LA REFONTE DE L'ÉCHANTILLON DE L'ENQUÊTE SUR LA POPULATION ACTIVE CANADIENNE

Normand Laniel[1]

1. Introduction

L'Enquête sur la population active (EPA) est un sondage mensuel auprès de ménages produisant un portrait du marché du travail canadien (e.g. chômage et emploi). Elle est la plus grande enquête-ménages menée par Statistique Canada (SC). Plusieurs enquêtes-ménages utilisent l'échantillon de l'EPA comme base de sondage, dont certaines font, en plus, la collecte de données simultanément à celle de l'EPA. Cela permet des économies quant à l'opération de ces enquêtes.

L'échantillon de l'EPA utilise un plan stratifié à plusieurs degrés. Pour les grands centres urbains, on utilise une base aréolaire et, pour les 18 plus grands centres au Canada, on utilise en plus une base-liste de grands immeubles à appartements (les individus se logeant dans ces appartements ont généralement des caractéristiques différentes). On utilise deux degrés d'échantillonnage dans les grands centres. Pour les petits centres urbains et les régions rurales, on se sert d'une base aréolaire et d'un échantillonnage à deux ou trois degrés.

Un sixième des ménages de l'échantillon de l'EPA est renouvelé à chaque mois. Les ménages sélectionnés restent six mois consécutifs dans l'échantillon avant d'être remplacés. Ce plan de renouvellement permet de produire des estimations efficaces des changements mensuels et maintient le fardeau des répondants à un niveau acceptable.

Pour l'EPA, la collecte des données se fait par interview en personne lors du premier mois dans l'échantillon et par téléphone pour les cinq mois subséquents. (Pour obtenir les détails de la méthodologie de l'EPA pour les années 1984 à 1990, voir Singh et coll., 1990).

On remanie l'EPA après chaque recensement décennal de la population. Dans les années 70, on a révisé le questionnaire afin de pouvoir produire de l'information plus

1 Division des méthodes d'enquêtes-entreprises, Statistique Canada, R.H. Coats, 11-O, Ottawa, K1A 0T6, CANADA

détaillée sur le marché du travail, accru l'utilisation de l'ordinateur pour rendre plus efficace le traitement des données et repensé le plan d'échantillonnage dans le but de produire des estimations provinciales plus fiables. La taille de l'échantillon est alors passée de 36 400 à 55 700 ménages. Après le recensement de 1981, on a essentiellement refait le plan d'échantillonnage afin d'être en mesure de produire des estimations infra-provinciales de qualité. Suite au gain d'efficacité obtenu, l'échantillon fut réduit à 51 700 ménages.

Lors de la refonte des années 90 (voir Drew et coll., 1991), on a entrepris une révision du questionnaire dans le but d'améliorer la mesure des composantes de la population active ainsi que d'augmenter l'information recueillie. On a aussi visé à faire usage des nouvelles technologies informatiques pour la collecte (maintenant assistée par ordinateur portatif), le traitement des données ainsi que la cartographie, ceci afin d'améliorer la qualité des données, réduire le temps de traitement, améliorer l'accessibilité des données pour fin d'analyse ainsi que réduire le temps pour préparer l'échantillon. De plus, on a repensé le plan d'échantillonnage pour améliorer son efficacité à un coût raisonnable et faire de l'EPA une meilleure base de sondage pour la réalisation d'enquêtes-ménages. Les changements apportés au plan sont le propos de la présente communication. Les études planifiées pour remanier d'autres éléments de l'enquête sont décrites dans Singh et coll. (1993).

Lors de la refonte du plan d'échantillonnage on a dû tenir compte de nouveaux éléments influençant grandement les changements apportés. Entre autres, avant 1989 les régions infra-provinciales pour lesquelles on devait produire des estimations étaient les régions économiques (RÉ), les régions métropolitaines du recensement (RMR) et certains centres urbains. Depuis 1989, on doit aussi produire des estimations pour les régions d'assurance-chômage (RAC) pour fin d'administration du programme d'assurance-chômage. Pour cela, le Ministère du Développement des ressources humaines (DHRC) finance un échantillon supplémentaire de 16 500 ménages. Comme on le verra, le besoin en données pour les RAC affecte l'approche utilisée pour répartir l'échantillon. Un autre facteur influençant la méthode de répartition est le résultat de deux réductions de la taille de l'échantillon de l'EPA afin de satisfaire à des contraintes budgétaires pour Statistique Canada. Une première réduction a eu lieu en 1986 et à cette occasion la taille est passée à 46 500 ménages. La deuxième fut introduite en 1993 et l'échantillon s'est retrouvé à 42 300 ménages pour la part financée par Statistique Canada. Dans le futur, il est plausible que d'autres réductions soient imposées à l'EPA. Par exemples, le financement de Statistique Canada pourrait être davantage diminué ou l'échantillon supplémentaire pour les RAC disparaître. Il faut tenir compte de ces éventualités lors de la répartition de l'échantillon.

