenant dans les équations (20) et (21), et un facteur de correction uniforme est appliqué au poids de sondage π_g

dans les carrés concernés[6]. Ce facteur est égal au rapport entre le nombre de bâtiments entièrement controlés et le nombre de bâtiments observés sur le terrain dans le carré. En toute rigueur, dans un carré touché par la non-réponse, les quantités x^g_{1+} et x^g_{+1} ne sont plus fixes puisqu'elles dépendent du mécanisme de réponse. L'effet de cette très rare non-réponse est négligé dans les estimations de variance et les calculs sont réalisés comme si ces quantités étaient fixes dans tous les carrés.

Estimation dans un domaine

On peut s'intéresser aux équivalents des taux de couverture définis par (17), (18) et (19) au sein de différents domaines. Les résultats sont par exemple ventilés par catégorie de bâtiment (exclusivement à usage d'habitation ou bien avec usage annexe), ou par grande région. L'estimation dans un domaine est réalisée en introduisant une variable indicatrice pour l'appartenance au domaine dans les estimateurs $\hat{x}_{1+}$, $\hat{x}_{+1}$ et $\widehat{C}$, conformément à ce qui est usuellement pratiqué en théorie des sondages.

6 Pondération pour l'échantillon de personnes

Pour les personnes, le taux de non-réponse[7] est estimé aux alentours de 21%, et la méthode de pondération est plus complexe que pour les bâtiments et logements. Notons s_P l'échantillon des personnes cibles (cf. section 3) atteintes sur le terrain à leur résidence principale et répondantes. Cet échantillon est souvent appelé "P-sample" dans la littérature. 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 de réponse de la personne cible i. Le poids w_i d'une personne cible i de s_P est alors donné par :

$$w_i = d_g / p_i. \tag{33}$$

La somme des $1/p_i$ sur le P-sample s_P fournit une estimation du nombre d'éléments de la liste B (selon les notations de la section 2), c'est-à-dire une estimation du nombre de personnes cibles que la procédure de l'enquête de couverture aurait permis d'enquêter en l'absence totale de non-réponse.
Afin d'estimer la probabilité p_i, on remarque tout d'abord que toutes les personnes occupant un même logement ℓ ont la même probabilité de réponse, puisqu'une même personne répond pour l'ensemble du ménage. On désigne par s_L l'échantillon de logements présents dans s_G et par s_L^{int} le sous-échantillon de s_L pour lesquels on a réussi à réaliser une interview. On a donc, pour toute personne i occupant le même logement ℓ :

$$p_i = \underbrace{\mathrm{P}\left(\ell \in s_L^{int} \mid \ell \in s_L\right)}_{p_\ell}, \tag{34}$$

6. En revanche, aucun bâtiment n'est exclu lors de l'estimation de la couverture des bâtiments.

7. Le taux de réponse est estimé par le rapport entre le nombre de logements répondants et le nombre de logements qu'il aurait fallu interroger. Ce taux de réponse ne peut être estimé que lorsque la pondération est terminée. En effet, si le nombre de logements répondants dans les zones sélectionnées est connu, le nombre de logements qu'il aurait fallu interroger ne l'est pas et est estimé lors de la pondération. Les informations relevées sur le terrain par les enquêteurs permettent de déterminer le nombre de logements, toutes catégories confondues, présents dans les zones sélectionnées. Pour obtenir une estimation du nombre de logements qu'il aurait fallu interroger, il faut encore estimer la part des logements vides et des résidences secondaires, et la déduire du nombre de logements toutes catégories confondues.

où P désigne la loi de probabilité jointe du plan de sondage et du mécanisme de réponse. De plus, on a pour tout logement ℓ :

$$p_\ell = \underbrace{\mathrm{P}\left(\ell \in s_L^{int} \mid \ell \in s_L^{cont}\right)}_{p_\ell^1} \underbrace{P\left(\ell \in s_L^{cont} \mid \ell \in s_L\right)}_{p_\ell^2}, \tag{35}$$

où s_L^{cont} désigne le sous-échantillon de logements que l'on a réussi à "contacter". Nous décrivons ci-après la façon dont nous procédons pour estimer les quantités p_ℓ^1 et p_ℓ^2 qui déterminent la probabilité de réponse p_i, cf. équations (34) et (35).

Estimation de p_ℓ^1

La méthode utilisée est celle de la régression logistique et les variables auxiliaires disponibles pour les logements que l'on a réussi à contacter sont les suivantes :

– la grande région,

– la taille du carré, qui est un proxy de la densité de population,

– le nombre de personnes présentes dans le ménage, qui est connu pour les ménages répondants ainsi que pour les ménages non-répondants qui ont néanmoins accepté de participer à l'enquête de non-réponse [8],

– la façon dont le contact avec le logement a été établi, à savoir par internet, par face à face lors du premier passage de l'enquêteur, par téléphone, ou par face à face lors de la troisième phase de l'enquête.

Cela sous-entend que l'on considère que la façon dont le contact a été établi est une variable déterministe. Nous faisons en fait l'hypothèse que cette variable est très liée à des facteurs déterministes tels que le degré de coopération ou la disponibilité des personnes qui occupent le logement. Par exemple, les personnes qui se connectent spontanément à internet pour commencer à remplir le questionnaire sont sans doute des personnes coopératives, tandis que les personnes contactées par face à face lors de la troisième phase de l'enquête, alors que toutes les tentatives de contact précédentes ont échoué, sont probablement des personnes peu disponibles et/ou peu coopératives [9].
Puisque l'on ne s'intéresse qu'aux personnes cibles atteintes à leur résidence principale, idéalement, le modèle permettant d'estimer p_ℓ^1 devrait être ajusté au sein de l'ensemble des résidences principales des personnes cibles situées dans l'échantillon de zones géographiques. Cela n'est cependant pas possible en pratique, car on ne peut pas délimiter cet ensemble. On ne peut en effet déterminer si un logement contacté est ou non la résidence principale d'au moins une personne cible que dans le cas où une interview a pu être réalisée. On ajuste donc le modèle au sein de l'ensemble de tous les logements contactés, et on fait l'hypothèse, qui semble raisonnable, qu'une fois le contact établi, la propension à répondre est la même que le

8. La façon de procéder pour tenir compte du fait que certains ménages refusent également de participer à l'enquête de non-réponse est décrite dans l'annexe B.

9. Malgré les arguments avancés, la façon dont le contact a été établi comporte tout de même une composante aléatoire. Nous avons reconduit toute la procédure de pondération et d'extrapolation, mais sans utiliser la façon dont le contact a été établi comme variable auxiliaire lors de l'estimation des probabilités de réponse p_ℓ^1. Les taux de couverture obtenus de cette façon sont extrêmement proches de ceux publiés, et les conclusions absolument identiques.

logement soit ou non la résidence principale d'une personne cible. Une fois le modèle ajusté, on ne conserve l'estimation de p_ℓ^1 que pour les logements pour lesquels on dispose d'une interview et qui sont la résidence d'une personne cible. Plus de détails sur la procédure appliquée pour estimer p_ℓ^1 se trouvent dans l'annexe B.

Estimation de p_ℓ^2

A ce stade, les seules informations auxiliaires disponibles pour tous les logements sont la grande région et la taille du carré. Nous sommes confrontés au même type de difficulté que pour l'estimation de p_ℓ^1 : on ne peut pas estimer p_ℓ^2 à l'aide d'une régression logistique sur l'ensemble des résidences principales des personne cibles, puisque l'on ne peut pas délimiter cet ensemble. La solution retenue est cependant différente de celle utilisée pour estimer p_ℓ^1. Elargir la régression logistique à l'ensemble de tous les logements de s_L n'est en effet sans doute pas la meilleure option. L'hypothèse que la probabilité de contact est la même pour les résidences principales des personnes cibles que pour les autres logements ne semble pas très raisonnable, puisque la probabilité de contact est sans doute bien plus faible pour une résidence secondaire que pour une résidence principale. De plus, certains logements sont totalement inoccupés et sont donc "hors-cible" pour l'enquête de couverture auprès des personnes. Pour contourner cette dificulté, nous utilisons des méthodes plus spécifiques aux enquêtes de couverture. Tout d'abord, on suppose que les probabilités de contact p_ℓ^2 sont homogènes au sein de sous-groupes. Concrètement, afin que la taille des sous-groupes ne soit pas trop petite, nous renonçons à croiser la grande région et la taille du carré pour les constituer et nous ne retenons que la grande région. La grande région est indicée par gr, et on note p_{gr}^2 la valeur commune des p_ℓ^2 au sein de la grande région gr. Pour chaque grande région gr :

– $N_{L,gr}^{cont}$ désigne le nombre de logements de s_L que l'on a réussi à contacter, parmi les résidences principales des personne cibles que la procédure de l'enquête de couverture aurait permis d'enquêter en l'absence totale de non-réponse,

– $N_{L,gr}^{\overline{cont}}$ désigne le nombre de ceux que l'on n'a pas réussi à contacter,

de sorte que :

$$p_{gr}^2 \approx \frac{N_{L,gr}^{cont}}{N_{L,gr}^{cont} + N_{L,gr}^{\overline{cont}}}. \tag{36}$$

Les quantités $N_{L,gr}^{cont}$ et $N_{L,gr}^{\overline{cont}}$ ne peuvent pas être dénombrées puisque l'on ne sait si un logement est la résidence principale d'une personne cible que lorsqu'une interview a pu être réalisée. Nous devons donc les estimer. Puisque p_ℓ^1 est la probabilité qu'une interview ait été réalisée pour le logement ℓ, sachant qu'un contact a pu être établi, la probabilité p_ℓ^1 permet de réaliser une extrapolation de l'ensemble des logements interviewés à l'ensemble des logements contactés. On peut donc estimer $N_{L,gr}^{cont}$ par :

$$\widehat{N}_{L,gr}^{cont} = \sum_{\ell \in \tilde{s}_{L,gr}^{int}} \frac{1}{p_\ell^1}, \tag{37}$$

où $\tilde{s}_{L,gr}^{int}$ désigne le sous-ensemble, parfaitement identifiable, des logements de la grande région gr pour lesquels on dispose d'une interview qui a montré qu'il s'agit de la résidence principale d'une personne cible. L'estimation de $N_{L,gr}^{cont}$ tient donc compte de toutes les informations auxiliaires utilisées pour estimer les probabilités p_ℓ^1 (cf. page 19).

