N° 11-522-XIF au catalogue

La série des symposiums internationaux de Statistique Canada - Recueil

Symposium 2003 : Défis reliés à la réalisation d'enquêtes pour la prochaine décennie

2003

Statistique Canada Statistics Canada

Canada

Recueil du Symposium 2003 de Statistique Canada
Défis reliés à la réalisation d'enquêtes pour la prochaine décennie

STRATÉGIE POUR L'OBTENTION D'UN SYSTÈME D'ÉCHANTILLONS AFIN D'ÉVALUER LA COUVERTURE D'UN RECENSEMENT INTÉGRÉ

Ronit Nirel, Hagit Glickman et Dan Ben-Hur[1]

RÉSUMÉ

Le Recensement de la population d'Israël de 2008 sera un recensement intégré (RI). Il sera fondé sur des sources de données administratives, que l'on complétera par des données d'enquête en vue d'estimer les erreurs de couverture des données administratives et de les corriger. Deux échantillons seront définis pour évaluer la couverture : l'échantillon U sera axé sur le sous-dénombrement et l'échantillon O, sur le surdénombrement. Le présent article décrit la création des échantillons et analyse les facteurs ayant une incidence directe sur l'étendue de la collecte des données. La méthode proposée est illustrée dans la première application du recensement intégré.

MOTS CLÉS : Couplage d'enregistrement, enquête postcensitaire, géocodage, recensement administratif, surdénombrement, sous-dénombrement.

1. INTRODUCTION

Le Recensement de la population d'Israël de 2008 sera un recensement intégré (RI). Il sera fondé sur des sources de données administratives, que l'on complétera par des données d'enquête en vue d'estimer et de corriger les erreurs de couverture (Blum, 1999). En 1995, le Bureau central de la statistique a réalisé le cinquième recensement de la population et du logement en Israël. Ce recensement était fondé sur le principe du dénombrement par secteur (un recensement habituel) et a été suivi par une enquête postcensitaire (EPC). En principe, la nouvelle approche améliorera la qualité et l'actualité des estimations, réduira le fardeau de réponse et, dans le long terme, permettra d'augmenter la fréquence des recensements. Les coûts devraient baisser et les dépenses, être réparties plus uniformément au cours du temps.

La principale source utilisée pour le recensement intégré est le registre de population (RP). Bien qu'il soit généralement assez fiable, le registre est susceptible de présenter des erreurs de couverture, dont le surdénombrement des émigrants (5 à 10 %) et la présence d'adresses périmées (10 à 30 %). Le problème du sous-dénombrement, bien connu en méthodologie des recensements, est habituellement résolu au moyen d'une enquête aréolaire sur la couverture. Le nouveau défi que pose le concept du recensement intégré est la mesure et la correction d'un surdénombrement de taille appréciable. La méthodologie repose sur un ajustement des chiffres du registre de population au moyen de deux enquêtes entrecroisées. La première, fondée sur un échantillon de secteurs, est axée sur le sous-dénombrement (échantillon U) et la seconde, fondée sur un échantillon de personnes sélectionnées à partir du fichier administratif, est axée sur le surdénombrement (échantillon O). Pour estimer les taux de sous-dénombrement, on utilise le modèle à système dual, où le premier dénombrement est fourni par les données administratives et le deuxième, par l'échantillon U. Pour estimer les taux de surdénombrement, on vérifie l'admissibilité et l'adresse le jour du recensement pour chaque personne comprise dans l'échantillon O, en utilisant des données recueillies sur le terrain, ainsi que des données administratives et la modélisation statistique.

La production d'un recensement intégré comprend cinq sous-processus : créer un fichier administratif intégré (FAI), recueillir des données au moyen d'enquêtes par sondage, coupler les deux sources, estimer les poids de recensement et évaluer les estimations a posteriori. Certains de ces processus sont utilisés plus d'une fois au cours d'un même

[1]Central Bureau of Statistics, 66 Kanfey Nesharim St., Jerusalem 95464, Israel, ronit@dev.cbs.gov.il; hagit@dev.cbs.gov.il; danb@dev.cbs.gov.il

cycle du recensement intégré. Par exemple, on procède à un couplage des données administratives et des données d'enquête après l'achèvement de chaque enquête.

Le *processus administratif* vise à améliorer les données du registre de population au moyen d'autres sources administratives. Celles-ci peuvent inclure les fichiers d'impôt municipaux, les fichiers des compagnies d'électricité et les fichiers des systèmes d'éducation. Les fichiers supplémentaires sont utilisés principalement pour valider ou pour remplacer les adresses figurant dans le registre de population. Puis, des coordonnées géographiques xy sont attribuées à chaque adresse figurant dans le fichier résultant selon une opération appelée « géocodage » (Blum et Calvo, 2001). Au cours d'un même cycle du recensement intégré, on crée trois versions du FAI :

- Le *FAI(s)* fournit les renseignements généraux sur la population et est utilisé pour la création des échantillons. Il est créé environ dix mois avant le jour du recensement.
- Le *FAI(l)* est créé peu de temps avant que se termine le travail sur le terrain pour l'échantillon U (environ un mois après le jour du recensement) et est utilisé pour définir le reste de l'échantillon O (voir la section 3). Il est fondé sur un registre de population mis à jour plusieurs semaines avant le jour du recensement.
- Le *FAI(e)* est le fichier final, qui est fondé sur un registre de population mis à jour le jour du recensement. Il est utilisé pour produire les estimations du recensement intégré (RI) et est créé environ quatre mois après le jour du recensement.

