Nations Unies ECE/CES/GE.41/2019/15

Conseil économique et social

Distr. générale
10 juillet 2019
Français
Original : anglais

Commission économique pour l'Europe

Conférence des statisticiens européens

Groupe d'experts des recensements de la population et des habitations

Vingt et unième réunion
Genève, 18-20 septembre 2019
Point 2 de l'ordre du jour provisoire
Résultats des essais menés, en ce qui concerne les méthodes, les techniques, la participation et d'autres aspects

Méthode d'estimation pour petites régions visant à corriger les erreurs de mesure dans les grands registres de la population avec application au recensement de la population en Israël

Note du Bureau central de statistique d'Israël*

Résumé

Comme nombre d'autres pays, Israël dispose d'un registre de la population assez précis au niveau national, où sont enregistrées environ 9 millions de personnes. Toutefois, le registre est beaucoup moins précis pour les petits domaines, pour lesquels l'erreur de dénombrement s'élève en moyenne à environ 13 %, principalement parce que les personnes qui quittent une région ou viennent s'y établir tardent souvent à déclarer leur changement d'adresse. La présente note traite des moyens de corriger les estimations sur petits domaines au moyen d'une enquête réalisée à partir du registre. Elle aborde en particulier les questions suivantes :

a) Quelle est la meilleure façon de prélever l'échantillon ?

b) Comment combiner les données d'échantillonnage avec celles du registre ?

c) Comment corriger la non-réponse ne manquant pas au hasard ?

d) Comment évaluer l'erreur quadratique moyenne des estimateurs du recensement qui en découlent ?

On trouvera des illustrations empiriques fondées sur les données du recensement de 2008.

* Document établi par Danny Pfeffermann, Dan Ben-Hur et Olivia Blum (voir aussi Pfeffermann *et al*, *Statistics in Transition (SiT)*, vol. 20, nº 1, Polish Statistical Association et Statistics Poland).

GE.19-11744 (F) 250719 260719
1911744

I. Introduction

1. Dans le présent document figure une proposition de nouvelle méthode de recensement fondée à la fois sur une enquête et sur l'exploitation de données administratives. On y examine de nouvelles façons d'intégrer les données d'enquête et les données administratives afin d'obtenir une unique estimation du recensement dans de petites régions géographiques qui tienne compte des erreurs présentes dans les deux sources de données et de la non-réponse ne manquant pas au hasard. La méthode proposée est illustrée à l'aide des données du recensement réalisé en Israël en 2008.

A. Description du dernier recensement réalisé en Israël (2008)

2. Israël dispose d'un registre central de la population relativement précis et presque parfaitement fiable au niveau national. Toutefois, ce registre est beaucoup moins précis pour les petites régions statistiques, dans lesquelles l'erreur de dénombrement s'élève en moyenne à 13 % et le 95e percentile à 40 %. Le territoire national est divisé en environ 3 000 régions statistiques, pour lesquelles des données de recensement telles que des chiffres et des renseignements socioéconomiques doivent être systématiquement fournis. L'inexactitude des données du registre au niveau des régions tient au fait que les personnes qui quittent une région ou qui viennent s'y établir tardent souvent à déclarer leur changement d'adresse, tandis que celles qui ont intérêt à conserver une adresse en particulier (que ce soit pour des considérations liées par exemple à des avantages fiscaux, à la carte scolaire ou au stationnement) ne signalent pas leur déménagement. En 2008, le Bureau central de statistique a procédé à un recensement intégré sur la base du registre de la population, corrigé des estimations obtenues à partir de deux échantillons de couverture pour chaque région : un échantillon (prélevé sur le terrain) des personnes vivant dans la région au moment du dénombrement, qui sert à estimer le sous-dénombrement du registre (échantillon U), et un échantillon de personnes enregistrées dans la même région pour estimer le sur-dénombrement du registre (échantillon O). L'échantillon U a également été utilisé pour la collecte d'informations socioéconomiques.

3. L'estimation finale du recensement a été calculée selon la formule ci-après, dans laquelle N_i et K_i désignent respectivement le nombre réel et le nombre théorique de personnes résidant dans la région i au moment du dénombrement. $p_{i,L|R}$ représente le nombre de personnes vivant dans la région i rapporté au nombre de personnes enregistrées, et $p_{i,R|L}$ le nombre de personnes enregistrées comme vivant dans la région i rapporté au nombre de personnes y vivant réellement :

$$N_i \times p_{i,R|L} = K_i \times p_{i,L|R} \Rightarrow \widehat{N}_i = K_i \times \frac{\hat{p}_{i,L|R}}{\hat{p}_{i,R|L}} \quad (1)$$

4. En utilisant un développement de Taylor, la variance conditionnelle (basée sur le plan de sondage) de $\widehat{N}_i$ peut être approchée comme suit :

$$Var\left(\widehat{N}_i \middle| K_i\right) \cong K_i^2 \left[\frac{Var(\hat{p}_{i,L|K})}{[E(\hat{p}_{i,K|L})]^2} + \frac{[E(\hat{p}_{i,L|K})]^2}{[E(\hat{p}_{i,K|L})]^4} \times Var(\hat{p}_{i,K|L})\right] \quad (2)$$

5. Au cours de la dernière décennie, Israël, comme nombre d'autres pays, a de plus en plus utilisé des données administratives pour la production de statistiques officielles en général, et il a en particulier amélioré sa capacité à les utiliser à des fins de recensement. En conséquence, la méthode du recensement de 2020 s'appuiera sur de nouvelles sources de données et, pour la première fois, un fichier administratif géodémographique servira de base de sondage pour un échantillon qui servira à corriger les données administratives portant sur de petites régions statistiques.

