Recueil du Symposium 2018 de Statistique Canada
Combiner pour vaincre : innovations dans l'utilisation de multiples sources de données

L'estimation sur petits domaines pour corriger les erreurs de mesure dans les registres de population de grande taille

Danny Pfeffermann et Dano Ben-Hur[1]

Résumé

Comme de nombreux pays, Israël a un registre de la population assez précis à l'échelon national, composé d'environ 9 millions de personnes. Toutefois, le registre est beaucoup moins précis pour les petites régions géographiques (secteurs statistiques), l'erreur moyenne de dénombrement des secteurs étant d'environ 13 %. La principale raison de cette inexactitude au niveau du secteur vient du fait que les personnes qui emménagent dans une région ou qui déménagent ailleurs tardent souvent à signaler leur changement d'adresse. Pour corriger les erreurs au niveau du secteur dans notre prochain recensement, nous étudions la procédure en trois étapes que voici :

A- constituer un échantillon à partir du registre afin d'obtenir des estimations préliminaires directes de l'échantillon sur le nombre de personnes qui habitent dans chaque région le « jour du recensement »,
B- appliquer le modèle de Fay-Herriot aux estimations directes dans le but d'en améliorer l'exactitude,
C- calculer une estimation finale pour chaque secteur statistique comme la combinaison linéaire de l'estimation obtenue à l'étape B et des chiffres du registre.

Nous envisageons également une procédure qui permettrait de traiter les cas de non-réponse pour la non-répartition au hasard des données manquantes (NMAR) à l'étape A. Nous illustrons les procédures proposées à l'aide des données du recensement de 2008 en Israël.

Mots-clés : Estimateur direct; modèle de Fay-Herriot; principe de l'information manquante; non-réponse NMAR

1. Introduction

Dans le présent article, nous proposons une nouvelle façon d'effectuer un recensement, qui combine une enquête et des mégadonnées administratives. Nous examinons d'autres manières d'intégrer les données d'enquête et les données administratives afin d'obtenir une seule estimation du recensement dans de petites régions géographiques, en tenant compte des erreurs dans les sources des données et de la non-réponse pour la non-répartition au hasard des données manquantes (NMAR). Nous illustrons la méthode que nous proposons à l'aide des données du recensement de 2008 en Israël.

1.1 Description du dernier recensement en Israël (2008)

Israël a un registre de la population assez précis, presque parfait au niveau du pays. Toutefois, le registre de la population est bien moins exact pour les petits secteurs statistiques, l'erreur de dénombrement moyenne s'établissant à 13 % et le 95^{e} centile, à 40 %. Israël compte environ 3 000 secteurs statistiques et des données du recensement, comme les chiffres et des données socioéconomiques, sont nécessaires pour tous les secteurs. La principale raison de l'inexactitude des chiffres du registre au niveau du secteur vient du fait que les personnes qui emménagent dans une région ou qui déménagent dans une autre tardent souvent à signaler leur changement d'adresse. En 2008, le Bureau central de la statistique d'Israël (ICBS) a tenu un recensement intégré, composé du registre de la population, corrigé à l'aide des estimations provenant de deux échantillons de la couverture pour chacun des secteurs : un échantillon sur le terrain (secteur) des logements pour estimer le sous-dénombrement du registre (« échantillon U ») et un échantillon

[1]Danny Pfeffermann, Israel Central Bureau of Statistic, Hebrew University of Jerusalem et University of Southampton; Dano Ben-Hur, Central Bureau of Statistic, 66 Kanfey Nesharim street, Jerusalem, Israel, 9546456

téléphonique des personnes inscrites dans le secteur pour estimer le surdénombrement du registre (« échantillon O »). L'échantillon U a également permis de recueillir des données socioéconomiques.

