Techniques d'enquête, décembre 1985
Vol. 11, n° 2, pp. 125–133
Statistique Canada

Échantillonnage des questionnaires manuscrits de recensement reproduits sur microfilm

D.R. BELLHOUSE[1]

RÉSUMÉ

Dans la première partie du document, nous procédons à une revue rétrospective des travaux d'historiens sur les dossiers manuscrits de recensement microfilmés. Les historiens ont utilisé plusieurs types de plan d'échantillonnage qui vont, par ordre croissant de complexité, de l'échantillonnage aléatoire en grappes et stratifié jusqu'à l'échantillonnage en grappes à deux degrés stratifié. Dans la deuxième partie, nous proposons une méthode permettant de créer une bande-échantillon à grande diffusion qui contiendrait des données du recensement du Canada de 1881. Cette recherche faisait partie d'un projet pilote exécuté pour les Archives publiques du Canada et a été réalisée par le Centre de données sur les sciences sociales de l'Université Western Ontario. Le projet pilote avait pour but de déterminer s'il était avantageux et possible, du point de vue économique et technique, de construire une base de données ordinolingue à l'aide des questionnaires microfilmés du recensement du Canada de 1881.

MOTS CLÉS: Échantillonnage aléatoire informatisé; dossiers microfilmés; sondages à plusieurs degrés; échantillons à grande diffusion; stratification.

1. INTRODUCTION

Pour écrire l'histoire d'une personne ou d'un peuple, l'historien a besoin de sources. De nos jours, beaucoup d'historiens veulent retracer la vie de l'*homme de la rue*. Pour mener à bien leur recherche, ils peuvent se servir de documents comme les questionnaires de recensement, les titres de propriété et les annuaires commerciaux. Dans le présent document, nous nous penchons sur l'utilisation des questionnaires de recensement comme source. Le principal inconvénient que présente l'utilisation de données de recensement est l'abondance de ces données. L'historien qui dispose d'un budget de recherche normal n'a ni le temps ni les ressources financières ou humaines nécessaires pour passer en revue tous les questionnaires du recensement. Pour contourner cette difficulté, il doit former un échantillon aléatoire de questionnaires. La plupart des questionnaires de recensement que peut consulter l'historien sont reproduits sur microfilm. Au Canada, les questionnaires reproduits sur microfilm sont ceux des recensements coloniaux de 1841, de 1851 et de 1861 et ceux des recensements du Canada de 1871 et de 1881. Le problème se résume donc pour l'historien à déterminer le plan de sondage approprié pour prélever un échantillon de questionnaires microfilmés.

Dans la deuxième section du document, nous passons en revue les méthodes d'échantillonnage appliquées par les historiens. L'application de ces méthodes a donné des résultats très inégaux. Dans certains cas, les résultats ont été très satisfaisants, les historiens ayant su adapter l'application des méthodes à l'objet de la recherche. Dans d'autres cas, toutefois, il semble que les historiens aient utilisé des plans de sondage inutilement complexes. Un plan de sondage complexe peut avoir des effets qui s'écartent sensiblement de 1 et, par conséquent, compliquer l'analyse des données. Voir, par exemple, Rao et Scott (1981) et Holt *et coll.* (1980) pour des commentaires sur l'analyse de données qualitatives, et Scott et Holt (1982) pour des commentaires sur l'analyse par régression. Enfin, les rapports de plusieurs des enquêtes analysées ci-dessous ne contiennent pas assez de renseignements pour nous permettre d'évaluer les raisons qui ont justifié le choix d'un plan de sondage particulier.

[1] D.R. Bellhouse, Département des sciences statistiques et actuarielles, Université Western Ontario, London Ontario), Canada N6A 5B9.

Dans la troisième partie de l'exposé, nous proposons une méthode permettant de prélever des questionnaires du recensement du Canada de 1881 dans le but de créer une bande-échantillon à grande diffusion. Cette recherche faisait partie d'un projet exécuté pour les Archives publiques du Canada. Elle a été confiée par contrat au Centre de données sur les sciences sociales de l'Université Western Ontario. Nous donnons ici une description du plan d'échantillonnage. On trouvera un rapport complet sur le projet dans Mitchell *et coll.* (1982). Le plan d'échantillonnage utilisé ressemble à certains égards aux plans qui ont servi à la création des bandes-échantillons à grande diffusion pour les recensements de 1971 et de 1976. Les plans d'échantillonnage reposent tous sur la stratification; pour le recensement de 1881, toutefois, la stratification n'a pu se faire qu'en fonction d'une répartition géographique.