6. Le changement prévu de méthode repose sur deux hypothèses clefs : a) les entrées dans le pays et les sorties du territoire sont correctement enregistrées ; b) l'ensemble des personnes présentes dans le pays sont inscrites dans les fichiers administratifs ; les citoyens sont inscrits dans le registre central de la population et les étrangers figurent dans des registres thématiques tels que celui des permis de travail ou celui des visas. D'un point de vue théorique et pratique, le passage à un processus entièrement administratif peut être envisagé, mais malheureusement pas pour le prochain recensement, qui aura lieu en 2021, pour lequel le moment du dénombrement a été fixé au 31 décembre 2020.

B. Nouvelle méthode prévue pour le prochain recensement

7. La méthode prévue pour le recensement de 2021 devrait constituer pour le Bureau central de statistique un pas vers la mise en œuvre de recensements entièrement fondés sur des données administratives. Les informations provenant d'un échantillon unique issu du fichier administratif géodémographique seront croisées avec les renseignements disponibles dans le registre de la population et d'autres fichiers administratifs, principalement pour corriger les chiffres obtenus à partir du fichier géodémographique. L'échantillon comportera des données géodémographiques sur tous les membres des ménages au moment du dénombrement, ainsi que des données socioéconomiques. Il est prévu que ces informations soient obtenues en ligne, ou à défaut par téléphone pour les personnes n'ayant pas répondu sur Internet. En cas de non-réponse par un de ces deux modes, des entretiens en face à face seront réalisés.

8. Les estimations directes obtenues à partir de l'échantillon seront améliorées grâce au modèle de Fay-Herriot et aux informations covariables pertinentes connues au niveau de la région, telles que le nombre de bâtiments et leur volume total, le volume étant défini comme la hauteur du bâtiment multiplié par la surface de son toit. D'autres covariables seront utilisées pour estimer les moyennes socioéconomiques pertinentes de la région.

9. Pour l'estimation du dénombrement de la région, le modèle de Fay-Herriot sera combiné avec les données correspondantes du fichier administratif géodémographique afin d'obtenir l'estimateur composite final du recensement (voir ci-dessous).

II. Estimateur en trois étapes proposé pour le recensement

A. Estimation par comptage direct (étape 1)

10. N désigne le nombre de résidents du pays au moment du dénombrement, et N_i le nombre de résidents de la zone i, de telle sorte que $N = \sum_i N_i$. $p_i = N_i/N$ représente la proportion de personnes inscrites dans le fichier administratif géodémographique qui résident réellement dans la région i, et $\hat{p}_i$ représente l'estimateur direct correspondant de l'échantillon, par exemple la proportion de l'échantillon en cas d'échantillonnage aléatoire simple (des méthodes d'échantillonnage et des estimateurs directs plus efficaces sont actuellement à l'étude). Enfin, $K \cong N$ désigne la taille du fichier administratif géodémographique au moment du dénombrement. L'estimateur direct pour le dénombrement de la région i est donc $\widehat{N}_i = K \times \hat{P}_i$. La variance conditionnelle basée sur le plan de sondage de $\widehat{N}_i$ est $Var_D(\widehat{N}_i|K) = K^2 Var_D(\hat{P}_i) = \sigma_{Di}^2$.

B. Estimation à l'aide du modèle « amélioré » de Fay-Herriot (étape 2)

11. La formule du modèle (standard) de Fay-Herriot (1979) est la suivante :

$$\widehat{N}_i = \alpha + x_i'\beta + u_i + e_i, \quad (3),$$

dans laquelle $\widehat{N}_i$ représente l'estimateur direct de l'échantillon, x_i les covariables de la région (nombre de bâtiments résidentiels et volume total de tous les bâtiments résidentiels

dans les exemples empiriques, des covariables plus puissantes étant actuellement recherchées), u_i un effet aléatoire et e_i l'erreur d'échantillonnage de l'estimateur direct.

12. Dans le modèle (3), le meilleur prédicteur linéaire sans biais amélioré et empirique du chiffre réel est :

$$\widehat{N}_{i,IMP} = \hat{\gamma}_i \widehat{N}_i + (1-\hat{\gamma}_i) x_i' \hat{\beta};\ \hat{\gamma}_i = \hat{\sigma}_u^2 (\hat{\sigma}_u^2 + \hat{\sigma}_{D_i}^2)^{-1}, \tag{4}$$

où $\hat{\beta}, \hat{\sigma}_u^2, \hat{\sigma}_{D_i}^2$ représentent les estimations d'échantillons appropriées.

C. Estimations finales du recensement

13. L'estimation finale du dénombrement d'une région i sera obtenue sous la forme d'une moyenne pondérée de l'estimation améliorée issue du modèle de Fay-Herriot (équation (4)) et de la valeur du chiffre issu du fichier administratif géodémographique. Pour cela, on part du principe que $K_i \sim Possion(N_i) \Rightarrow Var(K_i) = N_i$. L'estimateur composite final du recensement est donc

$$\widehat{N}_{i,COM} = \hat{\alpha}_i K_i + (1-\hat{\alpha}_i)\widehat{N}_{i,IMP};\ \hat{\alpha}_i = \frac{\hat{\sigma}_{i,FH}^2}{\hat{\sigma}_{i,FH}^2 + \hat{Var}(K_i)} \tag{5}$$

III. Autre méthode d'estimation des chiffres du recensement

A. Inclusion des données du registre dans les covariables en tant que chiffres fixes

14. Plutôt que de calculer l'estimateur composite (5), on peut inclure le chiffre issu du fichier administratif géodémographique en tant que covariable supplémentaire dans le modèle de Fay-Herriot (équation (3)). L'adaptation de ce modèle « tel quel » implique un conditionnement sur le chiffre issu du registre, sans tenir compte d'une possible erreur. Dans ce cas, l'estimation finale du chiffre du recensement correspond à l'estimation obtenue à l'aide du modèle de Fay-Herriot.