L'estimation finale du recensement a été calculée comme ceci : nous désignons par N_i le nombre réel de personnes qui vivent dans le secteur i le jour du recensement et par K_i le nombre de personnes inscrites comme vivant dans le secteur. Disons que $p_{i,L|R}$ représente la proportion de personnes vivant dans le secteur i parmi toutes celles inscrites comme vivant dans le secteur et $p_{i,\mathrm{R|L}}$ représente la proportion de personnes inscrites dans le secteur i parmi toutes celles qui vivent dans le secteur. Alors,

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

À développement en série de Taylor, la variance conditionnelle (fondée sur le plan) de $\hat{N}_i$ pourrait s'écrire comme ceci :

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

Il semblerait très intéressant d'utiliser les échantillons U et O avec les estimations résultantes, mais la réalisation réelle de l'échantillonnage du secteur U (sur le terrain) a été loin d'être simple. Parmi les principales difficultés, mentionnons les suivantes :

1. La méthode nécessite le listage de tous les appartements dans chaque secteur statistique, ou au moins un échantillon de cellules ou d'immeubles dans chaque secteur. Cela coûte très cher et nécessite de vérifier également que les appartements inscrits sont des unités d'habitation.

2. Des problèmes de couverture à tout endroit où il y a des problèmes d'accès, comme des étages fermés, des clôtures fermées, etc.

3. La difficulté à repérer les appartements de l'échantillon au moment de recueillir des données parce que tous les appartements ne sont pas repérés à l'étape du listage.

4. La réponse sur Internet est la méthode de choix, mais, en raison des problèmes susmentionnés, il est difficile de savoir quels ménages ont effectivement répondu.

5. Beaucoup de problèmes de logistique lors de l'exécution d'une enquête à une échelle aussi grande.

1.2 Nouvelle méthode prévue pour le prochain recensement en Israël

En raison des difficultés susmentionnées, qui sont liées au recensement de 2008, nous prévoyons adopter une méthode différente pour notre recensement de 2021 (le 31 décembre 2020 étant la date de référence du jour du recensement). Le recensement combinera de l'information d'un seul échantillon constitué à partir du registre de la population, à l'aide des données disponibles provenant du registre et d'autres fichiers administratifs. L'échantillon permettra de recueillir des renseignements sur le lieu de résidence de tous les membres d'un ménage administratif le jour du recensement, ainsi que des données socioéconomiques. Les renseignements devraient être obtenus d'abord sur Internet, puis au téléphone pour les personnes qui ne répondent pas par Internet et, dans les cas de non-réponse, par l'un ou l'autre de ces moyens, grâce à des interviews sur place.

Les estimations directes découlant de l'échantillon seront améliorées à l'aide de l'estimateur de Fay-Herriot (F-H), au moyen de l'information sur les covariables pertinentes connues au niveau du secteur, comme le nombre d'immeubles et le volume total de tous les immeubles dans le secteur, le volume étant défini comme la superficie du toit d'un immeuble multipliée par sa hauteur. D'autres covariables serviront à estimer la moyenne socioéconomique du secteur d'intérêt.

Pour estimer les chiffres du secteur, nous devons combiner l'estimateur de F-H et les chiffres correspondants du registre, afin d'obtenir notre estimateur du recensement composite final (voir ci-dessous).

2. Estimateur du recensement en trois étapes proposées

2.1 Estimation directe des chiffres (étape 1)

Nous désignons par N le nombre de résidents dans le pays le jour du recensement et par N_i le nombre de résidents dans le secteur i, de telle sorte que $N = \sum_i N_i$. Disons que $p_i = N_i / N$ – est la proportion réelle des résidents dans le registre qui vivent dans le secteur i et désignons par $\hat{p}_i$ l'estimateur de l'échantillon direct correspondant, p. ex. la proportion de l'échantillon dans le cas de l'échantillonnage aléatoire simple (des plans d'échantillonnage et des estimateurs directs plus efficaces sont actuellement à l'étude). Disons que $K \cong N$ représente la taille du registre le jour du recensement. L'estimateur direct des chiffres du secteur *i* est alors $\hat{N}_i = K \times \hat{P}_i$. La variance est la suivante : $Var_D(\hat{N}_i \mid K) = K^2 Var_D(\hat{P}_i) = \sigma^2_{Di}$.

2.2 Estimations de Fay-Herriot « améliorées » (étape 2)

Le modèle (standard) de Fay Herriot (F-H) (1979) est le suivant :

$$\hat{N}_i = \alpha + \mathrm{x}_i'\beta + u_i + e_i\,, \tag{3}$$

où $\hat{N}_i$ est l'estimateur de l'échantillon direct, x_i représente les covariables au niveau du secteur (le nombre d'immeubles résidentiels dans le secteur et le volume total de tous les immeubles résidentiels dans nos illustrations empiriques; nous sommes actuellement à la recherche de covariables plus solides), u_i est un effet aléatoire et e_i est l'erreur d'échantillonnage de l'estimateur direct.