Une approximation de $N_{L,gr}^{\overline{cont}}$ peut être obtenue en énumérant, pour chaque grande région gr, les logements de l'échantillon de zones géographiques qui sont la résidence principale d'une personne cible d'après nos données administratives, et que l'on n'a par ailleurs pas réussi à contacter. Cette liste est entâchée de sur- et de sous-couverture, que nous avons tenté de corriger, du moins approximativement, grâce à la procédure décrite dans l'annexe C. Les probabilités de contact p_{gr}^2 peuvent alors être estimées en remplaçant dans la formule (36) les quantités $N_{L,gr}^{cont}$ et $N_{L,gr}^{\overline{cont}}$ par leurs estimations.
Nous présentons dans le tableau 3 les estimations des probabilités de contact obtenues pour chaque grande région avec les deux méthodes décrites ci-après, afin de déterminer si la procédure relativement complexe qui a été retenue apporte une plus-value par rapport à une méthode plus basique.

– **Méthode "basique" :** estimer p_ℓ^2 grâce à une régression logistique réalisée sur l'échantillon de logements s_L (tous types de logement confondus, résidences principales ou secondaires ainsi logements inoccuppés) en utilisant la grande région comme variable auxiliaire.

– **Méthode retenue :** au sein de chaque grande région gr, estimer $p_\ell^2 = p_{gr}^2$ sur la base de l'équation (36).

Tableau 3 Estimation des probabilités de contact en fonction de la méthode utilisée

Grande région	Méthode basique	Méthode retenue
Région lémanique	0.792	0.859
Espace Mittelland	0.910	0.922
Suisse du Nord-Ouest	0.876	0.909
Zürich	0.863	0.885
Suisse orientale	0.809	0.923
Suisse centrale	0.911	0.945
Tessin	0.795	0.874

Les résultats présentés dans le tableau 3 doivent être interprétés avec précaution car ils ne sont pas accompagnés de leur écart-type. Dans le cas de la méthode retenue, l'estimation de la précision n'est en effet pas immédiate et ces calculs n'ont pas pu être réalisés durant le temps disponible pour mener à bien ce projet. Cependant, comme cela était prévisible, on constate que l'estimation de la probabilité de contact est plus élevée dans le cas de la méthode basique que dans celui de la méthode retenue, et que l'écart peut être relativement important dans certaines grandes régions, par exemple en Suisse orientale. Cela est très probablement dû au fait qu'avec la méthode basique, les probabilités de contact sont estimées comme si tous les logements, même les résidences secondaires et les logements totalement inoccupés, avaient dû être atteints, ce qui conduit à une sous-estimation du taux de contact. La méthode retenue, malgré sa relative complexité, présente donc un réel intérêt.

7 Procédure d'estimation pour les personnes

Définition des quantités à estimer

En accord avec les notations introduites à la section 2 :

N désigne ici la "vraie" taille de la population cible, qui est inconnue.

x_{1+} désigne ici la taille de la population cible obtenue d'après nos données administratives, mais nettoyé de la sur-couverture.

Nous désignons de plus par C la taille de la population cible obtenue d'après nos données administratives, y compris la sur-couverture. Les quantités à estimer sont les suivantes.

– Le taux de sous-couverture $(1 - F^-)$, où :

$$F^- = \frac{x_{1+}}{N}. \tag{38}$$

– Le taux de sur-couverture $(1 - F^+)$, où :

$$F^+ = \frac{x_{1+}}{C}. \tag{39}$$

– Le taux de sous-couverture nette $(1 - F)$, où :

$$F = \frac{F^-}{F^+} = \frac{C}{N}. \tag{40}$$

Estimateurs considérés

Nous rappelons que pour les raisons discutées à la section 4, nous nous plaçons sous les hypothèses du modèle de Petersen pour estimer le taux de sous-couverture des décomptes de personnes. Nous utilisons donc l'estimateur $\widehat{N}$ défini par (6) pour estimer la taille de la population, puis nous en déduisons une estimation $(1 - f^-)$ du taux de sous-couverture, la quantité f^- étant définie par (10). En notant

$$f_k^- = \frac{\hat{x}_{11}^k}{\hat{x}_{+1}^k}. \tag{41}$$

le facteur de correction pour la sous-couverture au sein de la cellule d'estimation k, l'estimateur $\widehat{N}$ peut se réécrire sous la forme :

$$\widehat{N} = \sum_{k=1}^{K} x_{1+}^k \frac{1}{f_k^-}. \tag{42}$$

Les estimations $\hat{x}_{+1}^k$ et $\hat{x}_{11}^k$ intervenant dans le calul du facteur de correction f_k^- sont définies par :

$$\hat{x}_{+1}^k = \sum_{i \in s_P \cap k} w_i = \sum_{i \in s_P} w_i \mathbf{1}_{\{i \in k\}} \tag{43}$$

et

$$\hat{x}_{11}^k = \sum_{i \in s_P \cap k} w_i \mathbf{1}_{\{i \in A\}} = \sum_{i \in s_P} w_i \mathbf{1}_{\{i \in A\}} \mathbf{1}_{\{i \in k\}} \tag{44}$$

où w_i désigne le poids d'extrapolation calculé à la section 6 pour la personne i énumérée lors de l'enquête de couverture, $\mathbf{1}_{\{i \in k\}}$ est une variable indicatrice de l'appartenance à la cellule d'estimation k, et où $\mathbf{1}_{\{i \in A\}}$ indique si la personne i du P-sample est considérée comme retrouvée ou non dans nos données administratives. Les critères selon lesquels une personne est considérée comme correctement appariée à nos données administratives sont discutés page 25. Les cellules d'estimation k sont construites à partir d'un algorithme de segmentation appliqué au P-sample s_P. Pour cet algorithme de segmentation, la variable d'intérêt est la variable $\mathbf{1}_{\{i \in A\}}$. Les variables auxiliaires sont des variables sociodémographiques comme l'âge, la nationalité, la grande région, le canton (ou regroupement de cantons), etc. La taille minimale des cellules est fixée à 100. Puisque l'algorithme de segmentation est appliqué au P-sample s_P en choisissant la variable $\mathbf{1}_{\{i \in A\}}$ comme variable d'intérêt, les cellules d'estimation sont construites de façon à ce que les probabilités $\mathrm{P}(i \in A \mid i \in s_P)$ y soient homogènes. D'après l'hypothèse d'indépendance entre les probabilités de capture et de recapture du modèle de Petersen (cf. page 8 section 2), la probabilité $\mathrm{P}(i \in A \mid i \in s_P)$ est égale à la probabilité $p_{i,1+}$ de faire partie de la liste A, de sorte que la première partie de l'hypothèse d'homogénéité du modèle de Petersen devrait être respectée au sein de chaque cellule d'estimation. On suppose que l'hypothèse d'homogénéité des probabilités $p_{i,+1}$ de faire partie de la liste B est également respectée au sein des cellules d'estimation.
La méthode d'estimation de la sous-couverture proposée ici diffère de celle appliquée par Renaud (2007), qui se base sur le taux global d'individus de s_P non retrouvés dans les données administratives, sans avoir recours aux cellules d'estimation. Dans Renaud (2007), les cellules d'estimation ne sont utilisées que pour l'estimation de la sous-couverture nette. L'avantage de la méthode que nous utilisons est que l'hypothèse d'homogénéité du modèle de Petersen est mieux respectée. De plus, elle garantit d'obtenir des estimations de la sous-couverture cohérentes avec celles de la sous-couverture nette. Puisque le taux de sous-couverture nette reflète le bilan entre la sous- et la sur-couverture, on s'attend en effet à ce que le taux de sous-couverture nette soit moins élevé que le taux de sous-couverture. Or comme le montre le tableau 2 de Renaud (2007), cela n'est pas toujours le cas pour le recensement de 2000, par exemple pour les étrangers temporaires ou le relevé SEMI-CLASSIC.
Pour estimer la composante de sur-couverture, deux types de contrôles sont effectués.

(i) Contrôles effectués sur l'ensemble des données administratives

On utilise le fichier de taille C qui contient la totalité de la population cible d'après nos données administratives. On effectue sur ce fichier un certain nombre de contrôles qui ne peuvent pas être réalisés lors de la production du décompte de la population qui sera publié. Par exemple, l'enquête de couverture étant réalisée quelques mois après la publication, on peut utiliser les données administratives qui nous sont parvenues après la publication des chiffres. On peut ainsi repérer certains décès ou départs à l'étranger intervenus avant la date de référence mais dont l'annonce nous est parvenue trop tardivement pour être pris en compte dans le décompte officiel.

(ii) Contrôles effectués sur un échantillon

Certains contrôles demandent plus de "temps machine" ou nécessitent des interventions manuelles. Ils ne peuvent donc pas être réalisés sur l'ensemble du fichier mais seulement sur un échantillon. L'échantillon sur la base duquel la composante de sur-couverture est estimée est souvent appelé E-sample dans la littérature. Notre E-sample est constitué de toutes les personnes qui, d'après les données administratives, font partie de la population cible et vivent dans une des zones sélectionnées. Sur la base de cet échantillon, on

a par exemple réalisé une recherche de doublons plus minutieuse que celle qui peut être menée pour la publication officielle. Pour la publication, la recherche de doublons est en effet effectuée sur la base du numéro AVS. Pour l'enquête de couverture en revanche, on a envisagé le fait que certains numéros AVS puissent être mal attribués. Il n'est donc pas exclu qu'une même personne soit enregistrée deux fois avec deux numéros AVS différents. A cet effet, nous avons fait appel à l'algorithme de Levenshtein (1966), qui permet de calculer une "distance" entre deux chaînes de caractères, ici les noms, prénoms, et dates de naissance. Cet algorithme a permis d'identifier 210 personnes dans l'E-sample ayant un doublon potentiel dans le fichier de taille C qui contient la totalité de la population cible. Ces cas sont ensuite analysés manuellement. Aucun doublon n'a pu être détecté.