B. Prise en compte des erreurs du registre

15. Conformément aux travaux d'Ybarra et Lohr (2008), le chiffre issu du fichier administratif géodémographique est ajouté à l'ensemble des covariables, mais une éventuelle erreur de mesure est prise en compte en posant l'hypothèse que $K_i \sim N(N_i, Var(K_i))$. On considère que $\tilde{x}_i = (x_i', K_i)$. En supposant que toutes les autres covariables sont mesurées sans erreur, on obtient :

$$C_i = Var(\tilde{x}_i) = \begin{bmatrix} 0\ldots0,\ldots,0 \\ 0\ldots0,\ldots,0 \\ \ldots\ ,\ . \\ \ldots\ ,\ . \\ \ldots\ ,\ . \\ 0\ldots0,\ldots,V(K_i) \end{bmatrix}, \text{ et}$$

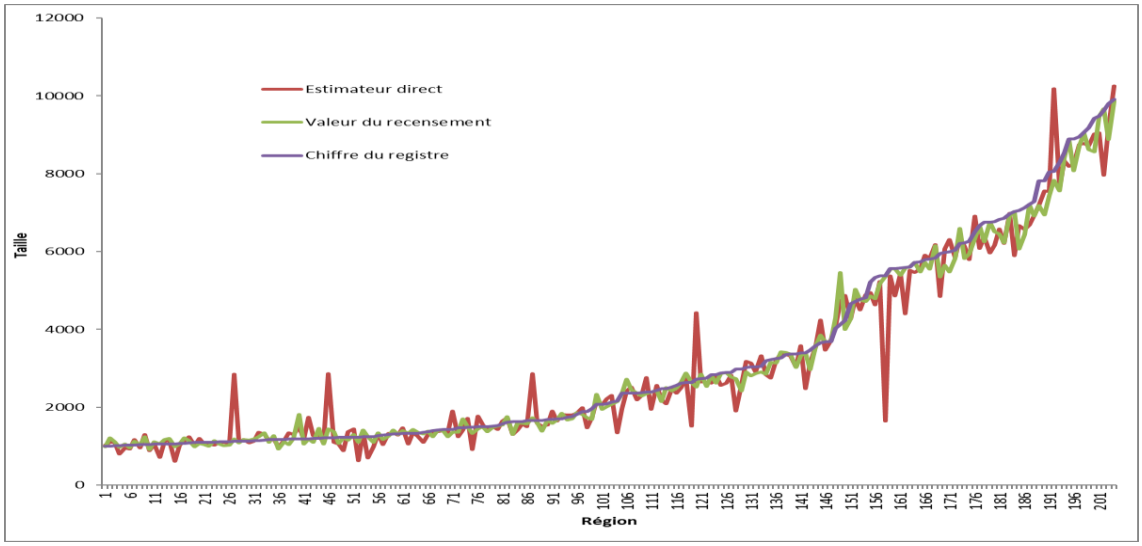
$$\hat{N}_{i,YL} = \hat{\delta}_i \hat{N}_i + (1-\hat{\delta}_i)\tilde{x}_i'\hat{\beta};\quad \hat{\delta}_i = \frac{\hat{\sigma}_u^2 + \hat{\beta}' C_i \hat{\beta}}{\hat{\sigma}_u^2 + \hat{\beta}' C_i \hat{\beta} + \hat{\sigma}_{Di}^2} \tag{6}$$

IV. Illustrations empiriques

16. Pour illustrer la méthode et ses différentes options, l'échantillon de sur-dénombrement O est tiré du registre central de la population utilisé pour le recensement de 2008. La taille totale de l'échantillon est d'environ 600 000 personnes. Les 205 régions comptant de 1 000 à 10 000 habitants estimées dans le recensement de 2008 sont prises en compte, car elles correspondent à la taille des régions d'intérêt du prochain recensement. L'échantillon a été prélevé par échantillonnage aléatoire simple stratifié. Les covariables utilisées pour les modèles sont le nombre de bâtiments résidentiels de la région et le volume total de l'ensemble de ces bâtiments. Les paramètres du modèle de Fay-Herriot ont été estimés par la méthode du maximum de vraisemblance, à l'aide du modèle mixte PROC du logiciel SAS, qui suppose la normalité des effets aléatoires et des erreurs d'échantillonnage. Les estimations du recensement de 2008 (basées sur les échantillons O et U) sont considérées comme étant les comptages réels (désignés dans les figures ci-dessous par les termes « valeur du recensement »).