Selon le modèle (3), la meilleure estimation linéaire sans biais empirique (EBLUP) améliorée des chiffres réels est :

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

où $\hat{\beta}, \hat{\sigma}^2_u, \hat{\sigma}^2_{D_i}$ sont les estimations de l'échantillon approprié.

2.3 Estimation des chiffres finals du recensement

L'estimation des chiffres finals dans le secteur i s'obtient en calculant la moyenne pondérée de l'estimation de F-H améliorée sous (4) et les chiffres du registre de la population. Pour ce faire, nous supposons que $K_i \sim Possion(N_i) \Rightarrow Var(K_i) = N_i$. L'estimateur composite final du recensement est alors le suivant :

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

3. Autre méthode d'estimation des chiffres du recensement

3.1 Extension du modèle

Au lieu de calculer l'estimateur composite (5), inclure les chiffres du registre comme des covariables supplémentaires dans le modèle de F-H (3). Pour ajuster ce modèle « tel quel », il faut qu'il soit fondé sur les chiffres connus du registre et ne pas tenir compte de l'erreur possible.

3.2 Extension du modèle, en tenant compte des erreurs du registre

Dans la foulée d'Ybarra et Lohr (2008), nous tenons compte des erreurs de mesure des chiffres du registre en supposant que :

$K_i \sim N(N_i, Var(K_i))$. Nous disons que $\tilde{x}_i = (x_i', K_i)$. En supposant que toutes les autres covariables sont mesurées sans erreur,

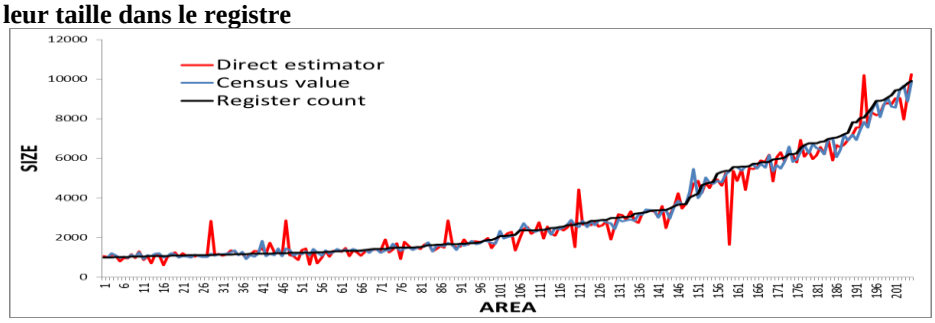
$$C_i = Var(\tilde{x}_i) = \begin{bmatrix} 0 \ldots 0, & \ldots, & 0 \\ 0 \ldots 0, & \ldots, & 0 \\ \ldots & , & . \\ \ldots & , & . \\ \ldots & , & . \\ 0 \ldots 0, & \ldots, & V(K_i) \end{bmatrix}, \text{ et}$$

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

4. Illustrations empiriques

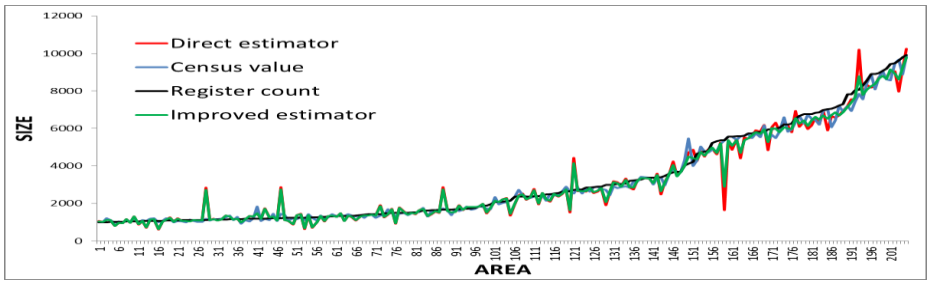
Pour illustrer cette méthode, nous nous servons de l'échantillon du surdénombrement (O) prélevé lors du recensement de 2008. La taille totale de l'échantillon est d'environ 600 000 personnes. Nous considérons les 205 secteurs, dont la taille va de 1 000 à 10 000 personnes, selon les estimations du recensement de 2008, parce que la taille de ces secteurs correspond à la taille des secteurs statistiques d'intérêt. L'échantillon a été recueilli à l'aide d'un échantillonnage aléatoire simple stratifié. Les covariables employées dans les modèles sont le nombre d'immeubles résidentiels dans le secteur et le volume total de tous les immeubles résidentiels. Les paramètres du modèle de F-H ont été estimés à l'aide de l'EMV, en utilisant la procédure PROC Mixed dans SAS, en supposant la normalité des effets aléatoires et des erreurs d'échantillonnage. Les estimations du recensement de 2008 (fondées sur les échantillons O et U) sont considérées être les chiffres réels (indiqués dans les figures comme les « valeurs du recensement »).