2. REVUE RÉTROSPECTIVE

Les travaux portant sur l'échantillonnage des documents manuscrits de recensement peuvent être classés suivant la méthode d'échantillonnage utilisée. Nous décrivons ci-dessous les diverses méthodes utilisées par ordre croissant de complexité du plan de sondage.

2.1 Échantillonnage en grappes

Ornstein et Darroch (1978) ont proposé une méthode simple et économique permettant d'échantillonner des dossiers de recensement et de les raccorder d'une période à une autre. Cette méthode consiste essentiellement à former des grappes de noms de famille et à constituer des échantillons à partir de ces grappes. Les grappes sont désignées par la première lettre du nom de famille. Si les mêmes grappes sont échantillonnées dans divers recensements à la fois, une personne dont le nom figure dans plus d'un recensement fera partie de l'échantillon choisi. Il y a donc moins de cas à analyser pour la liaison et le coût est par conséquent moins élevé. Ce plan de sondage se prête particulièrement bien aux études chronologiques de la migration ou de l'évolution démographique.

2.2 Échantillonnage stratifié

Aucun des plans de sondage avec stratification considérés dans la présente section ne prévoit une répartition optimale de l'échantillon. Cette situation s'explique par le fait qu'aucun des historiens n'était en mesure de connaître a priori les variations à l'intérieur des strates. Pour obtenir ce genre de renseignements, il aurait fallu supporter une hausse sensible du coût de chaque projet.

Hammarberg (1971) a utilisé une méthode de sondage à deux phases, ou méthode d'échantillonnage double, dans l'espoir de réduire le biais engendré par l'échantillonnage d'un ensemble incomplet de dossiers. Les dossiers échantillonnés dans la deuxième phase ont été les annuaires commerciaux de neuf comtés de l'Indiana. Dans la première phase du sondage, Hammarberg a prélevé un échantillon dans un ensemble de dossiers supposé complet, le recensement des États-Unis de 1870. Il a appliqué la méthode d'échantillonnage aléatoire stratifié avec répartition proportionnelle de sorte que l'échantillon soit autopondéré. Les strates correspondaient aux neuf comtés de l'Indiana. On trouve deux aspects de cette recherche dans des études ultérieures sur l'échantillonnage de documents anciens. Les strates correspondent à des régions géographiques et l'échantillon est autopondéré.

Hammarberg (1971) a également appliqué le test du khi carré classique à certaines variables pour vérifier dans quelle mesure la distribution des données de l'échantillon se rapprochait des distributions de population établies à partir des données du recensement. Dans beaucoup d'autres études, on ne s'est pas préoccupé de vérifier la *représentativité* de l'échantillon.

Soltow (1975) s'est servi d'échantillons des recensements américains de 1850, de 1860 et de 1870 pour faire une étude de la richesse aux États-Unis. Pour chaque année de recensement, il a prélevé un échantillon sur chaque bobine de microfilm de telle sorte que l'échantillon soit stratifié par bobine, ce qui équivalait à peu près à une stratification géographique. Le plan de sondage de Soltow semble correspondre à un échantillonnage systématique. Pour prélever un échantillon, il a déterminé un point sur l'écran de la visionneuse de microfilms puis a visionné le film. Il faisait avancer la pellicule par demi-tours successifs de manivelle jusqu'à ce qu'un questionnaire de recensement soit relativement centré sur le point désigné sur l'écran. Pour être sélectionné, le questionnaire devait concerner une personne de sexe masculin âgée de 20 ans ou plus. En outre, l'échantillon du recensement de 1860 comprenait 40 fois plus de personnes dont l'avoir s'élevait à $100,000 ou plus que de personnes dont l'avoir était inférieur à $100,000 (p. 5), de sorte que cet échantillon n'était pas autopondéré. Bien que Soltow n'en fasse pas mention, il peut avoir voulu *suréchantillonner* les personnes plus fortunées de manière à disposer d'un échantillon suffisamment grand pour lui permettre d'établir des comparaisons avec les classes moins aisées de la société. Il a également comparé les observations de ses échantillons aux distributions de fréquence publiées mais n'a effectué aucun test de validité de l'ajustement. Il a constaté que pour diverses variables, les données des échantillons se rapprochaient sensiblement de celles des recensements sur le plan des moyennes et des proportions. Ces observations s'appliquaient même à des variables comme le patrimoine moyen, chose surprenante compte tenu du suréchantillonnage de personnes plus fortunées et du fait que l'estimation de Soltow semble correspondre à la moyenne de l'échantillon.