Le *processus d'échantillonnage* englobe la création des échantillons afin d'évaluer la couverture et la collecte des données. Le présent article se concentre sur la création de l'échantillon U et de l'échantillon O, ainsi que sur la relation entre ces échantillons. La section 3 décrit la méthode principale de créer les échantillons et la section 4 donne l'analyse des facteurs qui influent sur l'étendue de la collecte des données. La méthode principale de créer les échantillons repose sur la capacité de géocoder les enregistrements du FAI. Certaines localités, principalement des villes arabes et d'autres petits villages, n'ont pas d'adresse dans le FAI. À la section 5, une approche de rechange est proposée pour ces localités.

Le *processus de couplage* consiste à apparier les données du FAI aux données d'enquête selon des méthodes de couplage d'enregistrements exactes et probabilistes (Yitzkov et Azaria, 2003). Le couplage repose principalement sur un numéro d'identification personnel à neuf chiffres et est appuyé par d'autres variables, comme le nom et l'âge.

Le *processus d'estimation* commence par l'estimation des paramètres de couverture d'un modèle à système dual étendu (Glickman et coll., 2003a, 2003b). Puisque ce modèle est le fondement de la création des échantillons afin d'évaluer la couverture, nous commençons par le décrire brièvement à la section 2. Ce processus inclut aussi des évaluations diagnostiques permettant de comparer les estimations démographiques, ainsi que des corrections pour les petits domaines d'intérêt. À la fin du processus d'estimation, un poids de recensement est calculé pour chaque enregistrement du FAI. Enfin, le *processus d'évaluation* a pour but d'évaluer la qualité des estimations finales à l'aide de divers outils, dont une enquête par sondage.

La première application du recensement intégré a été réalisée en 2002 dans une petite ville voisine de Jérusalem. La section 6 illustre la méthode principale du plan d'échantillonnage à l'aide de ces données. En principe, la deuxième application aura lieu en novembre 2004. Elle portera sur neuf localités du centre d'Israël d'environ 150 000 habitants. L'article se conclut, à la section 7, par la discussion de certaines questions supplémentaires qui seront abordées durant l'application de 2004.

2. LE MODÈLE DE COUVERTURE

L'objectif principal d'un recensement de la population est de fournir des chiffres de population fiables pour de petits sous-groupes géographiques et démographiques. Pour atteindre cet objectif, on crée une liste du recensement en vue de dénombrer toutes les personnes faisant partie de la population cible dans leur secteur de résidence. Dans le cas du recensement intégré, on utilise comme liste un fichier administratif intégré ou FAI.

Les erreurs de couverture du FAI sont définies par rapport aux secteurs statistiques (SS, voir la section 3). Pour un SS donné, le sous-dénombrement est dû aux personnes qui vivent dans ce secteur, mais qui sont énumérées sur la liste d'un autre secteur. Le surdénombrement correspond aux personnes qui ont émigré et aux personnes qui sont

dénombrées dans le secteur, mais qui vivent ailleurs. Les raisons principales des différences entre les adresses du FAI et les adresses résidentielles sont les retards de mise à jour des adresses et les considérations quant à certains avantages (p. ex. secteurs où les taxes sont réduites, secteurs d'enregistrement scolaire). Il convient de souligner qu'une discordance entre une adresse du FAI et une adresse résidentielle n'est pas considérée comme une erreur de couverture si les deux adresses appartiennent au même SS. L'approche du recensement intégré a pour but de fournir des estimations démographiques directes pour les SS. Les estimations à des niveaux géographiques plus élevés (p. ex., villes) sont obtenues par sommation des estimations au niveau des SS appropriés. Dans la suite, nous nous concentrerons sur l'estimation de la taille de la population d'un SS unique, disons *E*. Les personnes admissibles vivant dans *E* sont dites admissibles dans *E* et les personnes inadmissibles ou les personnes admissibles vivant en dehors de *E* sont dites inadmissibles dans *E*.

Habituellement, le sous-dénombrement des listes du recensement est estimé au moyen du modèle multinomial à système dual bien connu (Wolter, 1986). Ce modèle s'appuie sur deux dénombrements indépendants de la population. Le premier correspond à la liste du recensement et le deuxième est habituellement fondé sur un échantillon de secteurs géographiques, appelé secteurs de dénombrement (SD). Le recensement intégré utilise les chiffres administratifs du FAI et les chiffres fondés sur les secteurs de l'échantillon U, alors l'indépendance est facilement obtenu. Une différence importante entre le recensement intégré et un recensement classique est que, dans ce dernier cas, on suppose que le surdénombrement est faible lors des deux dénombrements. D'une façon générale, cela pourrait être correct pour un dénombrement par secteur avec contrôle et vérification appropriés des données; cependant, le FAI pourrait contenir, en moyenne, jusqu'à 25 % d'enregistrements n'ayant aucun rapport avec un secteur donné. Pour estimer le surdénombrement, on vérifie l'admissibilité de tous les enregistrements du FAI énumérés dans les SD échantillonnés. L'estimation est fondée sur le modèle étendu à système dual qui tient compte du surdénombrement lors du premier dénombrement (Glickman et coll., 2003a, 2003b).