Figure 4-1
Estimateur direct, valeurs du recensement et chiffres du registre pour les 205 secteurs, classés en fonction de leur taille dans le registre



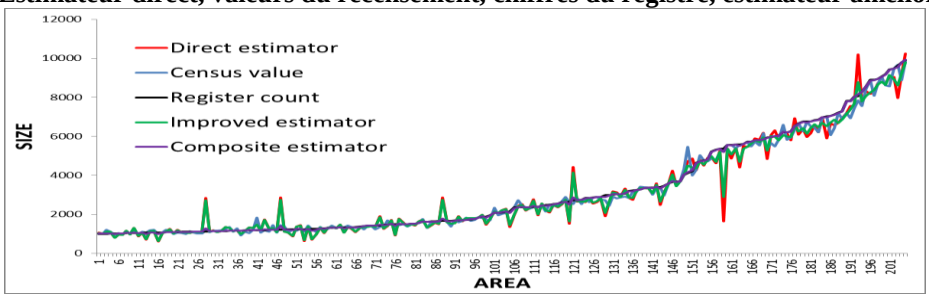
Comme nous pouvons le voir, l'estimateur direct ne présente pas de biais, mais il y a une variance importante.

Figure 4-2
Estimateur direct, valeurs du recensement, chiffres du registre et estimateur (de F-H) amélioré



L'estimateur de F-H amélioré ne réduit que légèrement la variance de l'estimateur direct. Nous sommes actuellement à la recherche de covariables plus solides.

Figure 4-3
Estimateur direct, valeurs du recensement, chiffres du registre, estimateur amélioré et estimateur composite



L'estimateur composite sert à estimer les chiffres réels bien plus précisément que les autres estimateurs. Le tableau 4.1 montre certaines statistiques sommaires sur le rendement des divers estimateurs que nous avons étudiés jusqu'à maintenant.

Tableau 4-1
Distance relative absolue des estimations par rapport aux valeurs du recensement

Estimation	Moyenne	10e centile	25e centile	50e centile	75e centile	90e centile
Directe	0,1047	0,0101	0,0243	0,0556	0,1084	0,2202
Chiffres 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

Enfin, la figure 4-4 et le tableau 4-2 montrent les résultats obtenus après l'ajout des chiffres du registre comme covariables supplémentaires dans le modèle de F-H, en tenant compte (FH_WME) et en ne tenant pas compte (FH_NME) de l'erreur de mesure. Dans le dernier cas, nous avons estimé σ_u^2 et β en nous servant de la méthode modifiée des moindres carrés (Ybarra et Lohr, 2008).

Figure 4-4
Estimations avec l'ajout des chiffres du registre aux covariables du modèle de Fay-Herriot, en tenant compte et en ne tenant pas compte de l'erreur de mesure

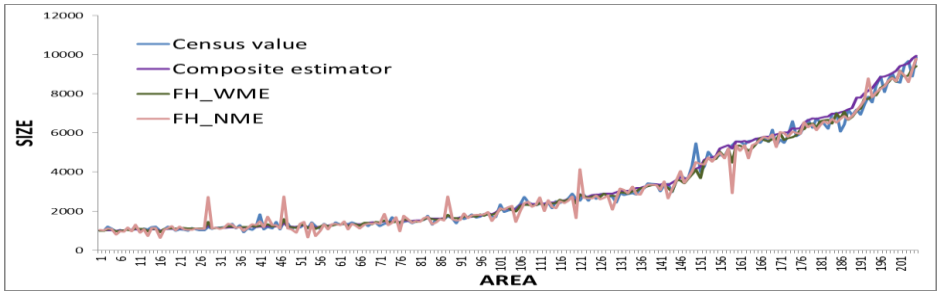


Tableau 4-2
Distance relative absolue des estimations par rapport aux valeurs du recensement