Darroch et Ornstein (1980) ont utilisé un échantillon du recensement du Canada de 1871 pour étudier la relation entre l'origine ethnique et la profession. La méthode d'échantillonnage utilisée est décrite dans Ornstein (1978). Pour les besoins des deux études, il a fallu suréchantillonner certains groupes ethniques de manière à ne pas obtenir un échantillon autopondéré. Ne tenant pas compte de ce suréchantillonnage, les deux historiens ont appliqué un échantillonnage aléatoire stratifié, la stratification étant fondée sur la structure géographico-hiérarchique des dossiers du recensement: provinces, districts, sous-districts et divisions. La division correspond au secteur de dénombrement actuel et semble être le critère naturel de stratification. Toutefois, Ornstein (1978) l'a subdivisée en strates et a prélevé deux ménages dans chaque strate. Il ne dit pas comment il opère cette subdivision mais il la justifie en affirmant que l'échantillonnage de deux unités par strate réduit au minimum la variance des estimations de certaines valeurs d'une population. Bien qu'Ornstein (1978) ne le précise pas, il semble qu'il ait voulu accroître l'efficacité de la stratification en formant des strates à l'intérieur d'une division aussi homogène que possible. Il s'est ainsi trouvé à accroître le coût de l'échantillonnage. Enfin, sa méthode exigeait que l'on fasse défiler le film au moins deux fois dans la visionneuse, la première fois pour obtenir le nombre de ménages par division et la deuxième pour échantillonner les ménages.

Dans leurs ouvrages, Johnson (1978b) et Graham (1980) indiquent comment ils ont obtenu un échantillon à grande diffusion du recensement des États-Unis de 1900. Dans un autre ouvrage, Johnson (1978a) reprend à peu près les mêmes éléments pour décrire comment il a échantillonné les questionnaires du recensement du Rhode Island de 1860. Dans les trois cas, l'échantillon est formé en déterminant au hasard des lignes sur le microfilm et en cherchant par la suite ces lignes à l'aide d'une visionneuse munie d'un compteur. Compte tenu de la méthode d'échantillonnage, la taille globale de l'échantillon est aléatoire. Graham (1980, p. 41) donne un certain nombre de critères permettant d'accepter ou de rejeter les lignes échantillonnées. Il s'agit d'un échantillonnage aléatoire stratifié et les bobines de microfilms servent de strates. La stratification est fondée sur une répartition géographique dans la mesure où les questionnaires de recensement d'une même région se trouvent tous sur la même bobine de microfilms. Cette méthode a l'avantage de rendre l'exécution efficace et de ne nécessiter

qu'un seul visionnement. Elle élimine en outre le problème des strates vides ou des strates à une unité lorsque la fraction de sondage pour une strate est faible. En revanche, comme elle ne comporte qu'un seul visionnement, il faut pouvoir résoudre de façon ponctuelle les principaux problèmes qui peuvent survenir.