La réduction du financement de l'EPA au cours des années 80 et au début des années 90 a aussi eu un impact sur le maintien de l'efficacité du plan d'échantillonnage dans les grands centres urbains. En effet, dans les grands centres, les tailles initiales des unités primaires d'échantillonnage, basées sur les données du recensement de 1991 et

servant à leur sélection avec probabilité proportionnelle à la taille (PPT), deviennent désuètes après un certain temps. Ceci est dû à la croissance inégale de la population à l'intérieur des strates. Comme la méthode de sélection des unités primaires, utilisée dans les grands centres, est celle avec probabilité proportionnelle à la taille sans remplacement par groupes aléatoires de Rao, Hartley et Cochran (1962), il est possible avec la méthode de Keyfitz (1951) de sélectionner un nouvel échantillon d'unités primaires, avec les tailles mises à jour, tout en retenant une majorité des unités de l'échantillon original. Cette approche permet de maintenir l'efficacité du plan tout en minimisant les coûts de listage d'unités nouvellement sélectionnées. Elle a été utilisée dans les années 70 et la proportion de rétention obtenue fut de 70 %. Malheureusement, dans les années 80, le financement de l'EPA n'a pas été suffisant pour permettre l'application de la méthode de Keyfitz avec, comme conséquence, une détérioration de l'efficacité du plan d'échantillonnage. Comme il n'y a pas de signes montrant que le financement pourrait être meilleur pour les années 90, il est donc important de revoir le plan pour les grands centres.

Pour les régions rurales et les petits centres urbains, l'ancien plan de l'EPA était à trois degrés avec des unités primaires géographiquement grandes et compactes dont on tirait, au troisième degré, un échantillon de logements de taille égale à la tâche d'un interviewer, soit environ 60 logements. Avant le remaniement des années 80, alors que les six interviews successives étaient en personne dans les régions rurales, l'utilisation de telles unités primaires permettait de minimiser les coûts. Après le remaniement des années 80, les interviews du deuxième au sixième mois dans ces régions sont devenues téléphoniques. Au début des années 90, on a observé que la correspondance un à un entre la tâche d'un interviewer et l'unité primaire compacte était moins fréquente que la non-correspondance. Cette observation a suggéré l'étude d'un plan, autre qu'un plan à trois degrés avec unité primaire compacte, qui pourrait donner un gain d'efficacité à coût raisonnable en plus d'être plus utile aux enquêtes de plus en plus nombreuses à utiliser l'échantillon de l'EPA comme base de sondage.

Dans les sections qui suivent, on abordera les questions et problèmes mentionnés dans cette introduction ainsi que les solutions suggérées de même que leur évaluation. Dans la section 2, la question de la répartition de l'échantillon sera discutée en tenant compte des contraintes énumérées ci-haut. Une alternative à l'ancien plan d'échantillonnage pour les grands centres urbains sera présentée à la troisième section ainsi qu'une comparaison empirique effectuée. À la section 4, on discutera d'une façon d'améliorer le plan pour les régions rurales et les petits centres urbains qui sera par la suite comparée à l'ancien plan à l'aide de simulations. Finalement, un sommaire et quelques plans futurs seront présentés à la cinquième et dernière section.

2. Répartition de l'échantillon

Dans un premier temps, on va présenter l'approche utilisée pour répartir l'échantillon lors du remaniement des années 80 puis, celle utilisée lors du remaniement des années 90 tout en faisant des comparaisons et en expliquant comment on y est arrivé. On discutera aussi des objectifs de qualité ainsi que des méthodes de répartition proprement dites. Pour des détails supplémentaires, on peut consulter Mian et Laniel (1994).

2.1 Remaniement des années 80

À la suite de consultations menées auprès des provinces et des principaux ministères fédéraux utilisant les données de l'EPA, il fut décidé d'améliorer les estimations infra-provinciales sans réduire la qualité de celles effectuées au niveau des provinces et on a alors dressé la liste d'objectifs de fiabilité suivante :

(i) maintenir la fiabilité des estimations mensuelles du chômage pour le Canada et les provinces à son degré actuel ;

(ii) produire des estimations mensuelles du chômage avec un coefficient de variation ne dépassant pas 20 % pour les 24 RMR ; et

(iii) produire des estimations mensuelles du chômage avec un coefficient de variation ne dépassant pas 25 % pour les 66 RÉ.