Estimation	Moyenne	10e centile	25e centile	50e centile	75e centile	90e centile
Directe	0,1047	0,0101	0,0243	0,0556	0,1084	0,2202
Chiffres 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
FH_NME	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
FH_WME	0,0603	0,0094	0,0227	0,0498	0,0793	0,1230

Comme nous le voyons clairement, le fait de ne pas tenir compte de l'erreur de mesure des chiffres du registre donne un estimateur du recensement qui ne représente qu'une amélioration mineure par rapport à l'estimateur direct des variables. Le fait de tenir compte de l'erreur dans les chiffres du registre améliore grandement le rendement de l'estimateur de F-H, mais il est étonnant de constater que l'estimateur composite s'en tire un peu mieux, malgré la propriété de l'EBLUP de l'estimateur d'Ybarra et Lohr (2008). Bien que les résultats ne reposent que sur une étude empirique, ils s'expliquent peut-être par le fait que, dans le dernier estimateur, le même facteur de pondération est attribué aux chiffres du registre et aux autres covariables (fixes), alors que l'estimateur composite est plus souple et qu'il permet d'utiliser des poids différents pour les chiffres du registre et les autres covariables. D'autres recherches théoriques et illustrations empiriques sont nécessaires pour valider ce résultat.

5. Prise en compte de la non-réponse pour la non-répartition au hasard des données manquantes (NMAR)

Sverchkov et Pfeffermann (2018) proposent une méthode qui fait appel au principe de l'information manquante d'Orchard et Woodbury (1972) pour estimer les probabilités de réponse dans de petits secteurs. L'idée de base est la suivante : il faut d'abord établir la probabilité qui serait obtenue si les valeurs résultantes manquantes étaient aussi connues pour les non-répondants. Or, comme les résultats manquants sont pratiquement inconnus, il faut remplacer la probabilité par son espérance pour la répartition des résultats manquants, en fonction de toutes les données observées. La dernière répartition est obtenue à partir de la répartition des résultats observés, qui ont été ajustés aux valeurs observées. Voir dans Sverchkov et Pfeffermann (2018) la relation entre la répartition des résultats observés et manquants, pour des covariables et des probabilités de réponse données.

Dans l'idéal, nous voudrions démontrer comment la méthode permet d'estimer le nombre réel de personnes qui habitent dans chaque secteur le jour du recensement, mais ces renseignements sont pratiquement inconnus pour les données de notre essai (l'échantillon O a été utilisé jusqu'à maintenant). Par conséquent, dans les paragraphes suivants, nous illustrons plutôt le rendement de la méthode pour prévoir le nombre réel de personnes divorcées et inscrites dans chaque secteur. L'échantillon O provient du registre de la population et le nombre réel de personnes divorcées inscrites dans chaque secteur est connu.

Nous définissons la variable de résultat, y_{ij}, comme étant 1 si la personne j inscrite dans le secteur i est divorcée, et, autrement, comme 0, et l'indicateur de réponse, R_{ij}, est 1, si la personne j dans le secteur i répond et, s'il ne répond pas, la valeur est 0. Nous restreignons l'analyse aux personnes âgées de 20 ans et plus. Les modèles ajustés pour les résultats observés des unités qui ont répondu et les probabilités de réponse sont définis dans les équations (7) et (8). Les covariables utilisées aux fins de cet exemple figurent au tableau 5.1.

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

$$\Pr(R_{ij}=1\,|\,y_{ij},\mathrm{x}_{ij},\mathrm{u}_i;\gamma)=\frac{\exp(\gamma_0+\mathrm{x}'_{ij}\gamma+\gamma_y y_{ij})}{1+\exp(\gamma_0+\mathrm{x}'_{ij}\gamma+\gamma_y y_{ij})}. \tag{8}$$

De toute évidence, pour $\gamma_y\neq 0$, l'équation (8) définit un mécanisme de réponse révélateur.