Soit *N* le nombre inconnu de personnes admissibles dans *E*, et supposons que *E* est partitionnée en *M* SD de taille $N_1,\ldots,N_M$, $\sum_{j=1}^{M} N_j = N$. Supposons que $\mathbf{Z}(j) = (Z_{11}(j), Z_{12}(j), Z_{21}(j), Z_{22}(j))$ décrit, respectivement, le nombre de personnes *admissibles dans E* comprises dans le j^{e} SD que l'on dénombre deux fois, que l'on dénombre uniquement sur la liste du FAI, que l'on dénombre uniquement sur la liste fondée sur les secteurs de dénombrement et qu'on ne dénombre sur aucune des deux listes. Soit $X(j)$ le nombre de personnes *admissibles dans E* dénombrées dans le FAI dans le j^{e} SD. Nous supposons que

$Z(j) \overset{ind}{\sim} Mult\left(N_j; p_{1+}p_{+1}, p_{1+}(1-p_{+1}), (1-p_{1+})p_{+1}, (1-p_{1+})(1-p_{+1})\right)$ est indépendant de $X(j) \overset{ind}{\sim} Poisson\left(\lambda N_j\right)$. Par conséquent, il existe deux paramètres qui caractérisent la couverture du dénombrement du FAI : p_{1+}, le taux de personnes admissibles dans *E* (paramètre de sous-dénombrement) et λ, le taux de personnes inadmissibles dans *E* (paramètre de surdénombrement). Le troisième paramètre, p_{+1}, caractérise le dénombrement fondé sur les secteurs de dénombrement. L'homogénéité peut être approximativement obtenue par une stratification démographique appropriée.

Soit *S* un échantillon aléatoire simple de *m* SD. En outre, définissons $Z_{1+}(j) = Z_{11}(j) + Z_{12}(j)$, $Z_{+1}(j) = Z_{11}(j) + Z_{21}(j)$. Les données provenant des échantillons U et O fournissent des mesures de $Z_{11}(j), Z_{1+}(j), Z_{+1}(j)$ et de $X(j)$ pour $j \in S$. L'information au sujet des SD non échantillonnés est limitée au nombre total de personnes comptées dans le FAI, $Z_{1+}(j) + X(j)$. Il convient de souligner que $X(j)$ et $Z_{1+}(j)$ ne sont pas observés pour $j \notin S$. Sous les hypothèses susmentionnées, les estimateurs asymptotiquement sans biais de N et des paramètres de couverture sont donnés par

$$\hat{N} = \frac{Z}{\hat{p}_{1+} + \hat{\lambda}}, \quad \hat{p}_{1+} = \frac{\sum_{j\in S} Z_{11}(j)}{\sum_{j\in S} Z_{+1}(j)}, \quad \hat{\lambda} = \frac{\sum_{j\in S} X(j)}{\sum_{j\in S} Z_{1+}(j) / \hat{p}_{1+}}, \tag{1}$$

où $Z = \sum_{j=1}^{M} Z_{1+}(j) + X(j)$ est le nombre total observé de personnes inscrites sur la liste dans le secteur *E* dans le FAI. L'approximation linéaire de la variance de $\hat{N}$ est donnée par

$$Var\left(\hat{N}\right) \doteq N\left[o_{1+}o_{+1} + \frac{M-m}{m}\left\{(1-r)r - o_{1+}\left(1-r-p_{+1}^{-1}\right)\right\}\right], \tag{2}$$

où $o_{\bullet} = (1 - p_{\bullet}) / p_{\bullet}$ représente le risque de ne pas être dénombré dans le dénombrement • et $r = p_{1+} / (p_{1+} + \lambda)$ est le paramètre de « rétrécissement ». La variance reflète les erreurs du modèle, ainsi que les erreurs d'échantillonnage.

3. CRÉATION DES ÉCHANTILLONS AFIN D'ÉVALUER LA COUVERTURE

Nous commençons par décrire le système de secteurs statistiques. Depuis le Recensement de 1961, le Bureau central de la statistique utilise un système de subdivision hiérarchique des localités urbaines comptant plus de 10 000 habitants en secteurs géographiques-statistiques (SS). Le système des SS couvre plus de 80 % de l'ensemble de la population d'Israël et est mis à jour avant chaque recensement. En moyenne, un SS contient de 3 000 à 4 000 personnes. Dans l'avenir, nous utiliserons l'abréviation SS pour représenter un secteur statistique, lorsque défini, ou autrement une localité.

Le but du recensement intégré est de fournir des estimations démographiques selon des sous-groupes d'âge et de sexe pour les secteurs statistiques. En prévision d'un recensement intégré, le pays est subdivisé en secteurs de dénombrement (SD). Chaque SD inclut en moyenne 170 personnes (environ 50 ménages). Les SD sont emboîtés dans les SS et sont utilisés comme unités d'échantillonnage pour les enquêtes afin d'évaluer la couverture, tel que décrit plus loin. Pour le premier recensement intégré, il a été prévu que l'échantillon sera environ un cinquième de la population (1,3 million de personnes ou environ 400 000 ménages), comme l'échantillon du « questionnaire détaillé » du Recensement de 1995. Dans le cas du recensement intégré, la fraction d'échantillonnage à l'intérieur des SS variera, contrairement à l'échantillonnage systématique uniforme utilisé en 1995 où un ménage sur cinq était sélectionné. Un échantillon aléatoire de SD sera sélectionné pour chaque SS.