Figure I

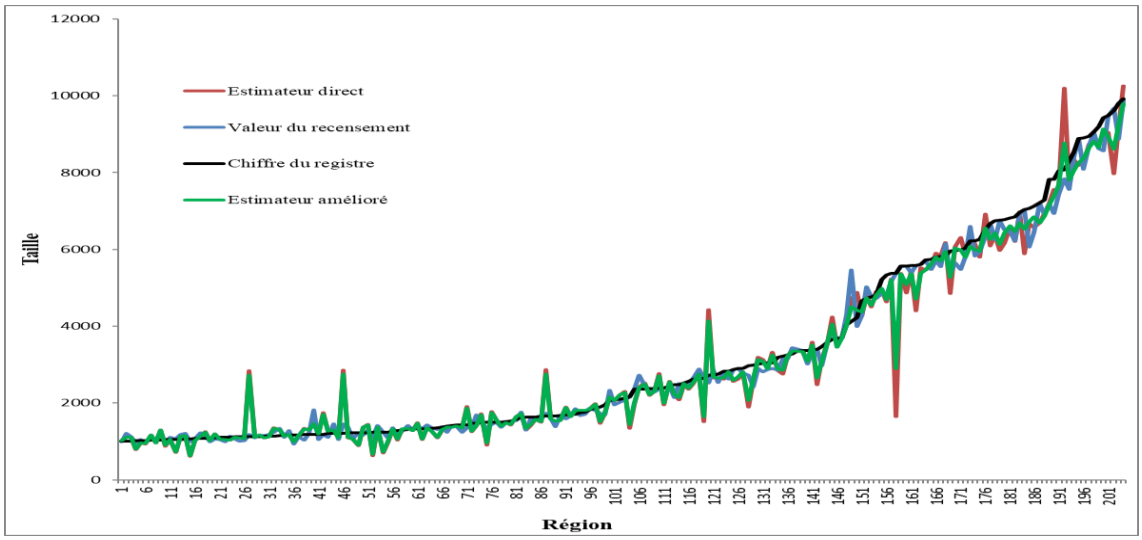
Estimateur direct, valeur de recensement et chiffre du registre pour les 205 petites régions, classées par taille



17. On constate que l'estimateur direct n'est pas biaisé, mais qu'il présente une grande variance.

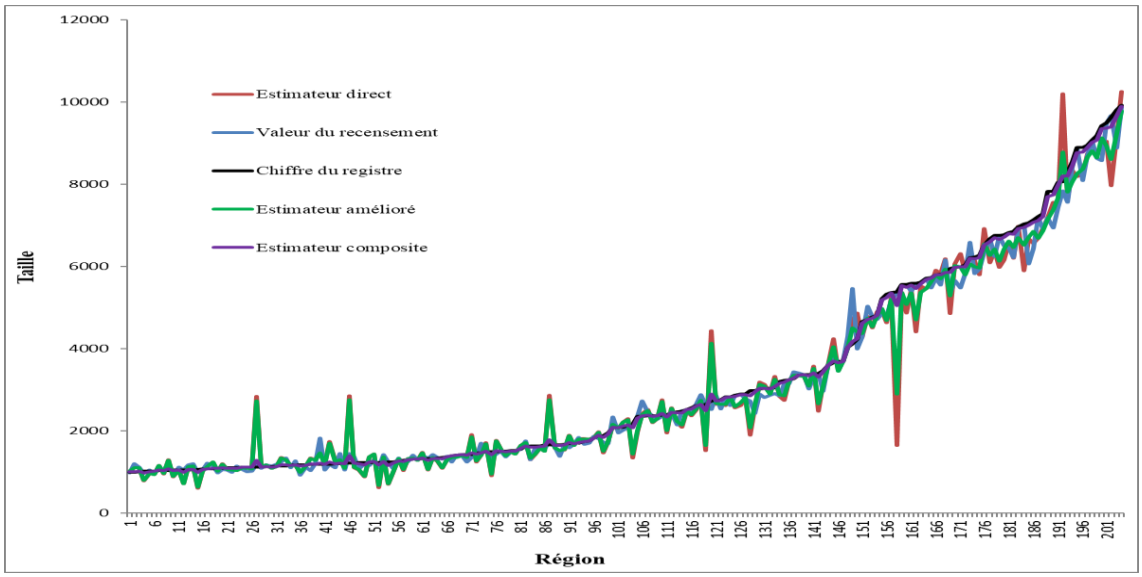
Figure II

Estimateur direct, valeur de recensement, chiffre du registre et estimateur amélioré du modèle de Fay-Herriot



18. L'estimateur amélioré du modèle de Fay-Herriot ne réduit que légèrement la variance de l'estimateur direct. Le Bureau central de statistique recherche actuellement des covariables plus puissantes. En particulier, la compagnie nationale d'électricité devrait fournir une liste de tous les appartements et maisons du pays, ce qui devrait améliorer très sensiblement l'estimateur du modèle Fay-Herriot par rapport à l'utilisation du seul nombre de bâtiments.

Figure III
Estimateur direct, valeur de recensement, chiffre du registre, estimateur amélioré et estimateur composite



19. On considère que l'estimateur composite permet d'estimer les chiffres réels avec beaucoup plus de précision que les autres estimateurs. Le tableau 1 présente quelques statistiques résumées relatives aux résultats obtenus à l'aide des différents estimateurs examinés jusqu'ici.

Tableau 1
Distance relative absolue des estimations par rapport aux valeurs du recensement I

Estimation	*Moyenne*	*10e percentile*	*25e percentile*	*50e percentile*	*75e percentile*	*90e percentile*
Directe	0,1047	0,0101	0,0243	0,0556	0,1084	0,2202
Chiffre du registre	0,0616	0,0010	0,0151	0,0507	0,0912	0,1344
Améliorée	0,0946	0,0112	0,0275	0,0573	0,0956	0,1959
Composite	0,0598	0,0056	0,0189	0,0469	0,0834	0,1257

20. Enfin, la figure IV et le tableau 2 présentent les résultats obtenus en ajoutant le chiffre du registre en tant que covariable supplémentaire dans le modèle Fay-Herriot, avec et sans prise en compte de l'erreur de mesure. Dans ce dernier cas, σ_u^2 et β sont estimés par la technique des moindres carrés modifiés (Ybarra et Lohr, 2008).