La taille de l'échantillon à répartir lors de ce remaniement était de 55 700 ménages. La répartition de cet échantillon entre les provinces, déterminée pendant les années 70, fut inchangée, donc, seulement la question des répartitions infra-provinciales avait besoin d'une réponse. Pour ce faire, on a choisi comme strates primaires pour chaque province le croisement des RÉ, des RMR et du type de région. On a défini deux types de régions désignés : (i) urbain, pour les grands centres urbains, et (ii) rural, pour les régions rurales et les petits centres. La première étape de la répartition infra-provinciale a consisté à répartir l'échantillon provincial de façon optimale (i.e. minimisation des coûts pour une variance déterminée) entre les deux types de régions en utilisant le modèle de Fellegi, Gray et Platek (1967) (voir aussi Choudhry et coll., 1985 et Singh et coll., 1990). À l'intérieur de chaque type de régions, l'échantillon fut ensuite réparti proportionnellement à la taille de la population. Dans un deuxième temps, on s'est assuré que les régions infra-provinciales rencontraient les objectifs de fiabilité en ajustant à la hausse les tailles obtenues à la première étape si, pour les RMR et RÉ, elles étaient inférieures à 300 ménages et, pour les autres centres urbains, à 120 ménages. Finalement, suite à des améliorations apportées au plan d'échantillonnage, on a décidé de réduire la taille de l'échantillon de sorte que la fiabilité des estimations provinciales soit celles d'avant le remaniement. On a ainsi

réduit les échantillons provinciaux de 5 à 9 % en réduisant essentiellement ceux des grandes RMR et RÉ. L'échantillon remanié avait alors une taille totale de 51 700 ménages.

2.2 Remaniement des années 90

Comme on l'a mentionné à l'introduction, l'échantillon de l'EPA a été deux fois réduit depuis le remaniement des années 80, avec comme résultat qu'au moment de le remanier après le recensement de 1991, il était de 42 300 ménages pour la partie financée par Statistique Canada (échantillon de base) et de 16 500 ménages pour la partie financée par Développement des Ressources Humaines Canada (échantillon supplémentaire). Il est important de souligner ici que le financement de DRHC peut disparaître si le besoin d'estimations pour les RAC n'est plus, suite, par exemple, à un changement dans la façon de gérer le programme d'assurance-chômage (ce programme est présentement en révision). Lors de la spécification des objectifs, on a tenu compte de cette situation et le résultat fut le suivant :

(i) maintenir la fiabilité des estimations mensuelles du chômage pour le Canada et les provinces au moins à son degré actuel avec l'échantillon de base ;

(ii) produire des estimations basées sur une moyenne de trois mois du chômage avec un coefficient de variation ne dépassant pas 25 % pour les 72 RÉ avec l'échantillon de base.

(iii) produire des estimations basées sur une moyenne de trois mois du chômage avec un coefficient de variation ne dépassant pas 15 % pour les 61 RAC avec l'échantillon de base et le supplémentaire.

On remarquera que les deux premiers objectifs font appel seulement à l'échantillon de base afin d'éviter des problèmes advenant le retrait de l'échantillon supplémentaire. On notera aussi que la fiabilité des RÉ est spécifiée pour des moyennes de trois mois au lieu d'estimations mensuelles comme au remaniement précédent. Ceci est dû au fait que l'échantillon de base n'est plus que de 42 300 ménages et qu'avec une telle taille il est plus difficile de produire des estimations infra-provinciales fiables. D'ailleurs, il a fallu grouper certaines des RÉ pour rencontrer l'objectif (ii). Enfin, aucun objectif n'a été explicitement fixé pour les RMR car elles sont aussi des RAC. Si l'échantillon supplémentaire venait à disparaître, il faudrait alors en spécifier en termes de l'échantillon de base.

L'échantillon a été réparti en deux temps : d'abord l'échantillon de base puis le supplémentaire. La première étape pour répartir l'échantillon de base fut d'étudier quelques schémas de répartition entre les provinces. Parmi les schémas étudiés, on retrouve : la répartition d'avant le remaniement, la répartition de Neyman, la répartition proportionnelle à la population, la répartition proportionnelle à la

racine carrée de la population et la répartition à coefficients de variation (CV) égaux ainsi que des mélanges de ces répartitions (e.g., un compromis entre les répartitions de Neyman et à CV égaux, voir Bankier, 1988). Comme les populations provinciales sont très différentes, les répartitions proportionnelle à la population et de Neyman donnent des tailles inacceptables pour les petites provinces. Dans le cas de la répartition à CV égaux, c'est la situation contraire, i.e. les petites provinces obtiennent des tailles d'échantillon trop grandes et le CV national augmente de façon importante. Les répartitions de compromis, quant à elles, donnent des tailles trop grandes aux trois plus grandes provinces. À la fin, on a décidé de garder la répartition d'avant le remaniement excepté pour quatre provinces. Pour les paires Manitoba - Saskatchewan et Alberta - Colombie-Britannique, l'échantillon de base a été ajusté pour que les provinces d'une même paire aient le même CV (voir le tableau 1).