Les deux types de contrôles ont été réalisés par la section POP. Tous les cas de sur-couverture détectés l'ont été lors de contrôles du type (i) effectués sur l'ensemble des données administratives. Par conséquent, les quantités en jeu sont supposées connues et il n'est pas nécessaire d'estimer le taux de sur-couverture $(1 - F^+)$. Le taux de sous-couverture nette est estimé par $(1 - f)$, où f est défini par :

$$f = \frac{C}{\widehat{N}}. \tag{45}$$

Estimation dans un domaine

On peut s'intéresser aux équivalents des taux de couverture définis par (38), (39) et (40) au sein de différents domaines. On désigne par N_d, x^d_{1+}, C_d l'équivalent des quantités N, x_{1+}, C respectivement au sein du domaine d. On considère :

– Le taux de sous-couverture $(1 - F^-_d)$ au sein du domaine d, où :

$$F^-_d = \frac{x^d_{1+}}{N_d}. \tag{46}$$

– Le taux de sur-couverture $(1 - F^+_d)$ au sein du domaine d, où :

$$F^+_d = \frac{x^d_{1+}}{C_d}. \tag{47}$$

– Le taux de sous-couverture nette $(1 - F_d)$ au sein du domaine d, où :

$$F_d = \frac{F^-_d}{F^+_d} = \frac{C_d}{N_d}. \tag{48}$$

Pour estimer les taux de couverture dans un domaine d, nous procédons comme suit. Nous commençons par estimer la taille N_d de la population au sein du domaine d :

$$\widehat{N}_d = \sum_{k=1}^{K} x^{kd}_{1+} \frac{1}{f^-_k},$$

où f^-_k est le facteur de correction pour la cellule d'estimation k défini par (41) et où x^{kd}_{1+} désigne la taille de la population cible dans l'intersection de la cellule d'estimaton k et du domaine d obtenue d'après nos données administratives, mais nettoyé de la sur-couverture. Cet estimateur

est un estimateur de type “synthétique” puisque l’on ne calcule pas de facteur de correction spécifique pour l’intersection de la cellule d’estimaton k et du domaine d, mais que l’on applique le facteur de correction f_k^-. Cela se justifie par le fait que les cellules d’estimation k sont construites de façon à respecter au mieux l’hypothèse d’homogénéité du modèle de Pertersen. Il n’y a donc pas lieu de calculer des facteurs de correction spécifiques pour l’intersection des cellules d’estimaton et des domaines. De plus, la taille de certaines intersections peut ne pas être suffisante pour pouvoir y produire des estimations qui ne soient pas trop instables. Le taux de sous-couverture est alors estimé par $1 - f_d^-$, où :

$$f_d^- = \frac{x_{1+}^d}{\widehat{N}_d}.$$

Là encore, il n’est pas nécessaire d’estimer le taux de sur-couverture $(1 - F_d^+)$ puisque les quantités en jeu sont connues. Le taux de sous-couverture nette est estimé par $(1 - f_d)$, où f_d est défini par :

$$f_d = \frac{C_d}{\widehat{N}_d}. \tag{49}$$

Choix des statuts d’appariement et d’énumération corrects

La variable $\mathbf{1}_{\{i \in A\}}$ qui, comme le montre la formule (44), intervient dans l’estimation de N, est définie sur la base de critères selon lesquels une personne du P-sample est considérée ou non comme correctement appariée à nos données administratives. Ces critères, appelés critères d’appariement correct, jouent donc un rôle clé dans les estimations de la couverture [10]. Il en va de même des critères d’énumération corrects selon lesquels une personne est considérée comme devant effectivement être énumérée dans le décompte officiel ou est au contraire en sur-couverture. Afin que ces critères soient clairement définis, un certain nombre de questions doivent être tranchées. Par exemple, si une personne du P-sample est retrouvée dans nos données administratives mais y figure comme faisant partie d’un ménage collectif, considère-t-on que $\mathbf{1}_{\{i \in A\}} = 1$? Lorsque l’on souhaite mettre en balance la sous- et la sur-couverture et calculer une sous-couverture nette, il convient lors de la définition des critères d’appariement et d’énumération corrects de prendre en compte la notion d’équilibrage discutée par exemple dans Renaud (2007) ou Hogan (2003). Concrètement, cela signifie dans l’exemple précédent que l’on ne peut refuser l’appariement d’une personne du P-sample avec une personne qui figure dans nos données administratives comme faisant partie d’un ménage collectif que si l’on est en mesure de repérer dans le décompte officiel la part de sur-couverture due aux personnes qui vivent en réalité dans un ménage collectif. Plus généralement, les critères selon lesquels un appariement n’est pas considéré comme correct doivent être les mêmes que les critères selon lesquels une personne énumérée dans notre décompte officiel de la population est considérée comme en sur-couverture car hors-cible en réalité.
Dans l’EC2013 comme en 2000, nous avons cherché à détecter la part de sur-couverture due aux personnes hors-cible car décédées ou parties à l’étranger avant la date de référence, mais nous ne vérifions pas si les personnes prises en compte dans notre décompte officiel font bien

10. Lorsqu’un numéro AVS est disponible, la personne du P-sample est recherchée dans l’ensemble des données administratives sur la base de cet identifiant. En l’absence de numéro AVS, la personne du P-sample est recherchée dans l’ensemble des données administratives sur la base de son nom, prénom, et date de naissance. Cette dernière procédure peut inclure des traitements manuels. Si la personne n’est pas retrouvée dans l’ensemble des données administratives, on considère que $\mathbf{1}_{\{i \in A\}}$=0. Si la personne est retrouvée dans l’ensemble des données administratives, il reste encore à décider si l’appariement est considéré ou non comme correct, selon des critères qui sont discutés sans ce paragraphe.

partie en réalité de la bonne catégorie de population. Par conséquent, si une personne du P-sample est retrouvée dans nos données administratives mais y figure comme faisant partie d'un ménage collectif ou bien comme disposant d'un permis de séjour non pris en compte dans la population cible, l'appariement est considéré comme correct et $\mathbf{1}_{\{i\in A\}} = 1$. Cela correspond à une notion plutôt "peu sévère" des défauts de couverture. Ce choix est motivé par le fait qu'en se basant sur les mêmes critères que pour le recensement de 2000, on peut évaluer l'effet du remplacement du recensement traditionnel par un recensement basé sur des données administratives. D'autre part, nous sommes confrontés aux mêmes type de limitations que pour le recensement de 2000 : nous ne disposons pas d'informations complémentaires qui nous permettraient de vérifier si une personne énumérée dans le décompte officiel faisait bien partie en réalité de la bonne catégorie de population.

8 Précision des estimateurs de la couverture des décomptes de personnes

Puisque le taux de sur-couverture est connu et n'a pas besoin d'être estimé, nous nous intéressons ici à la précision des estimateurs f^-, f, f_d^-, et f_d. Les calculs présentés dans ce qui suit permettent d'obtenir une approximation de la précision de f^-, mais ils se transposent facilement au cas des autres estimateurs. La précision de f^- est mesurée par l'erreur quadratique moyenne attendue définie par :

$$\mathrm{MSE}\left(f^-\right) = \mathbf{E}\left(f^- - F^-\right)^2, \tag{50}$$

où on rappelle que $\mathbf{E}(\cdot)$ désigne l'espérance sous l'aléa du modèle de Pertersen et du plan de sondage pour les individus, qui découle de la sélection des zones géographiques ainsi que de la non-réponse. Pour les raisons exposées dans l'annexe A, l'espérance $\mathbf{E}(\cdot)$ est définie par :

$$\mathbf{E}(\cdot) = \mathbf{E}_m\mathbf{E}_p(\cdot \mid A),$$

où on rappelle que $\mathbf{E}_m(\cdot)$ est l'espérance sous le modèle de Petersen et $\mathbf{E}_p(\cdot \mid A)$ est l'espérance sous le plan conditionnellement au mécanisme aléatoire qui conduit à la liste A. Afin de ne pas alourdir les notations de cette section, on utilise la notation $\mathbf{E}_p(\cdot)$ à la place de $\mathbf{E}_p(\cdot \mid A)$, de sorte que :

$$\mathrm{MSE}\left(f^-\right) = \mathbf{E}_m\mathbf{E}_p\left(f^- - F^-\right)^2. \tag{51}$$

Afin de proposer une procédure d'estimation de la précision de f^-, nous commençons par obtenir une approximation de $\mathrm{MSE}\left(f^-\right)$ donnée dans la formule (70). Cette formule sert ensuite de base pour proposer une estimation de l'erreur quadratique moyenne attendue (cf. page 30).

Approxination de l'erreur quadratique moyenne attendue

On remarque tout d'abord que :

$$\begin{aligned}
\mathrm{MSE}\left(f^-\right) &= \mathbf{E}_m\mathbf{E}_p\left[f^- - \mathbf{E}_p\left(f^-\right) + \mathbf{E}_p\left(f^-\right) - F^-\right]^2 \\
&= \mathbf{E}_m\mathbf{E}_p\left[f^- - \mathbf{E}_p\left(f^-\right)\right]^2 \\
&\quad +2\,\mathbf{E}_m\mathbf{E}_p\left\{\left[f^- - \mathbf{E}_p\left(f^-\right)\right]\left[\mathbf{E}_p\left(f^-\right) - F^-\right]\right\} \\
&\quad +\mathbf{E}_m\mathbf{E}_p\left[\mathbf{E}_p\left(f^-\right) - F^-\right]^2.
\end{aligned} \tag{52}$$