2.3 Échantillonnage en grappes stratifié

Bateman et Foust (1974) ont obtenu un échantillon des exploitations agricoles du nord des États-Unis à l'aide des données du recensement de 1860. Ils ont divisé le Nord en deux strates, est et ouest, et prélevé un échantillon aléatoire de comtés ruraux dans chaque strate. Ensuite, ils ont choisi au hasard, dans chaque comté, une commune rurale (grappe) et recueilli des données sur toutes les exploitations agricoles situées sur le territoire de la commune. Une des raisons du choix de l'échantillonnage en grappes est que cette méthode est économique. Comme les données sur les exploitations agricoles provenaient du recensement de l'agriculture et que les données démographiques sur les propriétaires de ces exploitations agricoles et ceux qui y travaillent provenaient du recensement de la population, il était plus facile de faire le lien entre les exploitations agricoles et leurs propriétaires respectifs en se limitant à une commune. Swierenga (1983) donne une deuxième raison pour justifier l'utilisation de l'échantillonnage en grappes. Il affirme que les données recueillies pour une commune ont permis d'estimer la productivité globale des facteurs en agriculture et de définir toute la main-d'oeuvre agricole, y compris les travailleurs agricoles qui demeurent à l'extérieur des 12,000 exploitations comprises dans l'échantillon (p. 793). Comme les grappes (communes) n'ont pas été choisies selon une probabilité proportionnelle à la taille, le plan de sondage n'a pas produit d'échantillons autopondérés.

Bateman and Foust (1974) ont aussi appliqué quelques tests pour vérifier la représentativité de leur échantillon. À l'instar d'Hammarberg (1971), ils ont appliqué le test du khi carré pour comparer les observations de l'échantillon aux observations probables de la population. Dans le cas des variables continues, ils ont utilisé le test t. Les estimations de la moyenne et de la variance étaient des estimations *simples*, c'est-à-dire qu'elles n'étaient pas fondées sur le plan de sondage.

2.4 Échantillonnage à deux degrés stratifié

Hammarberg (1977) a appliqué une méthode d'échantillonnage à deux degrés stratifié pour échantillonner des ménages dans le recensement du territoire de l'Utah de 1880. Les strates sont un amalgame assez complexe de cinq régions géographiques de l'Utah, de quelques comtés de régions populeuses et de quelques grandes municipalités. Hammarberg a prélevé dans chaque strate un échantillon de municipalités ou de wards. Les municipalités qui constituaient déjà des strates étaient automatiquement incluses dans l'échantillon. Les wards sont à l'Église mormone ce que les paroisses étaient à l'Église chrétienne médiévale. Un échantillon de ménages a ensuite été prélevé dans les municipalités ou les wards choisis. Cet échantillon était autopondéré. L'argument que fait valoir Hammarberg à la page 460 de son ouvrage pour justifier ce mode de stratification est convaincant:

> "Comme le mode d'organisation fondamental de la population repose sur une structure géographique et que la plupart des documents officiels - civils et religieux - correspondent à cette structure, on peut dire qu'un échantillonnage de la population suivant une répartition géographique équivaut dans une large mesure à un échantillonnage des dossiers produits pour cette population."

McInnis (1977) a aussi eu recours à l'échantillonnage à deux degrés stratifié pour obtenir un échantillon des dossiers du recensement colonial de 1861. Il étudiait alors la relation entre

le nombre d'enfants par famille et l'abondance des terres dans certaines régions. Il a commencé par répartir environ 300 cantons en strates selon l'année de colonisation. Il a ensuite prélevé un échantillon de cantons dans les strates et des échantillons de fermes dans les cantons. Il semble que McInnis ait choisi l'échantillonnage à deux degrés pour des raisons d'économie. En effet, comme les fermes échantillonnées étaient ensuite appariées au dossier correspondant dans le recensement de l'agriculture, il était moins long et, par conséquent, moins coûteux d'échantillonner quelques cantons et d'apparier les dossiers de plusieurs fermes d'un canton que de stratifier les cantons et d'apparier les dossiers d'un petit nombre de fermes dans chaque strate. Le même raisonnement s'applique aux travaux d'Hammarberg (1977). Lui aussi rattachait d'autres dossiers au ménage échantillonné.

2.5 Échantillonnage en grappes à deux degrés stratifié

Smith (1978) a appliqué une méthode d'échantillonnage en grappes à deux degrés stratifié pour étudier la population âgée dans le recensement des États-Unis de 1900. Les strates sont définies comme les régions de recensement, les comtés qui forment les régions étant les unités primaires d'échantillonnage. Celles-ci sont choisies selon une probabilité proportionnelle à la taille de leur population. Pour chaque comté, Smith a prélevé plusieurs pages de questionnaires du recensement. Il a ensuite relevé les noms des personnes de plus de 50 ans qui figuraient sur chacune des pages échantillonnées. L'échantillonnage en grappes était nécessaire du fait qu'il aurait été trop coûteux de déterminer toutes les personnes qui pouvaient être échantillonnées. Smith tente par la même occasion de comparer quelques distributions d'échantillons aux données publiées du recensement. Il se sert pour cela de la fonction des observations normalement utilisée dans les tests d'hypothèses portant sur une proportion simple même s'il s'agit de données multinomiales.