En ce qui concerne la répartition de l'échantillon de base à l'intérieur des provinces, on a considéré quelques schémas dont la répartition optimale (utilisée au remaniement précédent), la répartition proportionnelle à la population ainsi que la répartition proportionnelle à la racine carrée de la population. On a observé que la répartition proportionnelle à la population procurait une plus grande fiabilité au niveau provincial ainsi que national, alors que la répartition proportionnelle à la racine carrée donnait de meilleurs résultats pour les régions infra-provinciales. Comme l'échantillon de base a été grandement réduit au cours de la dernière décennie et qu'il peut encore l'être, on a préféré la répartition proportionnelle à la population afin d'assurer une fiabilité acceptable d'abord pour les estimations provinciales. Cependant, afin de rencontrer les objectifs de fiabilité pour les RÉ, que l'on s'était fixé avec un échantillon de 42 300 ménages, on a modifié la répartition obtenue pour assurer une taille d'au moins 200 ménages (300 pour l'Alberta) par RÉ. À l'intérieur des RÉ, la répartition est proportionnelle à la population.

Une fois la répartition de l'échantillon de base complétée, l'échantillon supplémentaire de 16 500 ménages, financé par DRHC, fut réparti sans égard aux frontières provinciales ; c'est-à-dire qu'on a attribué à chaque RAC une taille supplémentaire suffisante pour obtenir une fiabilité uniforme pour les RAC. Cela nous a permis d'atteindre une fiabilité correspondant à un CV de 10 % (calcul basé sur des moyennes de taux de chômage des années 1984 à 1992), ce qui est en deçà de l'objectif initial. De plus, on s'est assuré que chaque RAC obtenait au moins 600 ménages au total (i.e. base + supplémentaire) afin de garantir la fiabilité au cas où les taux de chômage au cours des prochaines années deviendraient plus bas que ceux utilisés pour la répartition.

Comme un examen du tableau 1 nous le montre, le fait d'avoir une répartition de base infra-provinciale essentiellement proportionnelle à la population procurera des estimations provinciales et nationale plus fiables qu'avec la répartition d'avant le remaniement. On observe, par exemple, que la réduction de variance au niveau national est de 19 % avec l'échantillon de base.

Tableau 1: Tailles des échantillons de base et total (i.e. base+supplémentaire) avec coefficients de variation correspondants selon la province pour la répartition d'avant le remaniement et celle après.

	Avant remaniement				Après remaniement			
Province	Base		Total		Base		Total	
	n	CV	n	CV	n	CV	n	CV
Terre-Neuve	2 240	5,4	2 582	5,1	2 240	5,2	2 240	5,2
Île-du-Prince-Édouard	1 421	7,5	1 421	7,5	1 421	7,5	1 421	7,5
Nouvelle-Écosse	3 101	5,4	4 002	5,0	3 101	5,1	4 050	4,7
Nouveau-Brunswick	3 095	5,2	3 441	5,0	3 095	5,2	3 480	5,0
Québec	6 474	4,1	11 356	3,5	6 474	3,7	11 630	3,2
Ontario	8 517	4,1	17 388	3,3	8 517	3,7	17 246	3,0
Manitoba	3 276	6,5	3 897	6,3	3 870	5,2	4 428	5,0
Saskatchewan	4 527	5,1	4 563	5,0	3 933	5,2	3 987	5,1
Alberta	5 205	4,5	5 225	4,5	4 745	4,3	4 745	4,3
Colombie-Britannique	4 454	5,1	4 975	4,6	4 914	4,3	5 623	4,1
Canada	42 310	2,0	58 850	1,7	42 310	1,8	58 850	1,5

3. Plan urbain

Comme souligné dans l'introduction, il y a peu de chances, dans le contexte budgétaire actuel, que l'efficacité du plan urbain (i.e. des grands centres urbains) puisse être maintenue avec la méthode de Keyfitz. Cette méthode permet de sélectionner un nouvel échantillon aréolaire, en utilisant des tailles de grappe mises à jour, tout en retenant une majorité des unités de l'échantillon original. Dans les paragraphes qui suivent, on présente des alternatives au plan aréolaire des années 80, puis on fait une comparaison de leur variance d'échantillonnage ainsi que de leur coût d'opération. D'autres détails sur cette étude sont donnés dans Laniel et Mohl (1994).