Puisque

$$\begin{aligned} & \mathbf{E}_m\mathbf{E}_p\left\{\left[f^- - \mathbf{E}_p\left(f^-\right)\right]\left[\mathbf{E}_p\left(f^-\right) - F^-\right]\right\} \\ = \; & \mathbf{E}_m\left\{\mathbf{E}_p\left[f^- - \mathbf{E}_p\left(f^-\right)\right]\left[\mathbf{E}_p\left(f^-\right) - F^-\right]\right\} \\ = \; & \mathbf{E}_m\left\{\left[\mathbf{E}_p\left(f^-\right) - \mathbf{E}_p\left(f^-\right)\right]\left[\mathbf{E}_p\left(f^-\right) - F^-\right]\right\} \\ = \; & 0, \end{aligned} \tag{53}$$

et

$$\mathbf{E}_m\mathbf{E}_p\left[\mathbf{E}_p\left(f^-\right) - F^-\right]^2 = \mathbf{E}_m\left[\mathbf{E}_p\left(f^-\right) - F^-\right]^2, \tag{54}$$

on obtient l'expression suivante de l'erreur quadratique moyenne attendue :

$$\begin{aligned} \mathrm{MSE}\left(f^-\right) &= \mathbf{E}_m\mathbf{E}_p\left[f^- - \mathbf{E}_p\left(f^-\right)\right]^2 + \mathbf{E}_m\left[\mathbf{E}_p\left(f^-\right) - F^-\right]^2 \\ &= \underbrace{\mathbf{E}_m\mathbf{Var}_p\left(f^-\right)}_{\mathrm{MSE}^A} + \underbrace{\mathbf{E}_m\left[\mathbf{E}_p\left(f^-\right) - F^-\right]^2}_{\mathrm{MSE}^B}, \end{aligned} \tag{55}$$

où $\mathbf{Var}_p$ désigne la variance sous le processus de sélection des individus. On observe ensuite que, si la taille du P-sample est suffisamment grande et si les cellules d'estimation k ne sont pas trop petites :

$$\mathbf{E}_p\left(f^-\right) \simeq f_0^-, \tag{56}$$

où f_0^- est le facteur de correction de la sous-couverture que l'on aurait obtenu si l'enquête de couverture avait pu être menée sur tout le territoire [11]. Par conséquent, le second terme de l'erreur quadratique moyenne attendue MSE^B est approximativement égal à l'erreur quadratique sous l'aléa du modèle de Petersen, dans le cas idéal où l'enquête de couverture aurait pu être menée sur tout le territoire :

$$\mathrm{MSE}^B \simeq \mathbf{E}_m\left(f_0^- - F^-\right)^2. \tag{57}$$

Il en découle que, dans notre cas où le taux de sondage des zones géographiques est de l'ordre de moins d'un demi-pourcent, la quantité MSE^B intervenant dans la formule (55) devrait être bien plus faible que MSE^A, qui dépend de l'erreur d'échantillonnage [12]. Nous négligeons donc dans ce qui suit le second terme de l'erreur quadratique moyenne attendue, et nous nous basons sur l'approximation suivante :

$$\mathrm{MSE}\left(f^-\right) \simeq \mathrm{MSE}^A = \mathbf{E}_m\mathbf{Var}_p\left(f^-\right). \tag{58}$$

11. Dans le cas où l'on ne partitionne pas la population en cellules d'estimation, le facteur f_0^- est défini par la formule (5) de la section 2.

12. Les calculs permettant de prouver rigoureusement cette approximation n'ont pas pu être réalisés durant le temps disponible pour mener à bien ce projet. Cependant, pour étayer nos propos, nous pouvons citer des calculs très proches réalisés par Wolter (1986). Les résultats (vi) des théorèmes 3.1 et 3.2 de Wolter (1986) montrent en effet que, si l'échantillon de zones géographiques est sélectionné selon un plan simple, et si la population n'est pas partionnée en cellules d'estimation, alors la variance $\mathbf{Var}_{mp}$ sous les deux sources d'aléa peut être décomposée sous la forme :

$$\mathbf{Var}_{mp}\left(\hat{N}\right) = V_A + V_B,$$

où

- V_A est la part de variance qui aurait été obtenue si l'enquête de couverture avait pu être menée sur tout le territoire et est égale à $N\frac{(1-p_{1+})(1-p_{+1})}{p_{1+}p_{+1}}$,
- V_B est la part de variance due à l'échantillonnage et est de l'ordre de $\frac{1}{0.005}\cdot N\frac{1-p_{1+}}{p_{1+}p_{+1}}$, c'est-à-dire $200\cdot N\frac{1-p_{1+}}{p_{1+}p_{+1}}$, si le taux de sondage des zones géographiques est de 0.5%.

Nous utilisons maintenant des techniques de linéarisation (cf. Deville, 1999) pour obtenir une approximation de la variance $\mathbf{Var}_p$ de l'estimateur f^- qui, comme le montrent les formules (6) et (10), est une fonction des estimateurs de totaux $\hat{x}^k_{+1}$ et $\hat{x}^k_{11}$:

$$f^- = F\left(\hat{x}^1_{+1}, \hat{x}^1_{11}, \ldots, \hat{x}^K_{+1}, \hat{x}^K_{11}\right), \tag{59}$$

où

$$F(u_1, v_1, \ldots, u_K, v_K) = \frac{x_{1+}}{\sum_{k=1}^{K} x^k_{1+} \dfrac{u_k}{v_k}}. \tag{60}$$

Puisque pour tout $k = 1, \ldots, K$:

$$\begin{cases} \dfrac{\partial F}{\partial u_k} = -\dfrac{x_{1+}}{\left(\sum_{k=1}^{K} x^k_{1+} \dfrac{u_k}{v_k}\right)^2} \dfrac{x^k_{1+}}{v_k} \\ \dfrac{\partial F}{\partial v_k} = -\dfrac{x_{1+}}{\left(\sum_{k=1}^{K} x^k_{1+} \dfrac{u_k}{v_k}\right)^2} x^k_{1+} u_k \left(-\dfrac{1}{v_k^2}\right) = \dfrac{x_{1+}}{\left(\sum_{k=1}^{K} x^k_{1+} \dfrac{u_k}{v_k}\right)^2} x^k_{1+} \dfrac{u_k}{v_k^2}, \end{cases} \tag{61}$$

la linéarisée de f^- s'écrit, pour tout individu i du P-sample s_P :

$$\ell_i^- = \frac{x_{1+}}{\left(\sum_{k=1}^{K} x^k_{1+} \dfrac{x^k_{+1}}{x^k_{11}}\right)^2} \sum_{k=1}^{K} x^k_{1+} \left[-\frac{1}{x^k_{11}} \mathbf{1}_{\{i \in k\}} + \frac{x^k_{+1}}{(x^k_{11})^2} \mathbf{1}_{\{i \in A\}} \mathbf{1}_{\{i \in k\}}\right]. \tag{62}$$

Cela permet l'approximation suivante de $\mathbf{Var}_p\left(f^-\right)$:

$$\mathbf{Var}_p\left(f^-\right) \simeq \mathbf{Var}_p\left(\sum_{i \in s_P} w_i \ell_i^-\right). \tag{63}$$

On observe ensuite, en utilisant la définition du poids d'extrapolation w_i donnée par (33), que :

$$\sum_{i \in s_P} w_i \ell_i^- = \sum_{g \in s_G} d_g \sum_{i \in s_P \cap g} \frac{1}{p_i} \ell_i^- = \sum_{g \in s_G} d_g \widehat{L}_g^-, \tag{64}$$

où

$$\widehat{L}_g^- = \sum_{i \in s_P \cap g} \frac{1}{p_i} \ell_i^-. \tag{65}$$

En combinant (63) et (64), il vient :

$$\mathbf{Var}_p\left(f^-\right) \simeq \mathbf{Var}_p\left(\sum_{g \in s_G} d_g \widehat{L}_g^-\right). \tag{66}$$

Grâce à la loi de variance totale, nous décomposons maintenant la variance sous le plan $\mathbf{Var}_p$ en une part attribuable à la non-réponse et une part attribuable à l'échantillonnage. A cet effet, on désigne par $\mathbf{E}_r(\cdot)$ l'espérance sous le mécanisme de réponse et par $\mathbf{E}_s(\cdot)$ l'espérance sous

le plan de sondage des zones géographiques, avec des notations analogues pour la variance. On a :

$$\begin{aligned}\mathbf{Var}_p\left(\sum_{g\in s_G} d_g\widehat{L}_g^-\right) &= \mathbf{Var}_r\mathbf{E}_s\left(\sum_{g\in s_G} d_g\widehat{L}_g^- \mid r\right) + \mathbf{E}_r\mathbf{Var}_s\left(\sum_{g\in s_G} d_g\widehat{L}_g^- \mid r\right)\\ &= \mathbf{Var}_r\left(\sum_{g=1}^{G}\widehat{L}_g^-\right) + \mathbf{E}_r\mathbf{Var}_s\left(\sum_{g\in s_G} d_g\widehat{L}_g^- \mid r\right). \qquad (67)\end{aligned}$$

Les quantités $\widehat{L}_g^-$ dépendent du mécanisme de réponse, qui peut être modélisé par un plan de Poisson sur les ménages, puisqu'une fois qu'un des occupants d'un logement est contacté, il est amenée à répondre pour l'ensemble de son ménage. Par conséquent, les $\widehat{L}_g^-$ sont indépendants d'une zone géographique à l'autre, ce qui conduit à :

$$\mathbf{Var}_p\left(\sum_{g\in s_G} d_g\widehat{L}_g^-\right) = \underbrace{\sum_{g=1}^{G}\mathbf{Var}_r\left(\widehat{L}_g^-\right)}_{\mathbf{Var}_p^A} + \underbrace{\mathbf{E}_r\mathbf{Var}_s\left(\sum_{g\in s_G} d_g\widehat{L}_g^- \mid r\right)}_{\mathbf{Var}_p^B}. \qquad (68)$$

La variance $\mathbf{Var}_p$ est ainsi décomposée en une part $\mathbf{Var}_p^A$ qui correspond à la variance due à la non-réponse si l'enquête de couverture avait pu être réalisée sur l'ensemble du territoire, et un terme d'erreur supplémentaire $\mathbf{Var}_p^B$ dû à la sélection d'un échantillon de zones géographiques. Observons que la décomposition de la variance est habituellement présentée en conditionnant par rapport à l'aléa de l'échantillonnage, plutôt que par rapport au mécanisme de réponse comme cela est fait dans la formule (68) :

$$\begin{aligned}\mathbf{Var}_p\left(\sum_{g\in s_G} d_g\widehat{L}_g^-\right) &= \mathbf{Var}_s\mathbf{E}_r\left(\sum_{g\in s_G} d_g\widehat{L}_g^- \mid s\right) + \mathbf{E}_s\mathbf{Var}_r\left(\sum_{g\in s_G} d_g\widehat{L}_g^- \mid s\right)\\ &= \underbrace{\mathbf{Var}_s\left(\sum_{g\in s_G} d_g L_g^-\right)}_{\mathbf{Var}_p^{A'}} + \underbrace{\mathbf{E}_s\mathbf{Var}_r\left(\sum_{g\in s_G} d_g\widehat{L}_g^- \mid s\right)}_{\mathbf{Var}_p^{B'}}, \qquad (69)\end{aligned}$$

où

$$L_g^- = \mathbf{E}_r\left(\widehat{L}_g^- \mid s\right) = \sum_{i\in \mathrm{B}\cap g}\ell_i^-,$$

la dernière somme portant sur la part de la liste B de personnes cibles contenue dans la zone géographique g. Dans la décomposition (69), la part $\mathbf{Var}_p^{A'}$ correspond à la variance attribuable à l'échantillonnage en l'absence de non-réponse, tandis que $\mathbf{Var}_p^{B'}$ est un terme d'erreur supplémentaire dû à la non-réponse. Nous nous basons dans la suite sur la décomposition (68) plutôt que sur la décomposition (69), car les calculs à développer pour proposer une procédure d'estimation de variance s'en trouvent facilités. Estimer $\mathbf{Var}_p^{A'}$ revient en effet à estimer la variance sous un plan équilibré. Deville and Tillé (2005) ont proposé plusieurs formules à cet effet. La difficulté est ici que la "variable d'intérêt" L_g^- est inconnue. Si la quantité inconnue L_g^- est remplacée par son estimation $\widehat{L}_g^-$ dans une des formules proposées par Deville and Tillé (2005), l'effet sur l'estimateur de variance devrait être estimé et corrigé. Comme le montre la suite du document, se baser sur la formule (68) permet de contourner cette difficulté.