Le nombre de SD dans l'échantillon est fondé sur les exigences de précision et sur le niveau des paramètres de couverture. Supposons que nous voulions estimer la taille N d'un sous-groupe dans un SS particulier, où l'homogénéité est vérifiée. Réécrivons (2) sous la forme $Var(\hat{N}) \doteq N\left(A + \frac{M-m}{m} B\right)$, avec $A = o_{1+} o_{+1}$ et $B = (1-r)r - o_{1+}(1 - r - p_{+1}^{-1})$. Pour une erreur-type relative (RSE pour *relative standard error*), $\alpha = \{\mathrm{Var}(\hat{N})\}^{1/2} / N$, la taille d'échantillon requise est

$$m(\alpha) = M \frac{B}{\alpha^2 N - A + B}. \qquad (3)$$

Pour résoudre (3), nous introduisons dans (3) les estimations des paramètres de couverture et la taille du sous-groupe. Toutefois, pour un SS et une RSE donnés, un nombre différent de SD pourrait être nécessaire pour divers sous-groupes. Habituellement, on choisit des estimations-clés pour élaborer le plan d'échantillonnage ou pour essayer d'éviter les « pires » scénarios. Dans notre cas, nous devons nous occuper des petits sous-groupes pour lesquels les taux de sous-dénombrement et de surdénombrement sont relativement élevés, comme les jeunes de 20 à 30 ans qui ont tendance à déménager fréquemment et à ne pas mettre à jour leur adresse dans le registre de population. Il convient de souligner que la variance de la taille estimée de tous SS , disons $\hat{N}$, qui est égale à la somme des estimations des sous-groupes appropriés, $\hat{N}_g$, $g = 1, \ldots, G$, est approximativement égale à $\mathrm{Var}(\hat{N}) \approx \sum_{g=1}^{G} \mathrm{Var}(\hat{N}_g) \le (N \alpha_{g_0})^2$, où g_0 est le sous-groupe dont la RSE est la plus élevée et que l'on utilise pour déterminer *m* (3). Par conséquent, $\mathrm{RSE}(\hat{N}) \le \alpha_{g_0}$.

Nous allons maintenant discuter d'un mécanisme de répartition de l'échantillon. Pour un SS i, l'équation (3) détermine la relation entre la RSE souhaitée α_i et la taille respective de l'échantillon m_i. Posons que la taille globale d'échantillon est $\boldsymbol{m} = \sum_i m(\alpha_i)$. Pour un $\boldsymbol{m}$ donné, la répartition de l'échantillon entre les divers secteurs, qui fournit une RSE uniforme sur tous les secteurs, s'obtient en calculant la racine carrée α de la fonction

$$f(\alpha) = \sum_i m_i(\alpha) - \boldsymbol{m} \qquad (4)$$

où $m_i(\alpha)$ est dérivé de (3). Si des niveaux différents de RSE sont nécessaires pour divers secteurs, ces différences peuvent être exprimées sous forme de multiples de α, c'est-à-dire $\alpha_i = c_i\alpha$, $i = 1,\ldots,I$. La répartition donnée par $m_i(c_i\alpha^*)$, où α^* est la racine de la fonction $f(\alpha) = \sum_i m_i(c_i\alpha) - \boldsymbol{m}$, permet de maintenir les RSE souhaitées.

Pour obtenir des estimations préliminaires des paramètres de couverture pour le plan d'échantillonnage, nous exécutons une analyse détaillée de la propension de la population à être enregistrée correctement dans le FAI. Le processus comporte un examen des caractéristiques socio-démographiques, telles que le pourcentage d'enfants, de jeunes, de personnes âgées, de personnes religieuses et de nouveaux immigrants. En outre, nous obtenons des estimations approximatives des paramètres de couverture a) par appariement des données du Recensement de 1995 à celles du registre de population respectif et b) d'après les données courantes sur la migration interne.

Tous les enregistrements du FAI sont géocodés, puis mis en grappes selon le SD. Les enregistrements qui ne sont pas géocodés (adresse incomplète ou inconnue) forment une strate distincte de l'échantillon O, la strate NG. Il existe également une strate distincte de l'échantillon U, appelée strate NN, pour les SD des nouveaux quartiers dont pratiquement aucun habitant n'est enregistré correctement dans le FAI. En général, on sélectionne les mêmes SD pour les deux enquêtes. Il existe toutefois deux exceptions : l'échantillon U peut comprendre un échantillon de SD provenant de la strate NN, et l'échantillon O peut inclure un échantillon d'enregistrements provenant de la strate NG. L'échantillon U est formé de toutes les personnes admissibles qui vivent dans les SD échantillonnés le jour du recensement. L'échantillon O comprend toutes les personnes qui sont énumérées dans le FAI dans les SD échantillonnés. La répartition calculée d'après (4) s'applique à la partie commune.