Foust (1968, chap. 2) décrit un deuxième cas d'échantillonnage en grappes à deux degrés stratifié. L'échantillon, dit échantillon Parker-Gallman, a été tiré du recensement des États-Unis de 1860 pour étudier les régions productrices de coton du sud du pays. Les strates étaient 405 *régions cotonnières* du Sud, où l'on avait produit au moins 1,000 balles de coton de 400 livres dans les douze mois qui précédaient le jour du recensement. Pour chaque région, on a prélevé un échantillon aléatoire systématique de pages de documents manuscrits du recensement, et un groupe de cinq plantations a été choisi au hasard sur une page donnée, le groupe étant considéré comme une grappe. On a recouru à l'échantillonnage en grappes car il fallait consulter trois questionnaires de recensement différents pour recueillir des données sur une plantation particulière. L'appariement des questionnaires a été qualifié de très laborieux. Fogel et Engerman (1974, p. 22-25) énumèrent plusieurs autres échantillons se rattachant à l'échantillon Parker-Gallman. Bode et Ginter (1984) formulent des critiques sur le contenu de l'échantillon.

De tous les échantillons considérés ici, l'échantillon Parker-Gallman et les échantillons de Bateman et Foust (1974) sont ceux qui ont été le plus étudiés. Swierenga (1983) a passé en revue une bonne partie des travaux fondés sur ces échantillons.

3. ÉCHANTILLONS À GRANDE DIFFUSION TIRÉS DU RECENSEMENT DU CANADA DE 1881

Au début des années 1980, les Archives publiques du Canada ont obtenu les copies du *questionnaire 1: Renseignements d'ordre général* du recensement du Canada de 1881. Les questionnaires ont été microfilmés, et on peut maintenant trouver des exemples de ces microfilms dans la plupart des bibliothèques de collège ou d'université et dans beaucoup de bibliothèques publiques. Après avoir produit les microfilms, les Archives publiques du

Canada ont voulu intégrer toutes les données du recensement dans une base ordinolingue ou créer des bandes-échantillons à grande diffusion semblables à celles qui avaient été produites pour les recensements de 1971 et de 1976 (voir Statistique Canada (1975, 1979) pour la documentation). Le Centre de données sur les sciences sociales de l'Université Western Ontario s'est donc vu accorder un contrat pour réaliser une étude de faisabilité, et on a demandé à l'auteur d'élaborer une méthode d'échantillonnage qui permettrait de former l'échantillon à grande diffusion. La présente section contient une description du plan de sondage proposé. On trouvera un rapport de l'étude de faisabilité dans Mitchell *et coll.* (1982).

Le questionnaire 1 contient des renseignements sur l'âge, le sexe, le pays de naissance, l'origine ethnique, la profession et l'état matrimonial de chaque personne et indique si cette personne souffre d'incapacité. Les sept autres questionnaires contiennent des renseignements sur l'industrie, l'agriculture, les forêts, la pêche et les mines. On trouvera une brève description des questionnaires dans *Recensement du Canada 1880-1881*, volume 1, p. v-xv.