Disons d'abord que $\gamma_y=0$, en supposant que le fait d'être divorcé n'a pas d'incidence sur la probabilité de réponse, ce qui équivaut à présumer une non-réponse pour les données manquantes au hasard (MAR), ce que nous mettons en place en omettant l'état civil, y_{ij}, du modèle de réponse (8).

Tableau 5-1
Rapports de côtes du modèle logistique estimatif des probabilités de réponse en supposant une non-réponse MAR

Variables	Rapport de cotes en cas de non-réponse MAR
Nombre de téléphones par famille	1,70
Taille de la famille administrative	1,15
20 à 29 ans	0,98
30 à 39 ans	0,87
40 ans et plus	1,00
Juif	1,04
Autre	1,00
Né en Israël	1,27
Autre	1,00

Comme prévu, le rapport de cotes de la réponse augmente à mesure que le nombre de téléphones appartenant à la famille administrative s'accroît et il en va de même pour la taille de la famille administrative. Le groupe d'âge qui affiche la plus petite probabilité de répondre est celui des personnes âgées de 30 à 39 ans (rapport de cotes=0,87) et les personnes nées en Israël enregistrent un rapport de cotes de la réponse bien plus élevé que les personnes nées à l'étranger. À partir de cette régression logistique, nous pouvons estimer la probabilité de répondre pour chaque personne.

Tableau 5-2
Répartition des probabilités estimatives de répondre à l'aide du modèle du tableau 5-1

État civil	Moyenne	5e centile	25e centile	75e centile
Autre	0,815	0,489	0,822	0,885
Divorcé	0,742	0,359	0,683	0,843
Total	0,812	0,487	0,819	0,885

Évidemment, l'hypothèse voulant que $\gamma_y = 0$ est inexacte. La probabilité de répondre parmi les personnes divorcées est bien plus basse que celle des autres personnes.

Ensuite, nous estimons les probabilités de répondre en ajoutant la variable binaire « divorcé » comme variable explicative.

Tableau 5-3
Rapports de cotes du modèle logistique estimatif des probabilités de répondre en tenant compte des cas de non-réponse NMAR

Variables	Rapport de cotes en cas de non-réponse MAR	Rapport de cotes en cas de non-réponse NMAR
Nombre de téléphones par famille	1,70	1,83
Taille de la famille administrative	1,15	1,11
20 à 29 ans	0,98	0,95
30 à 39 ans	0,87	0,86
Autre âge	1,00	1,00
Juif	1,04	1,05
Autre	1,00	1,00
Né en Israël	1,27	1,25
Autre	1,00	1,00
Divorcé	-	0,531

Comme le sous-entend déjà le tableau 5-3, le rapport de cotes de la réponse parmi les personnes divorcées est environ deux fois moins élevé que celui des autres personnes. Fait intéressant, les rapports de cotes des autres covariables sont très semblables aux rapports de cotes obtenus en supposant la non-réponse MAR.

Après avoir estimé les probabilités de répondre, nous pouvons les utiliser pour prévoir la moyenne réelle au niveau du secteur pour la variable cible (les proportions de personnes divorcées dans l'exemple actuel), à l'aide de l'estimateur approximativement sans biais sous le plan :

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

où $\pi_{j|i}$ désigne la probabilité d'échantillonnage. Sverchkov et Pfeffermann (2018) déterminent aussi indirectement le meilleur prédicteur empirique selon les modèles (7) et (8), mais nous ne tenons pas compte de ce prédicateur dans le présent document.

La figure 5-1 et les tableaux 5-2 et 5-3 comparent le rendement des trois prédicteurs suivants des proportions réelles de personnes divorcées dans les divers secteurs : la proportion de personnes divorcées dans l'échantillon observé, sans tenir compte de la non-réponse (ci-après appelée l'estimateur direct), l'estimateur obtenu en supposant la non-réponse MAR et l'estimateur obtenu en autorisant la non-réponse NMAR (équation 8).