En ce qui concerne les opérations, les travaux sur le terrain concernant l'échantillon U débutent un jour après le jour du recensement et durent environ six semaines. Il convient de souligner qu'en aucune circonstance l'information provenant du FAI ne sera utilisée durant le travail sur le terrain relatif à l'échantillon U. Cette mesure a pour but d'assurer l'indépendance entre les dénombrements du FAI et de l'échantillon U. Puis, le fichier de l'échantillon U (FEU) est apparié au FAI afin de créer la liste des autres personnes comprises dans les SD de l'échantillon O, c'est-à-dire la liste de personnes qui sont incluses dans l'échantillon O, mais non dans le FEU. Il pourrait s'agir de personnes inadmissibles dans *E*, ou des personnes admissibles dans *E* manquées par les recenseurs de l'échantillon U, où *E* est le SS respectif. Pour achever les travaux sur le terrain relatifs à l'échantillon O, toutes les personnes figurant sur la liste des personnes restantes sont dépistées et interviewées pour déterminer leur situation. Cette partie de l'échantillon O est appelée le *reste* de l'échantillon du surdénombrement (échantillon RO). Notons que, dans cet échantillon, l'unité de dénombrement est l'individu, tandis que dans l'échantillon U, il s'agit du ménage. En principe, l'échantillon RO devrait inclure environ 20 % de l'échantillon O. L'admissibilité et l'adresse au jour du recensement sont vérifiées pour les deux échantillons.

4. FACTEURS AYANT UNE INCIDENCE SUR L'ÉTENDUE DE LA COLLECTE DES DONNÉES

Dans la présente section, nous discuterons de divers facteurs qui ont une incidence directe sur l'étendue de la collecte de données et qui sont reliés au processus du recensement intégré, ainsi qu'aux méthodes appliquées sur le terrain. Dans le cas de l'échantillon U, nous nous préoccupons principalement du sous-dénombrement des personnes admissibles durant le dénombrement sur le terrain. Dans le cas de l'échantillon O, l'étendue réelle du travail sur le terrain est reliée au processus administratif, au sous-dénombrement de l'échantillon U et aux limites du travail sur le terrain relatif à l'échantillon O.

4.1 Échantillon U

Le sous-dénombrement de l'échantillon U comprend principalement les personnes « manquées » dans les ménages recensés (p. ex. les bébés et les soldats), les personnes sortantes (c.-à-d. celles ayant déménagé après le jour du recensement), les logements manqués par les recenseurs, et les non-répondants. Nous essayons de réduire ces problèmes au minimum grâce à des questions spécifiques dans le questionnaire, un questionnaire pour les non-enquêtés et une carte postale laissée au domicile des personnes non contactées. Certaines personnes manquantes

sont dénombrées dans l'échantillon RO. En outre, diverses mesures de contrôle et de gestion sont intégrées dans le processus de dénombrement afin d'atteindre la couverture maximale pour tous les logements.

En ce qui concerne le surdénombrement, il convient de souligner que les données sur l'échantillon U sont recueillies par interview sur place assistée par ordinateur et que la partie du questionnaire concernant la couverture ressemble à un questionnaire de l'EPC. Donc, les risques de dénombrements en double, de fabrications de données et d'autres erreurs de dénombrement sont très faibles. En principe, ce genre de dénombrements erronés est traité au moyen d'un système de contrôle et de vérification du FEU.

4.2 Échantillon RO

Cycle de vie du FAI. L'échantillon RO dépend du fichier en ce sens que divers fichiers peuvent produire des « restes » différents. De toute évidence, nous aimerions que l'échantillon RO soit défini d'après le FAI(e). En pratique, il est fondé sur le FAI(l). Par conséquent, le groupe final de personnes restantes définies d'après le FEU qui est utilisé pour l'estimation est légèrement différent de celui utilisé pour la collecte des données. Autrement dit, certaines personnes recensées ne figurent pas sur la liste du secteur *E* dans le FAI(e) (interviews « perdues »), tandis que d'autres auraient dues être recensées, mais n'ont pas été contactées sur le terrain. Le remède proposé consiste à essayer de réduire au minimum le décalage temporel entre le FAI(e) et le FAI(l), et à utiliser des méthodes d'imputation de pointe pour les personnes non incluses lors du travail sur le terrain.

Choix de l'adresse. L'adresse figurant dans le FAI est le produit d'un algorithme qui évalue un groupe d'adresses pour chaque personne. Toute adresse à laquelle est attribuée un score plus élevé qu'un seuil préétabli vient remplacer celle qui figure dans le registre de population. Si une personne vit dans E_0 et qu'une des adresses possibles est également dans E_0, mais que l'adresse *choisie* est dans E_1, alors cette personne est ajoutée à l'échantillon RO de E_1. Pour réduire au minimum cet effet, on examine minutieusement la fiabilité des sources autres que le registre de population avant de remplacer une adresse par une autre.

Géocodage des adresses. La définition de l'échantillon O s'appuie fortement sur le processus de géocodage, qui permet de repérer les adresses du FAI appartenant à un secteur statistique *E*. Une adresse comprise dans un SD échantillonné qui est géocodée incorrectement dans *E* agrandit l'échantillon RO. Il convient de souligner que le géocodage est généralement effectué au niveau de l'immeuble et qu'un géocodage de bonne qualité à ce niveau exige un effort considérable. Comme nous nous préoccupons surtout des erreurs provenant d'un mouvement des adresses entre les SS et les SD, nous accordons une haute priorité au géocodage de bonne qualité à ces niveaux.