Nous décrivons brièvement ci-dessous les conditions de base des échantillons à grande diffusion. Pour que ces échantillons soient conformes à ceux de 1971 et de 1976, il faudrait avoir deux échantillons indépendants, soit un échantillon de ménages et un échantillon de personnes. Si toutefois il n'était économiquement possible de produire qu'un échantillon, ce devrait être en priorité un échantillon de ménages. L'échantillon à grande diffusion tiré du recensement des États-Unis de 1900 et décrit par Johnson (1978b) et Graham (1980) est un échantillon de ménages. En outre, il semble que les historiens préfèrent avant tout le ménage comme unité d'échantillonnage. L'échantillon du recensement de 1900 indique également qu'une taille d'échantillon de l'ordre de 100,000 unités serait souhaitable pour l'échantillon de personnes ou l'échantillon de ménages. En ce qui concerne le recensement du Canada de 1881, cela impliquerait une fraction de sondage d'environ 2% pour l'un ou l'autre échantillon. Enfin, il est également souhaitable d'utiliser, pour l'un et l'autre échantillon, un échantillonnage stratifié avec répartition proportionnelle où les strates correspondent à des régions géographiques. Cette méthode est celle qui a été le plus couramment utilisée jusqu'à maintenant par les historiens, et elle produit un échantillon autopondéré. Le choix des unités dans une strate devrait se faire par échantillonnage aléatoire simple plutôt que par échantillonnage systématique. Johnson (1978a) soutient que l'échantillonnage systématique, bien que pratique, ne convient pas aux questionnaires manuscrits de recensement. Des voisins ont des caractéristiques similaires et ne pourraient jamais appartenir à un même échantillon systématique. Or les historiens pourraient vouloir étudier les personnes qui ont des caractéristiques similaires.

Compte tenu de ces conditions de base, nous proposons pour le prélèvement de l'échantillon de ménages un échantillonnage aléatoire stratifié où les strates correspondent, comme dans Ornstein (1978), aux divisions de recensement (secteurs de dénombrement actuels) plutôt qu'aux bobines de microfilm utilisées par Johnson (1978b) et Graham (1978). Les divisions de recensement sont des strates géographiques naturelles. En outre, les ménages sont numérotés successivement sur les listes des agents recenseurs, chaque page manuscrite comportant vingt-cinq noms. Il suffirait alors de faire un premier visionnement des microfilms pour connaître le nombre de ménages dans chaque strate. Avec une fraction de sondage de 2 à 2.5% et une répartition proportionnelle, on obtient des échantillons de moins de deux ménages dans des divisions (strates) comptant un peu moins de cent ménages. Dans ce cas, la division en question devrait être intégrée à des divisions contigües. Une stratification poussée au-delà de la division, comme celle d'Ornstein (1978), paraît inutile et ferait monter sensiblement le coût de l'échantillonnage.

L'échantillonnage peut être facilement informatisé. Pour un codeur assis à un terminal et utilisant une visionneuse de microfilms, l'échantillonnage est une opération simple. Lorsqu'un codeur échantillonne une division, il n'a qu'à taper le code de la division voulue et

le numéro du premier ménage à échantillonner apparaît à l'écran. Le codeur fait alors avancer le film jusqu'au numéro de ménage voulu. Une fois les données de ce ménage saisies, il appuie sur une touche appropriée et le deuxième numéro de ménage apparaît à l'écran. Lorsqu'il a fini d'échantillonner cette division, il appuie sur la même touche pour passer à une autre division. Il peut parfois y avoir des ménages manquants. Il se peut, par exemple, qu'une ou plus d'une feuille d'agent recenseur contenant vingt-cinq noms ait été perdue. Quand un codeur constate l'absence d'un ménage en échantillonnant une division, il introduit en mémoire le numéro correspondant en indiquant qu'il s'agit d'un ménage manquant et fait de même avec tous les autres numéros sous lesquels il ne voit aucun ménage d'inscrit. Il poursuit ensuite son échantillonnage jusqu'à ce qu'il ait couvert toute la division. Comme il manque au moins un ménage dans l'échantillon prévu, le codeur doit rembobiner le microfilm et poursuivre l'échantillonnage dans la même division. Le principal avantage de cette méthode est de permettre au codeur de faire défiler les microfilms dans une seule direction, sauf lorsqu'il y a des ménages manquants.