En combinant les équations (58), (66) et (68), on obtient finalement l'approximation suivante de l'erreur quadratique moyenne attendue :

$$\mathrm{MSE}\left(f^{-}\right) \simeq \mathbf{E}_m \Bigg[\underbrace{\sum_{g=1}^{G} \mathbf{Var}_r\left(\widehat{L}_g^{-}\right)}_{\mathbf{Var}_p^A} + \underbrace{\mathbf{E}_r \mathbf{Var}_s\left(\sum_{g \in s_G} d_g \widehat{L}_g^{-} \mid r\right)}_{\mathbf{Var}_p^B} \Bigg]. \tag{70}$$

On peut par ailleurs trouver une expression de $\mathbf{Var}_r\left(\widehat{L}_g^{-}\right)$. En utilisant la relation (34), on commence par réécrire la quantité $\widehat{L}_g^{-}$, qui est définie par (65), de la façon suivante :

$$\widehat{L}_g^{-} = \sum_{i \in s_P \cap g} \frac{1}{p_i} \ell_i^{-} = \sum_{\ell \in s_L^{int} \cap g} \frac{1}{p_\ell} \underbrace{\sum_{i \in \ell \cap s_P} \ell_i^{-}}_{L_\ell^{-}} = \sum_{\ell \in s_L^{int} \cap g} \frac{1}{p_\ell} L_\ell^{-}, \tag{71}$$

où on rappelle que s_L^{int} désigne l'échantillon de logements pour lesquels on dispose d'une interview. On observe ensuite que :

– le mécanisme de réponse est modélisé par un plan de Poisson sur les ménages,
– les poids $1/p_\ell$ permettent une extrapolation à l'ensemble des logements qui contiennent au moins une personne cible, et que la procédure de l'enquête de couverture aurait permis d'enquêter en l'absence totale de non-réponse. Cet ensemble correspond à la liste B pour les logements qui contiennent au moins une personne cible, est est dénoté par B_L.

Il découle alors de (71) que :

$$\mathbf{Var}_r\left(\widehat{L}_g^{-}\right) = \sum_{\ell \in \mathrm{B}_L \cap g} \frac{1-p_\ell}{p_\ell} \left(L_\ell^{-}\right)^2, \tag{72}$$

Estimation de l'erreur quadratique moyenne attendue

Tout d'abord, les quantités inconnues x_{11}^k et x_{+1}^k intervant dans la linéarisée ℓ_i définie par (62) doivent être remplacées par leurs estimations $\hat{x}_{11}^k$ et $\hat{x}_{+1}^k$.

Puisque par principe, une seule réalisation du modèle a été observée, nous estimons la quantité $\mathrm{MSE}\left(f^{-}\right)$, dont une approximation est donnée par (70), en estimant les deux termes de variance sous le plan $\mathbf{Var}_p^A$ et $\mathbf{Var}_p^B$ pour la réalisation du modèle qui est observée. En ce qui concerne le terme $\mathbf{Var}_p^A$, on peut proposer sur la base de la formule (72) l'estimateur suivant de $\mathbf{Var}_r\left(\widehat{L}_g^{-}\right)$:

$$\widehat{\mathbf{Var}}_r\left(\widehat{L}_g^{-}\right) = \sum_{\ell \in s_L^{int} \cap g} \frac{1-p_\ell}{p_\ell^2} \left(L_\ell^{-}\right)^2. \tag{73}$$

Le terme $\mathbf{Var}_p^A$ peut alors être estimé sans biais par :

$$\widehat{\mathbf{Var}}_p^A = \sum_{g \in s_g} d_g \widehat{\mathbf{Var}}_r\left(\widehat{L}_g^{-}\right). \tag{74}$$

En ce qui concerne le terme $\mathbf{Var}_p^B$ intervenant dans (70), on remarque tout d'abord qu'estimer $\mathbf{Var}_p^B$ revient à estimer la variance sous le plan de sondage des zones géographiques de

$$\widehat{T} = \sum_{g \in s_G} d_g \widehat{L}_g^{-},$$

en considérant les quantités $\widehat{L}_g^-$ comme fixes. Deville and Tillé (2005) ont proposé plusieurs formules permettant d'estimer la variance de l'estimateur d'un total sous un plan équilibré. Comme dans la section 5, nous utilisons la formule la plus simple correspondant au choix $c_g = (1-\pi_g)$ selon leurs notations, en traitant les quantités $\widehat{L}_g^-$ comme fixes.

9 Résultats

Le tableau 4 présente les taux de couverture obtenus au niveau national. La racine carrée de l'estimation de l'erreur quadratique moyenne attendue définie à la section 8 y est utilisée comme mesure de précision des estimateurs considérés et est donnée entre parenthèses. Pour la sur-couverture des personnes, la valeur donnée dans le tableau n'est pas accompagnée de mesure de précision, puisque comme cela est expliqué dans la section 7, la sur-couverture a été détectée lors de contrôles effectués sur l'ensemble des données administratives. On constate que, globalement, le nouveau système de recensement de la population basé sur des données administratives est de très bonne qualité. Ses défauts de sous-couverture sont encore réduits par rapport à 2000, puisque le taux de sous-couverture nette au niveau national passe de 1.64% en 2000 (cf. tableau 2 de Renaud, 2007) à 0.45%. Pour les bâtiments et les logements, nous rappelons que, pour les raisons discutéees à la section 5, les estimations fournies ne concernent que la strate des carrés classiques. On constate que la composante de sur-couverture est plus importante que la composante de sous-couverture, ce qui conduit à une sous-couverture nette négative. Ces deux composantes restent dans des ordres de grandeur satisfaisants. Des résultats plus détaillés peuvent être consultés dans Fritschi (2014).

Tableau 4 Estimations des taux de couverture au niveau national exprimées en pourcentage, accompagnées et leur précision

	Sous-couverture	Sur-couverture	Sous-couverture nette
Personnes	0.47 (0.04)	0.02	0.45 (0.04)
Bâtiments	0.18 (0.05)	0.71 (0.11)	-0.53 (0.12)
Logements	0.90 (0.13)	1.25 (0.12)	-0.36 (0.18)

Lors de l'exploitation des résultats, nous avons également estimé le gain de variance dû à l'emploi d'un plan de sondage équilibré pour les zones géographiques (cf. section 3). A cet effet, nous avons estimé la variance qui aurait été obtenue si le plan de sondage des zones géographiques avait été un plan de Poisson, avec les mêmes probabilités d'inclusion que celles utilisées. Il s'avère que le gain de précision est très faible. Le gain relatif sur l'écart-type est par exemple d'environ 2% pour l'estimation de la sous-couverture des bâtiments au niveau national. Cela s'explique par le fait que, comme le montre la formule (28), la variance estimée sous le plan équilibré dépend de la qualité de l'ajustement dans la régression des $\hat{\ell}_g/\pi_g$ sur les variables d'équilibrage divisées par π_g. Or le pouvoir explicatif de cette régression est très faible, le coefficient de détermination R^2 étant seulement de 5%. En pratique, cette faible corrélation pourrait être expliquée par les éléments suivants. La linéarisée $\hat{\ell}_g$ définie par (25) peut être ré-écrite sous la forme :

$$\hat{\ell}_g^- = \frac{1}{\hat{x}_{+1}}\left(x_{1+}^g - f_{db}^- x_{+1}^g\right) \tag{75}$$

$$= \frac{1}{\hat{x}_{+1}}\left(x_{1+}^g - x_{+1}^g + x_{+1}^g - f_{db}^- x_{+1}^g\right) \tag{76}$$

$$= \frac{1}{\hat{x}_{+1}}\left[\left(x_{1+}^g - x_{+1}^g\right) + x_{+1}^g(1-f_{db}^-)\right]. \tag{77}$$

Puisque l'estimation du taux de sous-couverture $(1 - f_{db}^{-})$ au niveau national pour les bâtiments est de 0.18% (cf. tableau 4), et que $x_{+1}^{g} = 81$ dans le carré dans lequel le plus grand nombre de bâtiments effectivement présents sur le terrain a été démombré, le terme $x_{+1}^{g}(1 - f_{db}^{-})$ intervenant dans (77) est au maximum égal à $0.0018 \cdot 81 \simeq 0.15$, et est souvent bien plus faible. Le terme $\left(x_{1+}^{g} - x_{+1}^{g}\right)$ est quant à lui égal au nombre de bâtiments en sous-couverture dans la zone g. Or les bâtiments en sous-couverture sont apparus aussi bien dans des zones comportant peu de bâtiments d'après nos données que dans des zones en comportant beaucoup. Plus généralement, le nombre de bâtiments en sous-couverture dans une zone donnée est donc peu corrélé aux variables auxiliaires prises en compte dans l'équilibrage.