3.1 Plans étudiés

Le plan urbain conçu lors du remaniement suivant le recensement de 1981, a, en général, deux niveaux de stratification ainsi que deux degrés d'échantillonnage. Le premier niveau de stratification consiste en des strates (primaires) compactes et formées d'unités géographiques contiguës. Ces strates primaires sont ensuite sous-stratifiées en strates non-compactes et formées d'unités non-contiguës. Dans la mesure où les tâches des interviewers sont géographiques et ces strates secondaires non-géographiques, on obtient une interpénétration des tâches et donc une représentation de la variance de réponse corrélée dans les estimations de la variance. Les strates secondaires font le deuxième niveau de stratification. Aux deux niveaux de stratification, une version modifiée (Drew et coll., 1985) de l'algorithme multivarié de stratification optimale de Friedman et Rubin (1967) est utilisée. Seize variables socio-économiques sont impliquées dans la formation de strates avec cet algorithme. L'unité primaire d'échantillonnage (UPÉ) du plan des années 80 est habituellement un îlot urbain de taille moyenne 50 logements mais pouvant varier de 20 à 100 logements. Comme mentionné dans l'introduction, ces UPÉ (aussi appelées grappes) sont sélectionnées avec la méthode des groupes aléatoires de Rao-Hartley-Cochran (1962). Au deuxième degré, on fait un sous-échantillonnage des grappes sélectionnées en tirant un échantillon systématique d'environ cinq à six logements par grappe. La taille du sous-échantillon de grappes est déterminée de sorte que le plan est auto-pondéré mais peut cependant varier avec le temps à mesure que de nouveaux logements ou des démolitions sont identifiés par les mises à jour semestrielles des listes de grappes sélectionnées.

Un inconvénient avec le plan ci-haut est que la croissance (ou décroissance) de la taille des grappes d'une même strate n'est pas uniforme, ceci étant dû à la non-uniformité de la croissance (ou décroissance) de la population dans cette strate. Comme les grappes sont sélectionnées en utilisant les tailles du dernier recensement décennal, elles deviennent désuètes et l'efficacité du plan diminue avec le temps. Comme souligné précédemment, dans ce cas une solution consiste à utiliser la méthode de Keyfitz, mais cela risque d'être impossible lors de la prochaine décennie. Une idée proposée dans le cours du remaniement des années 1990 est d'utiliser des grappes de taille plus grande. Le principe sous-jacent à cette proposition est qu'une grappe de grande taille aura un changement de taille relatif moins grand qu'une grappe de petite taille pour un même changement absolu dans sa taille. Par exemple, une augmentation de 50 logements constitue une croissance de 100 % pour une grappe dont la taille initiale est de 50 logements mais une croissance de seulement 25 % pour une grappe dont la taille initiale est de 200 logements. Comme c'est la taille relative qui sert à la sélection PPT, des grappes de grande taille devraient se traduire par un plan plus efficace pour les estimations mensuelles. Un autre avantage d'utiliser des grappes de plus grande taille est l'amélioration des estimations du changement de mois à mois. D'une part, les grappes étant plus grandes, elles seront renouvelées moins rapidement, assurant ainsi une plus grande longitudinalité et,

d'autre part, quand elles le seront par une grappe où il y a eu une croissance importante, l'impact sera moindre.

Pour étudier l'idée de l'utilisation de grappes de grande taille, on a choisi de comparer des grappes de taille moyenne : 50 logements avec d'autres de taille 100, 150, 200 et 250 logements. On s'est arrêté à 250 logements car l'utilisation de tailles plus grandes risquerait, selon notre expérience, de trop faire augmenter les coûts de listage. Notons que 250 logements est aussi la taille moyenne des secteurs de dénombrement (SD) urbains, utilisés pour le recensement de la population.

Lorsqu'on augmente la taille des grappes tout en gardant la taille totale de l'échantillon et la taille moyenne du sous-échantillon de grappe fixes, alors le volume de logements à lister augmente et les coûts de listage aussi. Ceci étant indésirable, pour la comparaison de plan il était donc important de faire varier la taille moyenne du sous-échantillon à l'intérieur des grappes. On a choisi de considérer les tailles moyennes de 5 à 16 logements, car, d'une part, plus petit que cinq ne présentait pas d'intérêt et, d'autre part, plus grand que seize risquait de donner un plan moins efficace.

3.2 Méthodologie de l'étude

Pour l'étude, on a choisi la région urbaine d'Ottawa, qui est une RMR de la province de l'Ontario, car les paramètres de son plan d'échantillonnage établi à partir du recensement de 1981, sont proches de ceux que l'on retrouve en général dans les grandes régions urbaines. Les tailles moyennes de grappe et de sous-échantillon de grappe sont respectivement environ 50 logements et 5 logements. La taille de l'échantillon pour la région d'Ottawa, allouée lors du remaniement des années 80, était de 390 logements. C'est ce que nous avons utilisé pour cette étude.