Appariement du FEU et du FAI. Toutes les personnes non appariées énumérées dans l'échantillon O sont incluses dans l'échantillon RO. Par conséquent, les valeurs manquantes dans le FEU et les erreurs de mesure pourraient réduire le nombre d'appariements corrects, donc augmenter la taille de l'échantillon RO. Bien que nous nous efforcions de ne pas augmenter cet échantillon inutilement, nous ne nous préoccupons surtout des appariements incorrects et, par conséquent, utilisons une erreur d'appariement assez conservatrice de type I (couplage des non appariements).

Difficultés sur le terrain. La collecte des données auprès de l'échantillon RO débute environ sept semaines après le jour du recensement. Pour chaque personne figurant sur la liste, nous vérifions l'admissibilité et l'adresse le jour du recensement afin d'obtenir une estimation du paramètre de surdénombrement λ. Nous répartissons l'échantillon en trois catégories de difficulté de dénombrement : A- facile, B- raisonnable et C- difficile. Les personnes de la catégorie A vivent habituellement dans des ménages qui ont déjà été interviewés dans l'échantillon U. L'appariement entre le FEU et le FAI les met en évidence et ainsi il est assez facile de les trouver. Nous pensons aussi que les ménages occupant des logements manqués sont faciles à trouver. Les personnes de la catégorie B sont habituellement celles que l'on peut retrouver dans d'autres parties du pays (p. ex., personnes ayant déménagé). Le noyau de la catégorie C comprend les émigrants, particulièrement ceux qui sont partis il y a de nombreuses années. Le processus de collecte de données est planifié de façon à tenir compte de ces trois catégories de personnes Premièrement, les recenseurs retournent dans les SD de l'échantillon U avec une liste correspondant à l'échantillon RO. En principe, durant cette phase, la plupart des personnes de la catégorie A sont dénombrées et certains renseignements au sujet des deux autres catégories sont recueillis. Cette phase dure environ deux semaines. Deuxièmement, des numéros de téléphone sont ajoutés au reste de la liste, dans la mesure du possible et les recenseurs essayent d'avoir un contact avec ces personnes par téléphone. Enfin, les cas qui restent font l'objet d'un

processus de dépistage minutieux en se servant de diverses sources d'information. On s'attend que 20 % à 30 % de l'échantillon RO soit des émigrants de longue date. Un grand nombre d'entre eux ne seront pas dépistés. À l'heure actuelle, nous menons un projet national ayant pour but de prédire, pour chaque personne figurant dans le registre de population, la propension à devenir un émigrant. Nous sommes en train de construire un modèle fondé sur les renseignements recueillis lors des contrôles frontaliers, les recensements antérieurs, les relations familiales, ainsi que diverses variables démographiques et socioéconomiques. Nous estimons que ce modèle nous aidera à résoudre les cas les plus difficiles de l'échantillon RO.

5. LOCALITÉS NE POSSÉDANT PAS DE SYSTÈME D'ADRESSES

La méthode principale du plan d'échantillonnage s'appuie sur le géocodage des enregistrements du FAI au moins au niveau du SS : le géocodage nous permet de repérer les enregistrements du FAI qui appartiennent à un SD échantillonné, donc à dénombrer la plupart des personnes figurant sur la liste de l'échantillon O durant le travail sur le terrain relatif à l'échantillon U. Certaines localités, principalement des villes arabes et d'autres petits villages, n'ont pas de système d'adresses. Pour les personnes habitant ces localités, les zones réservées à l'adresse dans le fichier administratif intégré contiennent uniquement le nom de la localité.

Nous classons ces localités en trois catégories :

- *Localités entièrement recensées*. Pour les localités entièrement recensées comprises dans l'échantillon U (principalement celles comptant trois à quatre SD), nous pouvons définir le reste de l'information manquante sur la localité dans le FAI et appliquer la méthode principale.
- *Localités pour lesquelles l'information de 1995 est utilisable*. Pour les localités arabes qui n'ont pas subi une grande expansion depuis le Recensement de 1995 (c.-à-d. qui ne comptent pas de nouveaux quartiers), les coordonnées géographiques peuvent être obtenues d'après le fichier du Recensement de 1995 et améliorées grâce à des mises à jour appropriées. Cette option est envisagée pour la population arabe, parce que celle-ci est assez peu mobile et que les jeunes couples ont tendance à vivre avec la famille du mari. Nous couplerons le fichier du Recensement de 1995 au FAI en utilisant les numéros d'identification, puis nous ajouterons l'information géographique provenant du dénombrement sur le terrain de 1995 dans le FAI et nous mettrons à jour les données sur les relations familiales afin d'y ajouter les nouvelles relations créées depuis 1995 (mariages et enfants). Une fois que les renseignements géographiques seront inclus dans le FAI et que l'utilité des données aura été confirmée, ce genre de localité pourra être traité selon la procédure principale.
- *Autres localités*. Cette catégorie comprend toutes les autres localités n'ayant pas de système d'adresses.