10 Conclusions

L'OFS a mis en œuvre une enquête complexe aussi bien du point de vue méthodologique que de celui de l'organisation sur le terrain. L'enquête a permis d'évaluer la qualité de la couverture des décomptes de bâtiments, logements et personnes que nous publions. Contrairement à ce qui est souvent pratiqué pour les enquêtes de couverture, nous ne nous sommes pas basés sur un bootstrap pour estimer la précision des estimations, mais avons développé une formule adaptée au plan de sondage et peu coûteuse en temps de calcul. La formule a été obtenue grâce à des techniques de linéarisation ainsi que sur la base de l'article de Deville and Tillé (2005), qui permet de calculer la variance pour des plan équilibrés. On peut remarquer que la précision estimée sur la base de notre formule est du même ordre de grandeur que celle qui avait été estimée pour l'année 2000 grâce à un jacknife (cf. Renaud, 2007, pour plus de précisions sur cette procédure).
Les résultats de l'enquête de couverture montrent que le nouveau système de recensement de la population basé sur des données administratives est de très bonne qualité. Les défauts de couverture sont encore réduits par rapport à 2000. En ce qui concerne la statistique des bâtiments et des logements, l'enquête a permis d'en évaluer la couverture dans la strate des carrés classiques. Au niveau national, la sur-couverture est plus prononcée que la sous-couverture. Ces deux composantes restent dans des ordres de grandeur satisfaisants qui montrent la qualité de la couverture de la StatBL dans la strate des carrés classiques.

Annexes

A Méthode duale et plan complexe

Nous montrons dans cette annexe comment la méthode proposée par Wolter (1986) peut être généralisée à notre plan de sondage. Dans notre cas, comme cela est explicité dans la section 3, les zones géographiques g sont sélectionnées avec des probabilités de sélection π_g qui dépendent du nombre d'unités présentes dans les zones selon la liste A. Dans le cas de la ré-énumération des personnes, il y a de plus une phase supplémentaire de non-réponse. Puisque les probabilités de sélection des zones géographiques dépendent du nombre d'unités présentes dans les zones selon la liste A, le plan de sondage est donc défini conditionnellement à la réalisation de la liste A. En revanche la sélection des zones géographiques ainsi que le mécanisme de réponse sont supposés indépendants du mécanisme aléatoire qui conduit à la liste B (inconnue lors de l'échantillonnage).
L'espérance totale $\mathbf{E}(\cdot)$ sous l'aléa du modèle de Petersen et du plan de sondage est définie par :

$$\mathbf{E}(\cdot) = \mathbf{E}_m \mathbf{E}_p(\cdot \mid m), \tag{78}$$

où $\mathbf{E}_p(\cdot \mid m)$ est l'espérance sous l'aléa de la sélection des zones géographiques et du mécanisme de réponse, conditionnellement aux réalisations du modèle de Petersen. Puisque le plan de sondage ne dépend que de la réalisation de la liste A et est indépendant du mécanisme aléatoire qui conduit à la liste B, l'équation (78) peut se ré-écrire sous la forme :

$$\mathbf{E}(\cdot) = \mathbf{E}_m \mathbf{E}_p(\cdot \mid m) = \mathbf{E}_m \mathbf{E}_p(\cdot \mid A), \tag{79}$$

où $\mathbf{E}_p(\cdot \mid A)$ est l'espérance sous le plan conditionnellement aux réalisations de la liste A.
Les quantités $\widehat{x}_{11}$ et $\widehat{x}_{+1}$ utilisées dans la section 2 peuvent être exprimées sous la forme :

$$\widehat{x}_{+1} = \sum_{i \in \mathcal{U}} \frac{1}{\pi_i} \frac{1}{p_i} \mathbf{1}_{\{i \in s\}} \mathbf{1}_{\{i \in B\}} \mathbf{1}_{\{i \in r\}}, \quad \widehat{x}_{11} = \sum_{i \in \mathcal{U}} \frac{1}{\pi_i} \frac{1}{p_i} \mathbf{1}_{\{i \in s\}} \mathbf{1}_{\{i \in B\}} \mathbf{1}_{\{i \in r\}} \mathbf{1}_{\{i \in A\}},$$

où s désigne l'échantillon d'unités de la population $\mathcal{U}$ présentes dans les zones sélectionnées, r désigne l'échantillon de répondants, $\pi_i = \pi_g$ pour toute unité i de la zone g, et p_i est sa probabilité de réponse. On remarque que :

$$\begin{aligned}
\mathbf{E}_p\left(\widehat{x}_{11} \mid A\right) &= \mathbf{E}_p\left(\widehat{x}_{11} \mid m\right) \\
&= \sum_{i \in \mathcal{U}} \frac{1}{\pi_i} \frac{1}{p_i} \mathbf{E}_p\left(\mathbf{1}_{\{i \in s\}} \mathbf{1}_{\{i \in B\}} \mathbf{1}_{\{i \in r\}} \mathbf{1}_{\{i \in A\}} \mid m\right) \\
&= \sum_{i \in \mathcal{U}} \frac{1}{\pi_i} \frac{1}{p_i} \mathbf{1}_{\{i \in A\}} \mathbf{1}_{\{i \in B\}} \mathbf{E}_p\left(\mathbf{1}_{\{i \in s\}} \mathbf{1}_{\{i \in r\}} \mid m\right) \\
&= \sum_{i \in \mathcal{U}} \frac{1}{\pi_i} \frac{1}{p_i} \mathbf{1}_{\{i \in A\}} \mathbf{1}_{\{i \in B\}} \pi_i p_i \\
&= \sum_{i \in \mathcal{U}} \mathbf{1}_{\{i \in A\}} \mathbf{1}_{\{i \in B\}} \\
&= x_{11}.
\end{aligned}$$

De même :

$$\mathbf{E}_p\left(\widehat{x}_{+1} \mid A\right) = x_{+1}.$$

Il en découle qu'il suffit de remplacer $\mathbf{E}_p(\cdot)$ par $\mathbf{E}_p(\cdot \mid A)$ dans les formules (7) et (8), et que la méthode duale reste valide pour ce plan de sondage.

B Estimation de la probabilité qu'un ménage participe à l'interview, sachant que l'on est parvenu à le contacter

Comme le montre les équations (33), (34) et (35), nous devons, pour les besoins de la pondération, estimer la probabilité p_ℓ^1 qu'un ménage participe à l'interview, sachant que l'on est parvenu à le contacter. Concrètement, l'estimation de p_ℓ^1 comprend trois étapes.

1. Nous commençons par estimer p_ℓ^1 pour les logements pour lesquels un contact a été établi par internet. Parmi ces logements, le taux de réponse est très élevé malgré quelques abandons ou interviews éliminées lors du traitement des données pour cause de trop nombreuses incohérences. La probabilité p_ℓ^1 est simplement estimée par la proportion de logements pour lesquels on dispose d'une interview parmi l'ensemble des logements pour lesquels un contact a été établi par internet, qu'il s'agisse ou non de la résidence principale d'une personne cible.
2. Pour les logements pour lesquels un contact a été établi par d'autres canaux qu'internet, nous commençons par décomposer p_ℓ^1 de la façon suivante :

$$p_\ell^1 = \underbrace{\mathrm{P}\left(\ell \in s_L^{int} \mid s \in s_L^{info}\right)}_{p_\ell^{11}} \underbrace{\mathrm{P}\left(s \in s_L^{info} \mid s \in s_L^{cont}\right)}_{p_\ell^{12}}, \tag{80}$$

   où $\in s_L^{info}$ désigne l'échantillon de logements pour lesquels le nombre de personnes dans le ménage est connu. Il s'agit donc des logements pour lesquels on dispose soit d'une interview, soit de l'enquête de non-réponse. Dans cette deuxième étape, on estime p_ℓ^{11} grâce à une régression logistique sur l'ensemble des logements de s_L^{info}, qu'il s'agisse ou non de la résidence principale d'une personne cible. Les quatre variables auxiliaires mentionnées page 19, ainsi que leurs interactions, sont testées dans la régression logistique. Les effets retenus sont sélectionnés grâce à l'option "selection=backward" de la "proc logistic" de SAS. Le taux de couples concordants calculés par la "proc logistic" de SAS est donné dans le tableau 5 (voir Hosmer and Lemeshow, 2000, page 163 pour plus de détails sur cette notion).

Tableau 5 Taux de couples concordants selon la proc logistic de sas

Taux de couples concordants	67.3%
Taux de couples discordants	29.6%
Taux de couples "tied"	3.2 %

3. Dans la troisième étape, on estime p_ℓ^{12} grâce à une régression logistique sur l'ensemble des logements de s_L^{cont}, qu'il s'agisse ou non de la résidence principale d'une personne cible. Lors de cette troisième étape, seules trois des informations auxiliaires mentionnées page 19 peuvent être utilisées, puique le nombre de personnes dans le ménage n'est pas connu pour tous les logements de s_L^{cont}. Ces trois variables auxiliaires, ainsi que leurs interactions, sont testées dans la régression logistique. Les effets retenus sont sélectionnés grâce à l'option "selection=backward" de la "proc logistic" de SAS. Le taux de couples concordants calculés par la "proc logistic" de SAS est donné dans le tableau 6.

Tableau 6 Taux de couples concordants selon la proc logistic de sas

Taux de couples concordants	67.4%
Taux de couples discordants	29.7%
Taux de couples “tied”	3.0 %

C Estimation de la probabilité de contact des ménages

Comme le montre les équations (33), (34) et (35), nous devons, pour les besoins de la pondération, estimer la probabilité p_ℓ^2 que l'on soit parvenu à contacter le logement ℓ. La formule (36) sur laquelle nous nous basons pour estimer cette probabilité de contact nécessite de disposer d'une estimation du nombre $N_{L,gr}^{\overline{cont}}$ de logements que l'on n'a pas réussi à contacter parmi les logements qui sont la résidence principale d'au moins une personne cible. Afin d'obtenir une approximation de $N_{L,gr}^{\overline{cont}}$, nous énumérons, pour chaque grande région gr, les logements de l'échantillon de zones géographiques qui sont la résidence principale d'une personne cible d'après nos données administratives, et que l'on n'a par ailleurs pas réussi à contacter. Cette liste est entâchée de sur- et de sous-couverture. Nous décrivons dans cette annexe comment les défauts de couverture de cette liste, qui joue le rôle de la liste A mentionnée dans la section 2, ont été corrigés.