Pour incorporer l'effet de détérioration des tailles initiales des grappes, on a besoin pour toute la population à l'étude des tailles des grappes à une date ultérieure au recensement de 1981. En appariant les grappes de 1981 avec les données des recensements de 1986 ou 1991, il était possible d'obtenir de telles données. On a choisi le recensement de 1991 car cela représente la situation à la demi-vie du plan basé sur 1981 qui fut introduit en 1985. Pour calculer la variance de variables comme l'emploi ou le chômage, on a aussi besoin de ces données au niveau de toute la population. Malheureusement, les valeurs de ces variables ne sont disponibles du recensement que pour un cinquième de la population, qui est l'échantillon auquel on demande de remplir le long questionnaire en plus du court. L'approche utilisée pour le calcul de la variance est présentée à la section 3.2.1.

Les coûts qui peuvent être affectés par un changement du plan d'échantillonnage sont ceux de dénombrement des personnes, de listage initial des grappes et de mise à

jour des listes de grappes. On a utilisé des modèles pour représenter ces coûts et prédire ceux des plans alternatifs. Ces modèles ainsi que l'estimation de leurs paramètres sont discutés à la section 3.2.2.

3.2.1 Calcul de la variance

L'EPA utilise un estimateur par régression, séparé pour chaque province, avec comme variables auxiliaires des projections démographiques par groupe âge-sexe au niveau provincial et des projections de la population de 15 ans et plus par région infra-provinciale (e.g., RMR et RÉ). Pour l'étude on a choisi de négliger l'effet des projections démographiques provinciales sur la région d'Ottawa, de sorte que l'estimateur de l'EPA se ramène dans ce cas à l'estimateur par quotient combiné :

$$\hat{Y}_c = \hat{R}_c X = \frac{\hat{Y}}{\hat{X}} X.$$

Pour calculer la variance de cet estimateur de façon précise, on a besoin des valeurs des variables à l'étude pour tous les individus de la population, cependant ces données ne sont disponibles que pour un échantillon. Alors, pour faire le calcul, on a utilisé une approximation de la variance de l'estimateur non-linéaire puis estimé l'expression avec les données échantillonnales du recensement de 1991. L'approximation de la variance utilisée est celle par linéarisation de l'estimateur par quotient combiné ; elle est donnée par :

$$V(\hat{Y}_c) = V(\hat{Y} - R_c \hat{X}) = V(\hat{U}) = \sum_h V(\hat{U}_h).$$

Dans ce qui suit, on suppose que l'échantillonnage aléatoire simple sans remplacement est une bonne approximation de l'échantillonnage systématique de logements au deuxième degré. Cette hypothèse est plausible pour les variables emploi et chômage, comme le confirment des études sur l'estimation de la variance pour l'EPA (Choudhry et Lee, 1987).

Choudhry et coll. (1985) donnent l'expression de la variance de l'estimateur par quotient combiné (linéarisé) pour le plan d'échantillonnage urbain de l'EPA décrit à la section 3.1. On a estimé cette expression en utilisant les données de l'échantillon 20 % du recensement de 1991 et en supposant que l'erreur d'échantillonnage sur l'estimation de R_C est négligeable. En utilisant un tilde au-dessus des quantités estimées à l'aide de l'échantillon du recensement, les formules pour les composantes au premier et deuxième degré s'écrivent respectivement :

$$V_1(\hat{U}_h) = \frac{\left(\sum_{g=1}^{n_h} N_{hg}^2 - N_h\right)}{N_h(N_h - 1)} \left(\sum_{i=1}^{N_h} \frac{u_{hi}^2 - V(u_{hi})}{z_{hi} / Z_h} - \left(U_h^2 - V(U_h)\right) \right)$$

et

$$V_2(\hat{U}_h) = \sum_{i=1}^{N_h} \left(W_h - 1 - \frac{\left(\sum_{g=1}^{n_h} N_{hg}^2 - N_h\right)}{N_h(N_h - 1)} \left(\frac{Z_h}{z_{hi}} - 1 \right) \right) M_{hi} S_{hi}^2$$

où N_h est le nombre de grappes dans la strate h ; N_{hg} est le nombre de grappes dans le groupe aléatoire g de la strate h ; Z_h est la somme des mesures de taille pour les grappes dans la strate h ; z_{hi} est la mesure de taille de la grappe i de la strate h (i.e., dans la présente étude le nombre de logements selon le recensement de 1981); W_h est l'inverse de la fraction de sondage combinée des deux degrés pour la strate h; $U_h=Y_h-R_cX_h$ est le total de la strate pour la variable u (selon le recensement de 1991) ; $u_{hi}=y_{hi}-R_cx_{hi}$ est le total de la grappe i de la strate h pour u ; M_{hi} est le nombre de logements dans la grappe i de la strate h au moment de tirer le sous-échantillon (e.g., un certain temps après le recensement de 1981) ; $u_{hij}=y_{hij}-R_cx_{hij}$ est le total du logement j de la grappe i de la strate h pour u; $\bar{u}_{hi}$ la moyenne par logement pour u dans la grappe i de la strate h ; S_{hi}^2 est la variance estimée, avec l'échantillon 20 %, de la variable u dans la grappe i de la strate h ; $V(u_{hi})$ est la variance estimée du total estimé u_{hi} ; et $V(U_h)$ est la variance estimée du total estimé U_h. Ces deux dernières variances ont pour but de corriger la surestimation obtenue en prenant le carré de l'estimation d'un total comme estimation du total au carré.