L'algorithme utilisé dans cette méthode d'échantillonnage est fondé sur un fichier contenant des données sur les divisions ou groupes de divisions de même que sur l'algorithme de Bebbington (1975) conçu pour le prélèvement d'un échantillon aléatoire simple sans remise. Après le premier visionnement des microfilms, on crée un fichier indiquant l'identificatif de la division et le nombre de ménages que contient la division. Si les divisions sont groupées, leur taille respective est enregistrée. Il suffit pour le codeur de taper l'identificatif de la division à échantillonner pour connaître la taille de cette division. Pour obtenir une répartition proportionnelle de l'échantillon de ménages, il faut que la taille de l'échantillon soit égale au produit de la taille de la division par la fraction de sondage s'appliquant à l'ensemble du recensement. Avec l'algorithme de Bebbington (1975), on effectue ensuite sondage progressif des unités de l'échantillon à partir d'une liste contenant dans l'ordre les numéros des ménages de la division donnée. Les numéros de ménage sont testés tour à tour, puis choisis ou rejetés. Lorsqu'un numéro de ménage est choisi, il est affiché sur l'écran et la sélection s'interrompt le temps de la saisie des données. Comme les numéros sont choisis dans l'ordre ascendant, le codeur n'a qu'à faire défiler le film dans une seule direction pour trouver les documents manuscrits voulus.

Cet algorithme permet aussi d'échantillonner des strates combinées ou divisions groupées. Supposons que L strates de taille $N_1, \ldots, N_L$ aient été combinées dans une seule strate de taille $N = N_1 + N_2 + \ldots + N_L$. Il suffit d'utiliser les tailles des strates pour connaître le ménage échantillonné dans chaque strate. Supposons,dans l'algorithme, que les unités $s(1), \ldots, s(n)$ aient été choisies pour former l'échantillon $1 \leq s(i) \leq N$. Si pour tout $i(i = 1, \ldots, n)$ $N_1 + \ldots + N_{h-1} < s(i) \leq N_1 + \ldots + N_{h-1} + N_h$, où $N_1 + \ldots + N_{h-1} = 0$ pour $h = 1$, l'unité $s(i)$ se trouve dans la strate h et le numéro de ménage correspondant est $s(i) - (N_1 + \ldots + N_{h-1})$.

On peut aussi modifier l'algorithme de sorte qu'il tienne compte des ménages manquants. La méthode décrite ci-dessous ne suppose pas l'énumération des ménages manquants avant l'échantillonnage. une fois terminé l'échantillonnage d'une strate au moyen de l'algorithme de Bebbington (1975), deux cas sont possibles: ou bien on a constaté l'absence d'un certain nombre de ménages, ou bien on n'a constaté aucune absence. Le deuxième cas ne pose aucun problème, l'échantillonnage étant complet. Dans le premier cas, la taille de l'échantillon formé, disons m, est inférieure à la taille voulue n. Pour obtenir un échantillon de ménages existants de taille n il faut échantillonner $n - m$ ménages additionnels. À cette fin, nous recommençons l'échantillonnage de la strate, mais cette fois avec une liste des ménages qui ont déjà été échantillonnés et des ménages manquants connus. Supposons que la liste contienne M ménages ($M \geq n$: le codeur peut avoir remarqué et enregistré des ménages manquants qui n'ont pas été échantillonnés). Définissons un vecteur v à N dimensions, où la valeur du

$u^{\text{ième}}$ élément est u, $v(u) = u$ pour $u = 1, \ldots, N$. Le $u^{\text{ième}}$ élément désigne le $u^{\text{ième}}$ ménage d'une division. Supprimons maintenant tous les éléments de v qui correspondent aux ménages qui ont déjà été échantillonnés ou à ceux qui ne figurent pas sur le microfilm et réduisons le vecteur v à un vecteur w à $(N - M)$ dimensions. Les valeurs $w(u), u = 1, \ldots, N - M$ correspondent aux numéros de ménages susceptibles d'être échantillonnés. Dans l'algorithme, il suffit de redéfinir la taille de la population comme $N - M$ et la taille de l'échantillon comme $n - m$.

Avec de légères modifications, la méthode d'échantillonnage des ménages peut facilement produire un échantillon distinct et indépendant de personnes. Cette flexibilité repose sur le fait que chaque page des listes de recensement est numérotée et contient vingt-cinq noms. Le premier visionnement des microfilms doit servir à connaître le numéro de la dernière page de chaque division et le nombre de lignes que contient cette page. Avec l'algorithme de Bebbington, l'ordinateur imprimera le numéro de la page et le numéro de la ligne où figure le nom de la personne échantillonnée.