Nous pouvons distinguer trois type d'erreurs de sur-couverture sur les logements :

– ceux qui n'existent pas ou plus sur le terrain,
– ceux en réalité totalement inoccupés, ni en tant que résidence principale ni en tant que résidence secondaire,
– ceux qui sont effectivement occupés mais qui sont des résidences secondaires ou des résidences principales de personnes qui ne font pas partie de la population cible.

Les erreurs de sur-couverture relevant des deux premiers types peuvent être identifiées, du moins en grande partie, grâce aux informations relevées par les enquêteurs sur le terrain (cf. section 4). Ces erreurs de sur-couverture sont retirées de la liste A. En ce qui concerne le troisième type d'erreur, il n'est possible de les identifier que dans le cas où une interview a été réalisée. Le même type de problème se pose pour évaluer la composante de sous-couverture. Il faudrait en effet être en mesure de dresser la liste des logements que l'on n'a pas réussi à contacter et qui sont bien la résidence principale d'au moins une personne cible. Or cette information ne peut être connue qu'en cas d'interview. Pour chaque grande région gr, nous calculons donc un facteur de correction f_{gr} pour la sous-couverture nette sur la base des logements pour lesquels une interview est disponible. Il est égal au rapport entre le nombre de logements qui sont bien sur le terrain la résidence principale d'une personne cible et le nombre de logements qui sont la résidence principale d'une personne cible selon nos données administratives. Le facteur de correction f_{gr} calculé sur la base des logements pour lesquels une interview est disponible fournit une approximation de celui qui s'applique à la liste A. Le facteur de correction f_{gr} est donc appliqué au nombre de logements de la liste A pour obtenir une estimation de $N_{L,gr}^{\overline{cont}}$.

Bibliographie

Deville, J.-C. (1999). Estimation de variance pour des statistiques et des estimateurs complexes : techniques de résidus et de linéarisation. *Techniques d'enquête 25*, 219–230.

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

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

Fritschi, R. (2014). Nouveau système de recensement, enquête de qualité. *Actualités OFS, Office fédéral de la statistique, Neuchâtel*. https://www.bfs.admin.ch/bfsstatic/dam/assets/349751/master,.

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

Hosmer, D. W. and S. Lemeshow (2000). *Applied logistic regression*. New York: Wiley.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady 10(8)*, 707–710.

Office for national statistics (2013). An assessment of the quality of the matching between the 2011 census and the census coverage survey. Technical report. http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-census-user-guide/quality-and-methods/methods/coverage-assessment-and-adjustment-methods/census-coverage-survey--ccs-/matching-quality-report-ns.pdf.

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, Office fédéral de la statistique, Neuchâtel*.

Renaud, A. (2004). Coverage estimation of the Swiss population census 2000. Estimation methodology and results. *Rapport de méthodes, Office fédéral de la statistique, Neuchâtel*.

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.

Renaud, A. and P. Eichenberger (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, Office fédéral de la statistique, Neuchâtel*.

Renaud, A. and J. Potterat (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 sondages pour les bâtiments. *Rapport de méthodes, Office fédéral de la statistique, Neuchâtel*.

Sanchez, P., A. Wakim, and D. Cronkite (2012). Results of the 2010 census coverage measurement field and matching operations. In *Proceedings of Joint Statistical Meetings (San Diego, July 28-August2 2012)*. http://www.amstat.org/sections/srms/proceedings/y2012/files/304305_72898.pdf.

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

Methodenberichte der Sektion Statistische Methoden des BFS
Rapports de méthodes de la section méthodes statistiques de l'OFS
Methodology reports published by the FSO's Statistical Methods Section

Massiani, A. (2017). Estimation de la couverture du nouveau recensement en Suisse en 2012. Numéro de commande : 338-0078

Nedyalkova, D., Assoulin, D. (2017). Construction of full-time equivalents for the Swiss structural business statistics 2011. Order number: 338-0077

Ferster, M. (2016). Energieverbrauchsstatistik EVS2014 - Stichprobe, Hochrechnung und Vergleichbarkeit mit der EVS2013. Bestellnummer: 338-0076

Panchard, C. (2014). Enquête sur la participation aux activités culturelles 2008. Plan de sondage et estimations. Bestellnummer: 338-0073

Assoulin, D. (2014). Wertschöpfungsstatistik. Revision 2009: Statistische Datenaufbereitung und Hochrechnung. Bestellnummer: 338-0072

Wilhelm, M. (2014). Echantillonnage boule de neige. La méthode de sondage déterminé par les répondants. Numéro de commande: 338-0071

Assoulin, D. (2013). Wertschöpfungsstatistik. Revision 2009: Stichprobenrahmen und Stichprobenplan. Bestellnummer: 338-0070

Ferster, M. (2013). EVS I - Energieverbrauchsstatistik 2002 bis 2007: Stichprobenplan und Hochrechnung. Bestellnummer: 338-0069

Assoulin, D. (2013). Zusatzerhebung für die landwirtschaftliche Betriebszählung 2010: Stichprobenplan und Hochrechnung. Bestellnummer: 338-0068

Potterat, J. (2012). Use of conversion keys for NOGA2002-2008. Order number: 338-0067-05

Andrade, B., Salamin P.-A. (2012). Enquête sur la situation sociale et économique des étudiant-e-s des hautes écoles suisses 2009. Cadre de sondage, plan d'échantillonnage et méthodes d'estimation. Numéro de commande: 338-0066

Potterat, J., Panchard, C., Kilchmann, D. (2012). Umweltschutzausgaben der Unternehmen 2009 (UWSA2009). Stichprobenplan, Einsetzungen, Gewichtung und Schätzverfahren. Bestellnummer: 338-0065

Potterat, J. (2012). Benutzung der Umsteigeschlüssel NOGA 2002-2008. Bestellnummer: 338-0064

Potterat, J. (2011). Kosten und Nutzen der Berufsbildung aus Sicht der Betriebe im Jahr 2009 (KNBB09). Stichprobenplan, Gewichtung und Schätzverfahren. Bestellnummer: 338-0063

Kilchmann, D., Potterat, J., Genoud, S. (2011). Gütertransporterhebung 2008. Stichprobenplan, Datenaufbereitung, Gewichtung und Schätzverfahren. Bestellnummer: 338-0062

Graf, E. (2010). Enquête suisse sur la santé 2007. Plan d'échantillonnage, pondérations et analyses pondérées des données. Numéro de commande: 338-0061

Eichenberger P., Hulliger B., Potterat J. (2010). Describing the Anticipated Accuracy of the Swiss Population Survey. Order number: 338-0060

Graf, E. (2010). Étude empirique de l'attrition du Panel Suisse de Ménages : vers une caractérisation du profil du non-répondant. Numéro de commande: 338-0059

Graf, E. (2009). Weightings of the Swiss Household Panel: SHP_I wave 9, SHP_II wave 4, SHP_I et SHP_II combined. Order number: 338-0058

Graf, E. (2009). Pondérations du Panel Suisse de Ménages: PSM_I vague 9, PSM_II vague 4, PSM_I et PSM_II combinés. Numéro de commande: 338-0057-05

Qualité, L., Tillé, Y. (2009). Estimation de la précision d'évolutions dans l'enquête sur la valeur ajoutée. Numéro de commande: 338-0056

Renaud, A., Panchard, C. et Potterat, J. (2008). Statistique de l'emploi. Révision 2007 : méthodes d'estimation. Numéro de commande: 338-0055

Graf, E. (2008). Pondérations du PSM. PSM_I vague 8, PSM_II vague 3, PSM_I et PSM_II combinés. Numéro de commande: 338-0054

Andrade, B., Graf, M. (2008). Enquête suisse sur la structure des salaires 2006. Aspects méthodologiques du modèle des salaires "Salarium". Numéro de commande: 338-0053

Renaud, A. (2008). Statistique de l'emploi. Révision 2007 : cadre de sondage et échantillonage. Numéro de commande: 338-0052

Graf, E. (2008). Pondérations du SILC pilote. SILC_I vague 2, SILC_II vague 1, SILC_I et SILC_II combinés. Numéro de commande: 338-0051

Kilchmann, D. (2008). Statistik der sozialmedizinischen Institutionen 1999-2004 und Krankenhausstatistik 1999-2002. Einsetzungen für fehlende Daten. Bestellnummer: 338-0050

Renaud, A. (2008). Technologies de l'information et de la communication. Estimations sur la base de la statistique de la valeur ajoutée. Numéro de commande: 338-0049

Assoulin, D. (2007). Wertschöpfungsstatistik. Einsetzungsversuche für fehlende Antworten grosser Unternehmen. Bestellnummer: 338-0048

Kilchmann, D. (2007). Beherbergungsstatistik Campingplätze. Stichprobenrahmen und Schätzverfahren 2005/06. Bestellnummer: 338-0047

Gabler, S., Häder, S. (2007). Haushalts- und Personenerhebungen. Machbarkeit von Random Digit Dialing in der Schweiz. Bestellnummer: 338-0046

Ferrez, J., Graf, M. (2007). Enquête suisse sur la structure des salaires. Programmes R pour l'intervalle de confiance de la médiane. Numéro de commande: 338-0045

Renaud, A. (2007). Harmonisation de la scolarité obligatoire en Suisse (HarmoS). Design général de l'enquête et échantillon des écoles. Numéro de commande: 338-0044

Potterat, J. (2007). Betriebszählung 2005. Statistische Methoden zur Schätzung der provisorischen Ergebnisse. Bestellnummer: 338-0043

Hulliger, B. (2006). Umweltschutzausgaben der Unternehmen 2003, Stichprobenplan, Datenaufbereitung und Schätzverfahren. Bestellnummer: 338-0042

Renfer, J.-P. (2006). Enquête sur les chiffres d'affaires du commerce de détail. Plan d'échantillonnage et méthodes d'estimation. Numéro de commande: 338-0041

Salamin, P.-A. (2006). Statistique de l'aide sociale dans le domaine de l'asile. Plan de sondage et extrapolations pour l'enquête pilote 2005. Numéro de commande: 338-0040

Renaud, A. (2006). Statistique suisse des bénéficiaires de l'aide sociale. Pondération des communes 2004. Numéro de commande: 338-0039