Afin de calculer la variance avec la formule ci-haut et les données du recensement, on a dans un premier temps formé des grappes de plus grande taille que celles disponibles à partir du plan de 1981. Pour obtenir ces grappes, on a simplement groupé les grappes de 50 logements qui étaient contiguës de façon à ce que le résultat soit de grandes grappes compactes. Lors de la formation des grandes grappes, un écart de 50 % par rapport à la taille moyenne était accepté.

La taille du sous-échantillon de grappe a été modifiée en changeant le nombre de grappes sélectionnées et donc le nombre de strates formées, car on s'est contraint à

sélectionner six grappes par strate comme c'est habituellement le cas avec le plan de l'EPA. Trois tailles moyennes de sous-échantillon ont été utilisées soit 4,7, 8,2 et 16,4 logements avec comme nombre de strates 14, 8 et 4 respectivement.

3.2.2 Calcul des coûts

Comme mentionné plus haut, les coûts de dénombrement, de listage initial et de mise à jour des listes peuvent être affectés par l'utilisation de grappes plus grandes. Pour ce qui est du listage initial, il n'est pas évident si de plus grandes grappes augmenteraient ou diminueraient ce coût. À première vue, ce coût doit augmenter puisque les grappes plus grandes contiennent plus de logements. Cependant, on renouvellera moins souvent de grandes grappes car leur durée de vie dans l'échantillon sera plus longue. En effet, la durée de vie d'une grappe est fonction du nombre de sous-échantillons qu'elle contient. Par exemple, une grappe de 150 logements de laquelle on tire des sous-échantillons de 10 logements, qui passent chacun 6 mois dans l'échantillon avant d'être renouvelés, aura une durée de vie maximale de sept ans et demi. Ceci illustre bien que la durée de vie des grappes est aussi un facteur important en ce qui concerne le coût de listage initial. Pour cette raison, les modèles de coût ci-après incluent la durée de vie du plan de l'EPA qui est de dix ans (i.e. la période entre deux remaniements).

Choudhry et coll. (1985) ont étudié l'effet de la variation de la taille des sous-échantillons de grappe sur les coûts de dénombrement. Ces coûts comprenaient le temps d'interview et les déplacements (i.e. les déplacements de la maison à l'aire de travail et le retour, de grappe-à-grappe et de logement-à-logement). Dans leur étude, Choudhry et coll. ont fait varier la taille du sous-échantillon de 2 à 10 logements et diminuer le nombre de grappes sélectionnées, quand la taille du sous-échantillon augmentait, de sorte que la taille totale de l'échantillon demeurait constante. Ils ont observé que le coût total de dénombrement demeurait essentiellement constant lorsque la taille du sous-échantillon variait. Aussi, dans la présente étude, on considère le coût total de dénombrement comme simplement proportionnel au nombre total de logements sélectionnés. Le modèle adopté est donc :

$$C_E = c_E m F L_D$$

où C_E est le coût total de dénombrement, c_E est le coût de dénombrement par logement sélectionné ; m est le nombre total de logements sélectionnés dans l'échantillon; F est la fréquence de l'enquête (i.e., 12 mois par an pour l'EPA) ; et L_D est la durée de vie du plan de l'enquête en années (i.e., 10 ans pour l'EPA). Le seul paramètre qui doit être estimé est c_E et il a été estimé à 6,20 $ lors d'une étude de temps et coût (voir Mantel et coll., 1994).

Le listage initial de grappe se fait dans deux situations : lorsqu'une grappe est sélectionnée dans l'échantillon initial ou lorsqu'une grappe en remplace une autre par renouvellement. Le nombre de fois qu'un listage initial est effectué dépend de la durée de vie des grappes. Pour une grappe dans l'échantillon initial, la durée de vie, en termes d'années, est un nombre aléatoire compris entre un minimum prédéterminé et le nombre de sous-échantillons (systématiques) qu'elle contient divisé par deux. Cette façon de déterminer la durée de vie d'une grappe initiale permet de préserver les probabilités de sélection des grappes lors de leur renouvellement. Pour ce qui est d'une grappe qui en remplace une autre, sa durée de vie est généralement égale au nombre de sous-échantillons qu'elle contient divisé par deux. On peut décomposer le coût de listage initial en deux parties : la première est indépendante de la taille de la grappe et comprend le déplacement de la maison à la grappe ainsi que le retour; la deuxième dépend de la taille de la grappe et inclut les déplacements de logement-à-logement en plus de l'enregistrement des adresses. Le modèle utilisé pour représenter le coût de listage initial pour la durée du plan est :