Pour les localités de cette dernière catégorie, nous sélectionnerons deux échantillons *indépendants* afin d'évaluer la couverture. Pour l'échantillon U, nous subdiviserons le secteur de la localité en SD en nous servant de données telles que le volume de bâtiments résidentiels extrait du registre des bâtiments. Nous sélectionnerons un échantillon de SD et dénombrerons toutes les personnes admissibles. Cet échantillon fournira une estimation du paramètre de sous-dénombrement p_{1+} de façon comparable à l'estimation proposée en (1) : l'estimation produite est le nombre de personnes qui sont comptées dans l'échantillon et énumérées dans le FAI pour la localité en question par rapport au nombre total de personnes dans l'échantillon. Il convient de souligner que, pour les localités qui sont subdivisées en SS, l'admissibilité dans le FAI est déterminée au niveau de la *localité* plutôt qu'au *niveau du SS*, comme dans le cas du processus principal; autrement dit, puisque nous ne pouvons identifier les SS dans le FAI, une personne est comptée comme étant énumérée dans le FAI si elle vit n'importe où dans la localité et non spécifiquement dans un SS. Cependant, nous obtenons des estimations distinctes du sous-dénombrement pour chaque SS.

Pour l'échantillon O, nous sélectionnons un échantillon stratifié indépendant de personnes. Les strates sont définies d'après des variables démographiques, comme le groupe d'âge, qui sont reliées à la propension à être admissible au niveau de la localité. Le paramètre du surdénombrement λ est estimé par le nombre de personnes inadmissibles dans l'échantillon par rapport au nombre de personnes admissibles dans l'échantillon O, *corrigé pour le sous-dénombrement dans le FAI*, tel qu'estimé au moyen de l'échantillon U. Enfin, pour les localités qui sont subdivisées en SS, l'échantillon O fournira aussi une estimation de la répartition de la population par SS. Les estimations sont égales au nombre de personnes admissibles dans chaque SS par rapport au nombre total de personnes admissibles dans la localité.

6. CRÉATION DES ÉCHANTILLONS AFIN D'ÉVALUER LA COUVERTURE POUR LA PREMIÈRE APPLICATION

Le modèle du recensement intégré a été mis en application pour la première fois en mai 2002 pour une ville voisine de Jérusalem. La population de cette ville est passée d'environ 25 000 habitants à la fin de 1995 à plus de 50 000 au moment de l'application. Environ 17 000 nouveaux venus peuplaient de nouveaux quartiers. En outre, la population est hétérogène et comprend, entre autres, de nouveaux immigrants provenant de l'ancienne Union soviétique et des résidents ultraorthodoxes. Par conséquent, nous nous attendions à ce que l'application fournisse un aperçu de toute une gamme de profils d'enregistrement dans les fichiers administratifs.

La ville comprend 12 SS, qui sont subdivisés en 247 SD (tableau 1). Les exigences concernant le plan d'échantillonnage sont qu'environ 50 SD doivent être sélectionnés et que la précision des estimations démographiques doit être à peu près la même pour tous les sous-groupes d'âge et sexe à l'intérieur des SS. Aux fins de l'application, il était également requis de sélectionner dans chaque SS un échantillon de taille minimale égale à trois SD. L'application de la formule (3) de répartition de l'échantillon nécessite certaines estimations de la taille de la population et des paramètres de couverture. Nous avons estimé la taille de la population à l'aide des totaux du FAI(s). Pour estimer les paramètres de couverture, nous avons apparié le fichier du Recensement de 1995 au fichier du registre de population de 1995. Nous avons construit des modèles de régression logistique pour les données de 1995 afin de prédire p_{1+}, p_{+1} et $r = p_{1+}/(p_{1+} + \lambda)$. Les variables explicatives incluses dans ces modèles comprennent des indicateurs pour les quartiers ultrareligieux, l'immigration et plusieurs groupes d'âge. Pour les nouveaux quartiers pour lesquels on ne dispose d'aucune donnée pour 1995, nous avons extrapolé les estimations en nous fondant sur d'autres secteurs dont les caractéristiques socio-démographiques sont comparables. Les colonnes (2) à (5) du tableau 1 montrent les valeurs des paramètres qui ont été utilisées pour le plan d'échantillonnage.

L'étape suivante consiste à déterminer le niveau de RSE qui satisfait globalement à la contrainte de taille de l'échantillon en calculant la racine carrée de $f(\alpha)$ de l'équation (4). La figure 1 donne la fonction $f(\alpha)$ et sa racine carrée pour divers scénarios ($\boldsymbol{m}=50$). La partie (a) montre la sensibilité de la racine carrée à des variations de ±10 % de la valeur du p_{1+} de la colonne (4) du tableau, tous les autres paramètres étant fixés à leur valeur dans le tableau. Les RSE résultantes sont égales à 2,1 % pour les valeurs de référence (ligne du milieu de la figure), 1,6 % pour les valeurs supérieures (ligne inférieure) et 2,5 % pour les valeurs inférieures. Donc, la sensibilité de la RSE est faible pour une variation relativement importante de la valeur de f p_{1+}. La partie (b) de la figure montre la sensibilité de la racine carrée aux variations de N (colonne (2) du tableau), tous les autres paramètres étant fixés à leur valeur dans le tableau. Nous voyons que, pour les sous-groupes dont la taille est d'environ $N/8$ (p. ex. quatre groupes d'âge par sexe), la RSE résultante est de 5,8 % et que pour $N/16$, elle passe à 8,3 %.