Graf, M. (2006). Swiss Earnings Structure Survey 2002-2004. Compositional data in a stratified two-stage sample: Analysis and precision assessment of wage components. Order number: 338-0038

Potterat, J. (2006). Pensionskassenstatistik 2004. Statistische Methoden zur Schätzung der provisorischen Ergebnisse. Bestellnummer: 338-0037

Potterat, J. (2006). Kosten und Nutzen der Berufsbildung aus Sicht der Betriebe im Jahr 2004. Stichprobenplan, Gewichtung und Schätzverfahren. Bestellnummer: 338-0036

Kilchmann, D. (2006). Vierteljährliche Wohnbaustatistik. Stichprobenplan, statistische Datenaufarbeitung und Schätzverfahren 2005. Bestellnummer: 338-0035

Kilchmann, D. (2006). Erhebung über Forschung und Entwicklung in der schweizerischen Privatwirtschaft 2004. Bereinigung der Stichprobe, Ersatz fehlender Werte und Schätzverfahren. Bestellnummer: 338-0034

Kilchmann, D., Eichenberger, P., Potterat, J. (2005). Volkszählung 2000. Statistische Einsetzungsverfahren Band 2. Bestellnummer: 338-0033

Kilchmann, D., Eichenberger, P., Potterat, J. (2005). Volkszählung 2000. Statistische Einsetzungsverfahren Band 1. Bestellnummer: 338-0032

Graf, M., Matei, A. (2005). Enquête suisse sur la structure des salaires 2002. La précision du salaire brut standardisé médian. Numéro de commande: 338-0031

Graf, E., Renfer, J.-P. (2005). Enquête suisse sur la santé 2002. Plan d'échantillonnage, pondération et estimation de la précision. Numéro de commande: 338-0030

Potterat, J. (2005). Mietpreis-Strukturerhebung 2003. Gewichtung und Schätzverfahren. Bestellnummer: 338-0029

Potterat, J. (2005). Landwirtschaftliche Betriebszählung 2003. Schätzverfahren für die Zusatzerhebung. Bestellnummer: 338-0028

Renaud, A. (2004). Coverage estimation for the Swiss population census 2000. Estimation methodology and results. Order number: 338-0027

Kilchmann, D. (2004). Revision des Schweizerischen Lohnindex. Schätzmethoden der Lohnindices und deren Varianzschätzer. Bestellnummer: 338-0026

Graf, M. (2004). Enquête suisse sur la structure des salaires 2002. Plan d'échantillonnage et extrapolation pour le secteur privé. Numéro de commande: 338-0025

Renaud, A. (2004). Analyse de données d'enquêtes. Quelques méthodes et illustration avec des données de l'OFS. Numéro de commande 338-0024

Renaud, A., 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. Numéro de commande: 338-0023

Graf, M. (2004). Fusion de données. Etude de faisabilité. Numéro de commande: 338-0022

Potterat, J. (2003). Mietpreis-Strukturerhebung 2003. Entwicklung des Stichprobenplans und Ziehung der Stichprobe. Bestellnummer: 338-0021

Potterat, J. (2003). Landwirtschaftliche Betriebszählung 2003. Stichprobenplan der Zusatzerhebung. Bestellnummer: 338-0020.

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). Numéro de commande: 338-0019

Hulliger, B. (2003). Bereinigung der Stichprobe, Ersatz fehlender Werte und Schätzverfahren. Erhebung über F+E in der schweizerischen Privatwirtschaft 2000. Bestellnummer: 338-0018

Renfer, J.-P. (2003). Enquête 2000 sur la recherche et le développement dans l'économie privée en Suisse. Plan d'échantillonnage. Numéro de commande: 338-0017

Potterat, J. (2003). Kosten und Nutzen der Berufsbildung aus Sicht der Betriebe. Schätzverfahren. Bestellnummer: 338-0016

Graf, M., Matei, A. (2003). Stratégie de choix des modèles de désaisonnalisation. Application aux séries de l'emploi total. Numéro de commande: 338-0015

Potterat, J., Salamin, P.A. (2002). Betriebszählung 2001. Methoden für die Datenbereinigung. Bestellnummer: 338-0014

Renaud, A. (2002). Programme international pour le suivi des acquis des é lèves (PISA). Plans d'échantillonnage pour PISA 2000 en Suisse. Numéro de commande: 338-0013

Renfer, J.-P. (2002). Enquête 2001 sur les coûts et l'utilité de la formation des apprentis du point de vue des établissements. Plan d' échantillonnage. Numéro de commande: 338-0012

Potterat, J., Salamin, P.A. (2002). Betriebszählung 2001. Stichprobenplan und Schätzverfahren für die provisorischen Ergebnisse. Bestellnummer: 338-0011

Graf, M. (2002). Enquête suisse sur la structure des salaires 2000. Plan d'échantillonnage, pondération et méthode d'estimation pour le secteur privé. Numéro de commande: 338-0010

Renaud, A., 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. Numéro de commande: 338-0009

Kilchmann, D., Hulliger, B. (2002). Stichprobenplan für die Obstbaumzählung 2001. Bestellnummer: 338-0008

Graf, M. (2002). Passage du concept établissement au concept entreprise. Numéro de commande: 338-0007

Salamin, P.A. (2001). La technique de la double enquête pour la statistique du transport routier de marchandise. Numéro de commande: 338-0006

Peters, R., Renfer, J.-P. et Hulliger, B. (2001). Statistique de la valeur ajoutée 1997-1998. Procédure d'extrapolation des données. Numéro de commande: 338-0005

Potterat, J., Hulliger, B. (2001). Schätzung der Sägereiproduktion mit der Sägerei-Erhebung PAUL. Bestellnummer: 338-0004

Graf, M. (2001). Désaisonnalisation. Aspects méthodologiques et application à la statistique de l'emploi. Numéro de commande: 338-0003

Hüsler, J., Müller, S. (2001). Schlussbericht Betriebszählung 1995 (BZ 95), Mehrfach imputierte Umsatzzahlen. Bestellnummer: 338-0002

Renaud, A. (2001). Statistique suisse des bénéficiaires de l'aide sociale. Plan d'échantillonnage des communes. Numéro de commande: 338-0001

Hulliger, B., Eichenberger, P. (2000). Stichprobenregister für Haushalterhebungen: Umstellung auf Telefonnummern ohne Namen und Adressen, Abläufe für Erstellung und Stichprobenziehung. Bestellnummer: 338-0000

Programme des publications de l'OFS

En tant que service statistique central de la Confédération, l'Office fédéral de la statistique (OFS) a pour tâche de rendre les informations statistiques accessibles à un large public. Il utilise plusieurs moyens et canaux pour diffuser ses informations statistiques par thème.

Les domaines statistiques

00 Bases statistiques et généralités
01 Population
02 Espace et environnement
03 Travail et rémunération
04 Économie nationale
05 Prix
06 Industrie et services
07 Agriculture et sylviculture
08 Énergie
09 Construction et logement
10 Tourisme
11 Mobilité et transports
12 Monnaie, banques, assurances
13 Protection sociale
14 Santé
15 Éducation et science
16 Culture, médias, société de l'information, sport
17 Politique
18 Administration et finances publiques
19 Criminalité et droit pénal
20 Situation économique et sociale de la population
21 Développement durable, disparités régionales et internationales

Les principales publications générales

L'Annuaire statistique de la Suisse

L'Annuaire statistique de la Suisse de l'OFS constitue depuis 1891 l'ouvrage de référence de la statistique suisse. Il englobe les principaux résultats statistiques concernant la population, la société, l'État, l'économie et l'environnement de la Suisse.

Le Mémento statistique de la Suisse

Le mémento statistique résume de manière concise et attrayante les principaux chiffres de l'année. Cette publication gratuite de 52 pages au format A6/5 est disponible en cinq langues (français, allemand, italien, romanche et anglais).

Le site Internet de l'OFS: www.statistique.ch

Le portail «Statistique suisse» est un outil moderne et attrayant vous permettant d'accéder aux informations statistiques actuelles. Nous attirons ci-après votre attention sur les offres les plus prisées.

La banque de données des publications pour des informations détaillées

Presque tous les documents publiés par l'OFS sont disponibles gratuitement sous forme électronique sur le portail Statistique suisse (www.statistique.ch). Pour obtenir des publications imprimées, vous pouvez passer commande par téléphone (058 463 60 60) ou par e-mail (order@bfs.admin.ch).
www.statistique.ch → Trouver des statistiques → Catalogues et banques de données → Publications

Vous souhaitez être parmi les premiers informés?

Abonnez-vous à un Newsmail et vous recevrez par e-mail des informations sur les résultats les plus récents et les activités actuelles concernant le thème de votre choix.
www.news-stat.admin.ch

STAT-TAB: la banque de données statistiques interactive

La banque de données statistiques interactive vous permet d'accéder simplement aux résultats statistiques dont vous avez besoin et de les télécharger dans différents formats.
www.stattab.bfs.admin.ch

Statatlas Suisse: la banque de données régionale avec ses cartes interactives

L'atlas statistique de la Suisse, qui compte plus de 3 000 cartes, est un outil moderne donnant une vue d'ensemble des thématiques régionales traitées en Suisse dans les différents domaines de la statistique publique.
www.statatlas-suisse.admin.ch

Pour plus d'informations

Service de renseignements statistiques de l'OFS
058 463 60 11, info@bfs.admin.ch

Depuis 2010, 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 couverture, ainsi que celle du registre des bâtiments et des logements. À cet effet, l'OFS a mené début 2013 une enquête de couverture sur la base d'un échantillon de zones géographiques au sein desquelles les bâtiments, logements et habitants sont ré-énumérés. Les données relevées lors de l'enquête sont alors comparées aux données de l'OFS afin de calculer des taux de sur- et de sous-couverture pour les résultats relatifs à l'année 2012. Les résultats montrent que le nouveau système est de très bonne qualité. Ce document décrit les aspects méthodologiques de cette enquête complexe. Les thèmes couverts concernent notamment le plan d'échantillonnage, la pondération, ainsi que les méthodes d'estimation utilisées.

Commandes d'imprimés
Tél. 058 463 60 60
Fax 058 463 60 61
order@bfs.admin.ch

Prix
Gratuit

Téléchargement
www.statistique.ch (gratuit)

Numéro OFS
338-0078

ISBN
978-3-303-00570-5