Figure IV
Estimations lors de l'addition du chiffre du registre aux covariables du modèle de Fay-Herriot, avec et sans prise en compte de l'erreur de mesure

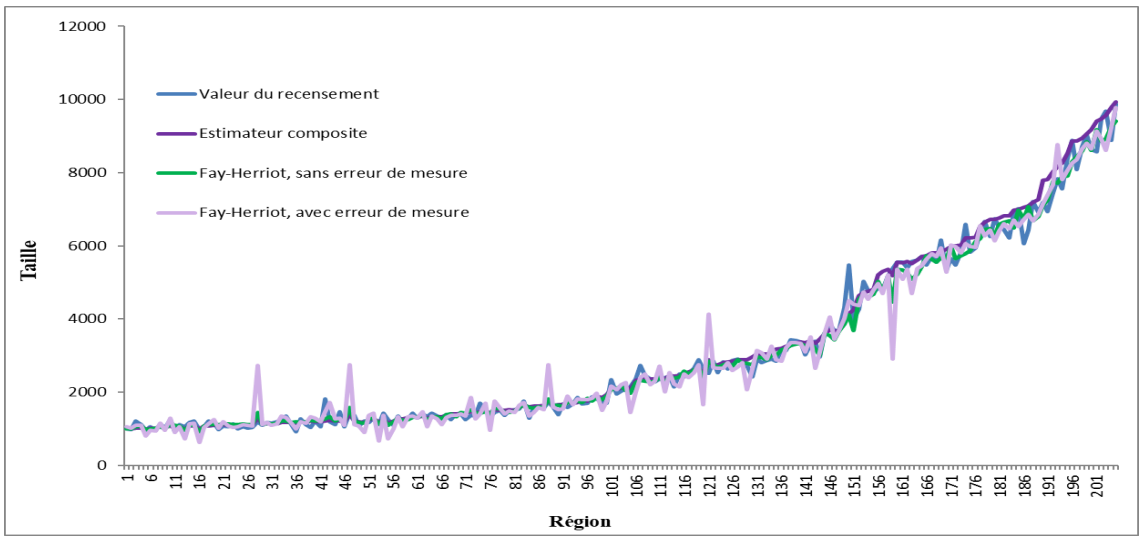


Tableau 2
Distance relative absolue des estimations par rapport aux valeurs du recensement II

Estimation	*Moyenne*	*10e percentile*	*25e percentile*	*50e percentile*	*75e percentile*	*90e percentile*
Directe	0,1047	0,0101	0,0243	0,0556	0,1084	0,2202
Chiffre du registre	0,0616	0,0010	0,0151	0,0507	0,0912	0,1344
Améliorée	0,0946	0,0112	0,0275	0,0573	0,0956	0,1959
Modèle de Fay-Herriot sans prise en compte de l'erreur de mesure	0,0893	0,0100	0,0261	0,0540	0,0931	0,1877
Composite	0,0598	0,0056	0,0189	0,0469	0,0834	0,1257
Modèle de Fay-Herriot avec prise en compte de l'erreur de mesure	0,0603	0,0094	0,0227	0,0498	0,0793	0,1230

21. Il apparaît clairement que si l'on ne tient pas compte de l'erreur de mesure du chiffre du registre, l'estimateur du recensement ne présente alors qu'une amélioration mineure par rapport à l'estimateur de l'échantillon direct. La prise en compte de l'erreur du chiffre du registre améliore très sensiblement la qualité de l'estimateur Fay-Herriot, mais, de façon assez surprenante, l'estimateur composite donne de meilleurs résultats, bien que l'estimateur d'Ybarra et Lohr (2008) soit le meilleur prédicteur linéaire empirique sans biais. Bien qu'il ne repose que sur une seule étude empirique, ce résultat peut s'expliquer par le fait que, dans ce dernier estimateur, une pondération identique est attribuée au chiffre du registre et aux autres covariables (fixes), alors que l'estimateur composite est plus souple et autorise l'attribution de pondérations différentes selon les variables. Le résultat obtenu devra être validé par de nouveaux travaux de recherche et d'autres illustrations empiriques.

V. Prise en compte de la non-réponse ne manquant pas au hasard

22. Sverchkov et Pfeffermann (2018) proposent une méthode qui utilise le principe de l'information manquante d'Orchard et Woodbury (1972) pour estimer les probabilités de réponse dans de petites régions. L'idée de base est la suivante : il faut d'abord construire la probabilité qui serait obtenue si les valeurs manquantes des résultats étaient également connues pour les non-répondants. Toutefois, comme les résultats manquants sont pratiquement inconnus, il faut remplacer la probabilité par les attentes en ce qui concerne la distribution des résultats manquants, en tenant compte de toutes les données observées. Cette distribution s'obtient à partir des résultats observés, tels qu'ajustés aux valeurs

observées. On pourra se référer à Sverchkov et Pfeffermann (2018) pour la relation entre les distributions des résultats observés et des résultats manquants, pour des covariables et des probabilités de réponse données.