Nous calculons la répartition au moyen de l'équation (3). La colonne (7) du tableau illustre la répartition résultante pour les estimations au niveau des SS, en utilisant les valeurs des paramètres figurant dans le tableau. La répartition réelle (colonne (8)) est fondée sur une analyse et des considérations plus poussées, mais elle est fort semblable aux résultats de la colonne précédente. Un quartier entièrement nouveau, autoreprésentative, a été ajouté à l'échantillon. Enfin, les estimations résultantes et leur RSE sont présentées, respectivement, aux colonnes (9) et (10). En général, les RSE sont conformes aux exigences du plan d'échantillonnage. Pour le SS n° 5, la RSE est nettement plus faible qu'il n'est requis (0,9 %), ce qui donne à penser qu'un plus petit échantillon aurait suffi pour ce secteur.

7. CONCLUSION

Le présent article traite des caractéristiques principales des échantillons afin d'évaluer la couverture dans le contexte du recensement intégré en Israël. D'autres idées seront examinées lors de futures applications. Par exemple, nous avons l'intention d'évaluer la faisabilité d'un plan d'échantillonnage à deux étapes, où les SS sont stratifiés en sous-groupes définis selon les conseils régionaux ou les localités, selon le cas, et un échantillon de SS est sélectionné à l'intérieur de ces strates.

La population cible comprend toutes les personnes vivant en Israël depuis plus d'un an le jour du recensement, si bien qu'une partie des travailleurs étrangers et d'autres non-Israéliens devrait être recensés. Comme le registre de

population ne couvre pas ces personnes et qu'il n'existe aucune autre donnée fiable à leur sujet, nous avons décidé de séparer ce groupe spécial de la population cible et d'estimer sa taille directement d'après les données recueillies auprès de l'échantillon U. Donc, nous augmenterons les fractions d'échantillonnage dans les secteurs où la proportion de travailleurs étrangers est forte afin de favoriser la production d'estimations fiables pour cette population.

Un autre objectif important consiste à fournir des renseignements socioéconomiques. Les préoccupations concernant la qualité de ces estimations seront également intégrées dans le plan de sondage. Et en dernier, mais non le moindre, il convient de mentionner que les personnes placées en établissement sont entièrement recensées dans le cadre d'une opération distincte. Ces personnes sont déduites du FAI pour l'estimation des paramètres de couverture, puis rajoutées par la suite.

RÉFÉRENCES

Blum, O. (1999). "Combining Register-Based and Traditional Census Processes as a Pre-defined Strategy in Census Planning", *Proceedings of the Federal Committee on Statistical Methodology (FCSM) Research Conference*. http://www.fcsm.gov/99papers/blum2.html

Blum, O. et R. Calvo (2001). "Geospatial Data Collection and Analysis as Crucial Processes in an Integrated Census", *Proceedings of the Federal Committee on Statistical Methodology (FCSM) Research Conference*. http://www.fcsm.gov/01papers/Blum.pdf

Glickman, H., R. Nirel et D. Ben-Hur (2003a). "False Captures in Capture-Recapture Experiments with Application to Census Adjustment", paper presented at the 54th Biennial Session of the International Statistical Institute, Berlin, Germany.

Glickman, H., R. Nirel et D. Ben-Hur (2003b). "Estimation of Population Size Based on Contaminated Capture-Recapture Data with Application to Census Adjustment", in preparation.

Wolter, K. M. (1986). "Some Coverage Error Models for Census Data". *Journal of the American Statistical Association*, 81, pp. 338-346.

Yitzkov, T., et H. Azaria (2003), "Record Linkage in an Integrated Census", *Proceedings of the Federal Committee on Statistical Methodology (FCSM) Research Conference*.

Tableau 1 : Création des échantillons afin d'évaluer la couverture pour l'application de 2002. Taille de la population, *N(s)*, valeurs des paramètres utilisées pour l'établissement du plan d'échantillonnage, nombre de SS, *M*, répartition proposée pour α=0,02 et pour un échantillon de taille *m=50 SD*, répartition actuelle, estimations résultantes et leur RSE

(1) SS	*(2)* *N(s)*	*(3)* p_{+1}	*(4)* p_{1+}	*(5)* *r*	*(6)* *M*	*(7)* *m(0,02)*	*(8)* *m* réel	*(9)* $\hat{N}$	*(10)* $\hat{\alpha}$
1	2 670	0,85	0,80	0,75	12	4	3	2 337	0,018
2	4 705	0,85	0,75	0,80	21	4	4	4 689	0,017
3	4 411	0,85	0,85	0,75	20	3	4	3 559	0,021
4	3 747	0,90	0,75	0,90	19	4	4	4 819	0,019
5	6 323	0,50	0,80	0,95	31	6	8	5 838	0,009
6	4 062	0,92	0,70	0,85	20	5	4	3 355	0,016
7	2 155	0,90	0,75	0,90	10	3	3	2 537	0,014
8	1 390	0,93	0,85	0,70	5	2	3	1 323	0,012
9	2 456	0,91	0,70	0,90	12	4	4	2 453	0,015
10	4 959	0,93	0,85	0,75	24	3	3	4 526	0,020
11	4 531	0,80	0,70	0,95	26	6	5	16 804	0,009
12	9178	0,80	0,70	0,95	47	6	7		
Total	50 587				247	50	52	52 240	0,005

Figure 1 : Sensibilité de la racine carrée de *f(α)* selon : a) les variations de p_{1+} et b) les variations de *N* (*m=50*).