Figure 5-1
Pourcentage des personnes divorcées dans les secteurs : valeur réelle, estimateur direct et estimateurs obtenus en supposant la non-réponse MAR et la non-réponse NMAR

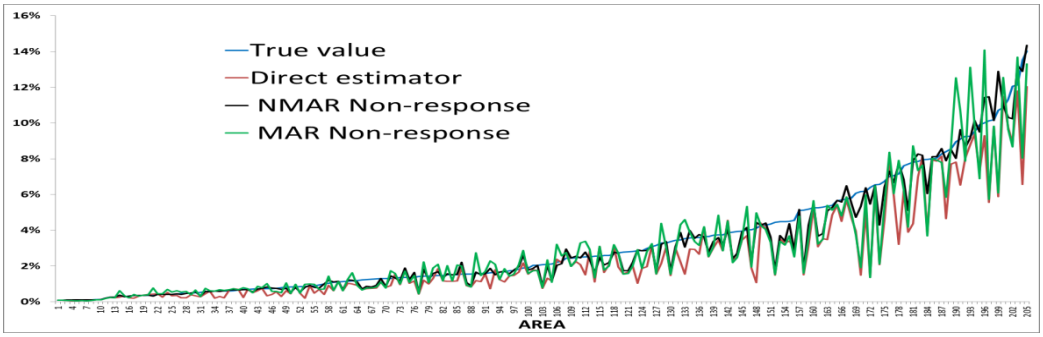


Tableau 5-4
Différence entre les valeurs réelles et les estimations (BIAIS) dans l'ensemble des secteurs

Estimateur	Moyenne	10e centile	25e centile	50e centile	75e centile	90e centile
Direct	0,0075	-0,0005	0,0006	0,0036	0,0099	0,0211
MAR	0,0033	-0,0077	-0,0018	0,0004	0,0057	0,0168
NMAR	0,0019	-0,0027	-0,0004	0,0001	0,0032	0,0094

Tableau 5-5
Distance relative absolue des estimations par rapport aux valeurs réelles

Estimateur	Moyenne	10e centile	25e centile	50e centile	75e centile	90e centile
Direct	0,270	0,042	0,121	0,233	0,406	0,551
MAR	0,256	0,032	0,113	0,216	0,379	0,472
NMAR	0,118	0,004	0,022	0,055	0,156	0,362

Comme l'indiquent clairement la figure 5.1 et les tableaux 5-4 et 5-5, les estimations obtenues en tenant compte de la non-réponse NMAR donnent de loin les biais les plus petits et la distance relative absolue la plus petite par rapport aux valeurs réelles. Les estimations directes, qui font fi de la non-réponse, donnent des biais importants et une grande distance relative par rapport aux valeurs réelles.

6. Conclusion

Dans le présent article, nous examinons une nouvelle façon d'effectuer un recensement, en combinant des estimations de l'échantillon et des mégadonnées administratives. L'un des principaux avantages de cette méthode vient du fait qu'elle ne nécessite pas d'interviews sur place, sauf pour les non-répondants. Israël n'a toujours pas de registre des logements suffisamment fiable et l'utilisation d'un échantillon sur le terrain nécessite le listage antérieur de tous les appartements dans un échantillon de cellules pour chaque secteur statistique, ce qui est plutôt compliqué du point de vue logistique, et cela coûte très cher. Il faut aussi vérifier que chacun des appartements est une unité d'habitation.

Selon la nouvelle méthode, un seul échantillon de personnes est recueilli à partir du registre, qui est connu comme étant généralement exact à l'échelon national, sauf pour certaines petites sous-populations « aberrantes », comme les Bédouins ou les immigrants illégaux. Nous étudions d'autres façons de combiner les données d'enquête et le registre de la population afin de créer un seul estimateur final du recensement, en tenant compte des erreurs d'échantillonnage de l'enquête et des erreurs dans le registre. Par ailleurs, nous proposons une procédure descriptive simple pour tester l'utilité des données-échantillons manquantes, ainsi qu'une méthode pour tenir compte des cas de non-réponse NMAR. Nous illustrons toutes les questions susmentionnées à l'aide de données empiriques réelles.
Pour le moment, nous prévoyons effectuer un « recensement pilote » l'an prochain dans deux secteurs statistiques d'Israël, ce qui devrait nous donner l'occasion de tester les idées avancées dans le présent article, avec des données plus actuelles.

Bibliographie

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

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

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

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