23. Dans l'idéal, il faudrait montrer de quelle façon la méthode permet d'estimer le nombre réel de personnes résidant dans chaque région au moment du dénombrement, mais cette information est pratiquement inconnue pour les données d'essai (l'échantillon O utilisé jusqu'ici). Par conséquent, ce qui suit illustre plutôt les résultats obtenus à l'aide d'une méthode visant à prédire le nombre réel de personnes divorcées enregistrées dans chaque région. L'échantillon O est tiré du registre central de la population et le nombre réel de personnes divorcées enregistrées dans chaque région est donc connu.

24. Il faut assigner à la variable de résultat y_{ij} la valeur 1 si la personne j enregistrée dans la région i est divorcée, et 0 si elle ne l'est pas. L'indicateur de réponse R_{ij} doit prendre la valeur 1 si l'unité j de la région i répond, et 0 si elle ne répond pas. Cette analyse est limitée aux personnes âgées de 20 ans et plus. Le modèle ajusté pour les résultats observés des unités répondantes et le modèle supposé pour les probabilités de réponse sont définis dans les équations (7) et (8). Les covariables utilisées pour cette illustration sont énumérées dans le tableau 3.

$$Pr(y_{ij} = 1 | x_{ij}, u_i, R_{ij} = 1) = \frac{exp(\beta_0 + x_{ij}^{'}\beta + u_i)}{1 + exp(\beta_0 + x_{ij}^{'}\beta + u_i)}; \quad u_i \sim N(0, \sigma_u^2) \tag{7}$$

$$Pr(R_{ij} = 1 | y_{ij}, x_{ij}, u_i; \gamma) = \frac{exp(\gamma_0 + x_{ij}^{'}\gamma + \gamma_y y_{ij})}{1 + exp(\gamma_0 + x_{ij}^{'}\gamma + \gamma_y y_{ij})} \tag{8}$$

25. Pour $\gamma_y \neq 0$, il est clair que l'équation (8) définit un mécanisme de réponse informatif. La première valeur $\gamma_y = 0$ est imposée, ce qui présume que le fait d'être divorcé n'a pas d'incidence sur la probabilité de réponse et correspond à l'hypothèse de non-réponse manquant au hasard. Pour ce faire, on omet le statut matrimonial y_{ij} dans le modèle de réponse (8). Les résultats sont présentés dans les tableaux 3 et 4.

26. Le tableau 3 présente les rapports des cotes du modèle logistique estimé des probabilités de réponse pour ce cas. Comme prévu, le rapport des cotes pour la réponse augmente à mesure que le nombre de téléphones appartenant à la famille administrative augmente. Il en va de même pour la taille de la famille administrative. Le groupe d'âge ayant la plus faible probabilité de réponse est celui des 30-39 ans (rapport des cotes de 0,87), et la probabilité que les personnes nées en Israël répondent est beaucoup plus élevée que pour les personnes nées à l'étranger.

Tableau 3
Rapports des cotes du modèle logistique estimé des probabilités de réponse en supposant une non-réponse manquant au hasard

Variable	*Rapport des cotes en cas de non-réponse manquant au hasard*
Nombre de téléphones par famille	1,70
Taille de la famille administrative	1,15
Tranche d'âge 20-29 ans	0,98
Tranche d'âge 30-39 ans	0,87
Tranche d'âge 40 ans et plus	1,00
Juif/juive	1,04
Autre	1,00
Né(e) en Israël	1,27
Né(e) ailleurs	1,00

27. Le tableau 4 présente la distribution des probabilités de réponse estimées selon le modèle du tableau 3.

Tableau 4
Distribution des probabilités de réponse estimées selon le modèle présenté au tableau 3

Situation matrimoniale	*Moyenne*	*5e percentile*	*25e percentile*	*75e percentile*
Autre	0,815	0,489	0,822	0,885
Divorcé(e)	0,742	0,359	0,683	0,843
Total	0,812	0,487	0,819	0,885

28. Il ressort clairement du tableau 4 que l'hypothèse $\gamma_y = 0$ est erronée. La probabilité que les personnes divorcées répondent est significativement plus faible que pour les autres. Par conséquent, les probabilités de réponse ont été estimées en incluant la variable binaire « divorcé(e) » en tant que variable explicative supplémentaire.

Tableau 5
Rapports des cotes du modèle logistique estimé des probabilités de réponse permettant une non-réponse ne manquant pas au hasard

Variable	*Rapport des cotes en cas de non-réponse manquant au hasard*	*Rapport des cotes en cas de non-réponse ne manquant pas au hasard*
Nombre de téléphones par famille	1,70	1,83
Taille de la famille administrative	1,15	1,11
Tranche d'âge 20-29 ans	0,98	0,95
Tranche d'âge 30-39 ans	0,87	0,86
Autres tranches d'âge	1,00	1,00
Juif/juive	1,04	1,05
Autre	1,00	1,00
Né(e) en Israël	1,27	1,25
Né(e) ailleurs	1,00	1,00
Divorcé(e)	---	0,531

29. Le tableau 5 montre que le ratio de probabilité de réponse des personnes divorcées est environ deux fois plus faible que celui des autres personnes, ce qui corrobore les résultats du tableau 6. Il est intéressant de noter que les rapports de cotes des autres covariables sont très semblables aux rapports de cotes obtenus en supposant une non-réponse manquant au hasard.