Cette méthode d'échantillonnage a été programmée et testée avec succès par le Centre de données sur les sciences sociales. Selon l'étude de faisabilité, par exemple, la recherche des unités échantillonnées a représenté environ 6% de la durée totale estimée de la saisie de données pour l'échantillon de ménages et 18.5% de cette durée pour l'échantillon de personnes. Voir Mitchell *et coll.* (1982, p. 20-21).

BIBLIOGRAPHIE

BATEMAN, F., et FOUST, J.D. (1974). A sample of rural households selected from the 1860 manuscript censuses. *Agricultural History*, 48, 75-93.

BEBBINGTON, A.D. (1975). A simple method of drawing a sample without replacement. *Applied Statistics*, 24, 135.

BODE, F.A., et GINTER, D.E. (1984). A critique of land holding variables in the 1860 census and the Parker-Gallman sample. *Journal of Interdisciplinary History*, 15, 277-295.

DARROCH, A.G., et ORNSTEIN, M.D. (1980). Ethnicity and occupational structure in Canada in 1871: the vertical mosaic in historical perspective. *Canadian Historical Review*, 61, 305-333.

FOGEL, R.W., et ENGERMAN, S.L. (1974). *Time on the Cross: Evidence and Methods*. Boston: Little, Brown and Co.

FOUST, J.D. (1975). *The Yeoman Farmer and Westward Expansion of U.S. Cotton Production*. New York: Arno Press.

GRAHAM, S.N. (1980). *1900 Public Use Sample: User's Handbook*. Seattle: Centre for Studies in Demography and Ecology, University of Washington.

HAMMARBERG, M.A. (1971). Designing a sample from incomplete historical lists. *American Quarterly*, 23, 542-561.

HAMMARBERG, M.A. (1977). A sampling design for Mormon Utah, 1880, *Journal of Interdisciplinary History*, 7, 453-476.

HOLT, D., SCOTT, A.J., et EWINGS, P.D. (1980). Chi-squared tests with survey data. *Journal of the Royal Statistical Society*, série A, 143, 303-320.

JOHNSON, R.C. (1978a). A procedure for sampling manuscript census schedules. *Journal of Interdisciplinary History*, 8, 513-530.

JOHNSON, R.C. (1978b). The 1900 census sampling project: methods and procedures for sampling and data entry. *Historical Methods*, 11, 147-151.

McINNIS, R.M. (1977). Childbearing and land availability: some evidence from individual household data. *Population Patterns in the past*, (R.D. Lee ed.), New York: Academic Press, 201-227.

MITCHELL, S.P., LING, D.G., et HANIS, E.H. (1982). *Final Report: Determination of Procedures and Costs for the Production of a Machine Readable Edition of the 1881 Census of Canada*. MAS contrat n° OSU80-00326.

ORNSTEIN, M.D. (1978). The design of a sample of households from the 1871 census of Canada. Manuscrit non publié, Université York, Toronto.

ORNSTEIN, M.D., et DARROCH, G.O. (1978). National mobility studies in past time: a sample strategy. *Historical Methods*, 11, 152-161.

RAO, J.N.K., et SCOTT, A.J. (1981). The Analysis of categorical data from complex surveys: chi-square tests for goodness of fit and independence in two-way tables. *Journal of the American Statistical Association*, 76, 221-230.

SCOTT, A.J., et HOLT, D. (1982). The effect of two-stage sampling on ordinary least squares methods. *Journal of the American Statistical Association*, 17, 848-854.

SMITH, D.S. (1978). A community-based sample of the older population from the 1880 and 1900 United States manuscript census. *Historical Methods*, 11, 67-74.

SOLTOW, L. (1975). *Men and Wealth in the United States 1850-1870*. New Haven: Yale University Press.

STATISTIQUE CANADA (1975). *Recensement du Canada de 1971: Bandes-échantillon à grande diffusion, Documentation des utilisateurs*, Ottawa.

STATISTIQUE CANADA (1979). *Recensement du Canada de 1976: Bandes-échantillon à grande diffusion, Documentation des utilisateurs*, Ottawa.

SWEIRENGA, R.P. (1983). Quantitative methods in rural landholding. *Journal of Interdisciplinary History*, 13, 787-808.