30. Les modèles estimés des tableaux 3 et 5 permettent d'estimer la probabilité de réponse pour chaque répondant de l'échantillon. Si l'on considère la réponse comme une étape supplémentaire de l'échantillonnage, les probabilités de réponse estimées peuvent être utilisées pour prédire les moyennes régionales réelles de la variable cible (proportion de personnes divorcées dans le présent exemple) selon la théorie de l'échantillonnage standard, par exemple en utilisant l'estimateur approximativement sans biais sous le plan,

$$\hat{\bar{Y}}_i^{HB} = \sum_{j,(i,j)\in R} (y_{ij}/\tilde{\pi}_{j|i}) \Big/ \sum_{j,(i,j)\in R} (1/\tilde{\pi}_{j|i}); \tilde{\pi}_{j|i} = \pi_{j|i}\hat{p}_r(y_{ij}, x_{ij}; \hat{\gamma}) \qquad (9)$$

où $\pi_{j|i}$ désigne la probabilité d'échantillonnage. Sverchkov et Pfeffermann (2018) obtiennent également le meilleur prédicteur empirique avec les modèles (7) et (8), mais ce prédicteur n'est pas étudié dans le présent document.

31. La figure V et les tableaux 6 et 7 comparent les résultats obtenus à l'aide des trois prédicteurs suivants de la proportion réelle de personnes divorcées dans les diverses régions : la proportion de personnes divorcées dans l'échantillon observé, sans tenir compte de la non-réponse (l'estimateur direct) ; l'estimateur obtenu dans l'hypothèse de la non-réponse manquant au hasard et l'estimateur obtenu en tenant compte de la non-réponse ne manquant pas au hasard (équation 8).

Figure V
Pourcentage de personnes divorcées dans les régions : valeur réelle, estimateur direct et estimateurs obtenus dans l'hypothèse d'une non-réponse manquant au hasard ou ne manquant pas au hasard

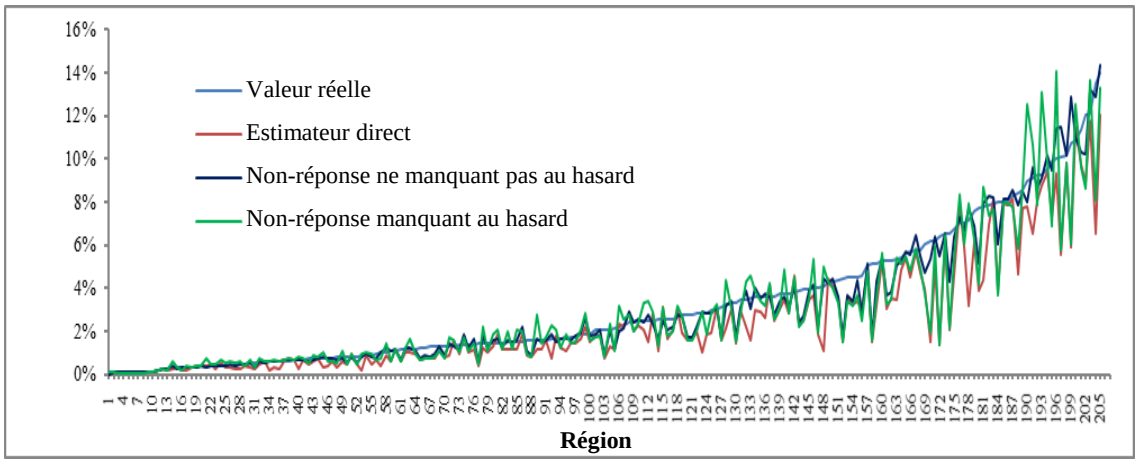


Tableau 6
Différence entre les valeurs réelles et les estimations sur l'ensemble des régions

Estimateur	*Moyenne*	*10e percentile*	*25e percentile*	*50e percentile*	*75e percentile*	*90e percentile*
Direct	0,0075	-0,0005	0,0006	0,0036	0,0099	0,0211
Manquant au hasard	0,0033	-0,0077	-0,0018	0,0004	0,0057	0,0168
Ne manquant pas au hasard	0,0019	-0,0027	-0,0004	0,0001	0,0032	0,0094

Tableau 7
Distance relative absolue des estimations par rapport aux valeurs réelles

Estimateur	*Moyenne*	*10e percentile*	*25e percentile*	*50e percentile*	*75e percentile*	*90e percentile*
Direct	0,270	0,042	0,121	0,233	0,406	0,551
Manquant au hasard	0,256	0,032	0,113	0,216	0,379	0,472
Ne manquant pas au hasard	0,118	0,004	0,022	0,055	0,156	0,362

32. Comme le montrent la figure V et les tableaux 6 et 7, les estimations obtenues en tenant compte de la non-réponse ne manquant pas au hasard présentent de loin le biais le plus faible et la plus faible distance relative absolue par rapport aux valeurs réelles. Les estimations directes, qui ne tiennent pas compte de la non-réponse, présentent un biais important et une grande distance relative par rapport aux valeurs réelles.

VI. Estimation de l'erreur quadratique moyenne en cas de non-réponse ne manquant pas au hasard

33. Comme dans toute publication de statistique officielle, il est nécessaire de publier des chiffres et d'autres estimations sociodémographiques, mais aussi d'évaluer leur précision. Une série de mesures d'évaluation (qui n'est certainement pas la seule possible) consiste à estimer pour chaque région l'erreur quadratique moyenne de l'estimation finale.

Cette opération serait relativement simple si les estimations directes de l'échantillon devaient être utilisées comme estimations finales et si toutes les unités d'échantillonnage répondaient, mais ce cas de figure ne peut évidemment pas se produire. L'estimation de l'erreur quadratique moyenne est plus compliquée lorsqu'on utilise les estimations du modèle de Fay-Herriot et que l'on tient compte de la non-réponse ne manquant pas au hasard, et encore plus lorsqu'on utilise l'estimateur composite décrit à la section II.

34. Sverchkov et Pfeffermann (2018) proposent une méthode bootstrap pour l'estimation de l'erreur quadratique moyenne des non-réponses ne manquant pas au hasard dans les estimations relatives à de petites régions, qui tient compte des processus aléatoires supposés générer les valeurs de population et des processus d'échantillonnage et de réponse. Cela signifie que les paramètres de la région cible (la proportion réelle de personnes résidant dans la région au moment du dénombrement par rapport à l'ensemble des personnes inscrites dans le registre central de la population) sont considérés comme aléatoires, ce qui diffère des méthodes classiques d'échantillonnage d'enquête, dans lesquelles les valeurs de la population finie et donc les paramètres cibles sont considérées comme fixes. Les utilisateurs des estimations de l'enquête par sondage (statistique officielle) connaissent bien les mesures de l'erreur, qui ne tiennent compte que de la variabilité découlant du caractère aléatoire de la sélection de l'échantillon (appelée distribution randomisée) et de la non-réponse. En d'autres termes, ils sont habitués aux estimations de l'erreur quadratique moyenne basée sur le plan de sondage (randomisation), pour toutes les sélections d'échantillons possibles, les valeurs de population des variables de l'enquête (et donc les valeurs des paramètres cibles) restant fixes. L'estimation et la publication de l'erreur quadratique moyenne basée sur le plan de sondage (ou de sa racine carrée) sont une pratique courante dans les organismes nationaux de statistique du monde entier.

35. Dans un article récent, Pfeffermann et Ben-Hur (2019) ont proposé une nouvelle méthode d'estimation de l'erreur quadratique moyenne basée sur le plan de sondage des prédicteurs pour les petites régions fondés sur un modèle, qui s'est révélée efficace dans une étude de simulation à grande échelle et qui surpasse les autres méthodes proposées dans la littérature. Cette procédure est en train d'être étendue à l'estimation de l'erreur quadratique moyenne basée sur le plan de sondage des estimateurs composites proposés, compte tenu, en particulier, de l'inévitable non-réponse ne manquant pas au hasard et de l'utilisation de l'estimateur composite, qui combine l'estimation de l'enquête et le chiffre issu du registre administratif de la population.

VII. Conclusion

36. Les auteurs de la présente note proposent une nouvelle méthode de recensement qui combine des estimations par sondage et l'utilisation de données administratives. L'un des principaux avantages de cette méthode est qu'elle ne nécessite pas de réaliser des entretiens en face à face, sauf dans le cas des non-répondants. Israël ne dispose toujours pas d'un registre des logements suffisamment fiable, et l'utilisation d'un échantillon prélevé sur le terrain nécessite l'établissement préalable d'une liste de toutes les unités d'habitation d'un échantillon de cellules dans chaque région statistique, ce qui est assez complexe sur le plan logistique et très coûteux. Il importe également de vérifier que chaque appartement est habité.

37. Selon la nouvelle méthode, un seul échantillon de personnes est tiré du fichier administratif géodémographique, dont on sait qu'il est généralement exact au niveau national, à l'exception de quelques petites sous-populations « périphériques » telles que celle des immigrés clandestins. D'autres moyens sont envisagés pour combiner les données de l'enquête avec celles du fichier administratif afin de constituer un unique estimateur final du recensement, en tenant compte des erreurs d'échantillonnage de l'enquête et des erreurs d'adresses dans le fichier. Une procédure descriptive simple de vérification de la valeur informative des données d'échantillonnage manquantes est également proposée, ainsi qu'un moyen de prendre en compte la non-réponse ne manquant pas au hasard. Tous ces sujets sont illustrés par l'utilisation de données empiriques réelles.

38. Une répétition du recensement doit avoir lieu l'année prochaine dans deux régions statistiques d'Israël, ce qui devrait donner au Bureau central de statistique une nouvelle occasion de tester, à l'aide de données actualisées, les propositions examinées dans la présente note.

39. En s'appuyant davantage sur les sources administratives d'information, on ouvre de nouvelles perspectives pour le processus et les résultats du recensement. Se référer à une population identifiée au sein d'une base de population connue implique un changement substantiel de paradigme. Les limites d'une région deviennent, d'une certaine manière, une entité virtuelle plutôt que la principale entité physique d'un recensement. Il conviendra d'étudier plus avant les implications théoriques et socioéconomiques de ce changement, ainsi que ses effets sur l'élaboration des politiques.

Références

Blum, O. (2018). « From physical area to virtual lists : Toward an administrative census in Israel ». Groupe d'experts des recensements de la population et des habitations de la CEE, Genève.

Fay, R. E. et Herriot, R. A. (1979). « Estimation of income from small places : An application of James Stein procedures to census data ». *Journal of the American Statistical Association*, vol. 74, p. 269-277.

Orchard, T., et Woodbury, M. A. (1972). « A missing information principle : theory and application ». *Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, p. 697-715.

Pfeffermann, D. et Ben-Hur, D. (2018). « Estimation of Randomization Mean Square Error in Small Area Estimation ». *International Statistical Review*, vol. 87, numéro spécial 1, p. 31-49.

Sverchkov, M., et Pfeffermann, D. (2018), « Small area estimation under informative sampling and not missing at random non-response », *Journal of the Royal Statistical Society*, série A, vol. 181, p. 981-1008.

Ybarra, L. M. R., et Lohr, S. L. (2008), « Small area estimation when auxiliary is measured with error », *Biometrika*, vol. 95, p. 919-931.