$$C_L = (c_{0L} + c_{1L} M) n R_L$$

avec $R_L = 1 + (L_D - 1) / L_{clu,max}$ et $L_{clu,max} = (M / \overline{m}) L_s$

où C_L est le coût total de listage initial, c_{0L} est le coût indépendant moyen par grappe; c_{1L} est le coût dépendant moyen par logement; M est la taille (moyenne) d'une grappe; n est le nombre de grappes sélectionnées dans l'échantillon; R_L est le nombre espéré de fois que des grappes ont à être listées durant la vie du plan ; $L_{clu,max}$ est la vie maximale moyenne d'une grappe en années ; $\overline{m}$ est la taille moyenne d'un sous-échantillon de grappe ; L_s est la durée de vie d'un sous-échantillon en années (e.g., ½ pour l'EPA). On notera que le modèle utilisé ne tient pas compte des probabilités de sélection des grappes mais utilise plutôt une taille moyenne de grappes. Bien que cela ne soit pas théoriquement correct, nous croyons que l'approximation est suffisante pour les besoins de cette étude. On doit ici estimer deux paramètres soient c_{0L} et c_{1L}. Pour ce faire, on s'est basé sur des données de coûts de listage que le bureau central obtient régulièrement. L'inconvénient avec ces données est que les composantes, indépendante et dépendante, de la taille des grappes ne sont pas séparées. Cependant, après discussion avec les responsables de la collecte de données des hypothèses raisonnables ont pu être faites quant à la désagrégation des composantes. On a donc estimé c_{0L} à 15,13 $ par grappe et c_{1L} à 0,32 $ par logement.

La mise à jour des listes de grappes se fait généralement lors de la semaine de collecte, en même temps que les visites en personne pour les ménages qui en sont à leur premier mois dans l'échantillon. Par conséquent, ce coût n'inclut pas de

déplacement et n'a donc pas de composante indépendante de la taille des grappes comme le coût de listage initial. Le modèle que l'on a utilisé est :

$C_U = c_U M n R_U$ avec $R_U = L_D/L_U$

où C_u est le coût total de mise à jour des grappes, c_U est le coût moyen de mise à jour par logement; R_U est le nombre de fois que des grappes auront leur liste mise à jour pendant la vie du plan; et L_U est la périodicité de mise à jour en années (e.g. ½ pour l'EPA). Notez que le modèle reflète le fait que les grappes nouvellement dans l'échantillon sont mises à jour dès leur premier mois car le listage initial est effectué quelques mois avant l'entrée dans l'échantillon. Dans ce modèle, il y a un paramètre à estimer c_U. Malheureusement, il n'y a pas de données disponibles sur le coût unitaire de mise à jour. Après consultation avec les responsables des opérations de l'enquête, on a établi que ce coût est environ le tiers du coût dépendant unitaire de listage initial, donc on a pris c_U= 0,11 $ par logement pour la présente analyse.

3.3 Résultats

En se servant des méthodes de calcul pour la variance et les coûts présentés ci-haut, on a obtenu les résultats qui suivent pour des grappes de taille moyenne 50, 100, 150, 200 et 250 logements et une taille de sous-échantillon variant de 5 à 16 logements.

3.3.1 Variance

La variance a été calculée pour deux variables : l'emploi et le chômage. Les résultats se trouvent représentés graphiquement aux figures 1 et 2. La figure 1 nous montre l'efficacité relative au plan de la région d'Ottawa (i.e., grappe moyenne de 50 logements et sous-échantillon moyen de 4,7 logements) pour la variable emploi en fonction de la taille du sous-échantillon de grappe. La figure 2 montre les mêmes résultats pour la variable chômage.

Comme on peut l'observer dans les deux figures, pour une taille de sous-échantillon donnée, la variance diminue lorsque la taille de la grappe augmente ; cela s'explique par un accroissement de la fraction de sondage au premier degré. Aussi, pour une taille de grappe donnée, la variance augmente lorsque la taille du sous-échantillon augmente, reflétant ainsi l'effet de réduire le nombre de grappes dans l'échantillon. On remarquera que les courbes pour le chômage ont des pentes plus faibles que celles pour l'emploi, ce qui traduit un gain d'efficacité plus faible. Par exemple, pour une taille de grappe disons de 220 logements et une taille de sous-échantillon de 8 logements, le gain pour l'emploi est de 9 %, à la demi-vie du plan, contre seulement 1 % pour le chômage. Ce résultat peut s'expliquer par le fait que la contribution du

premier degré d'échantillonnage à la variance est plus faible pour le chômage que pour l'emploi, variable plus corrélée avec le nombre de logements dans la grappe.

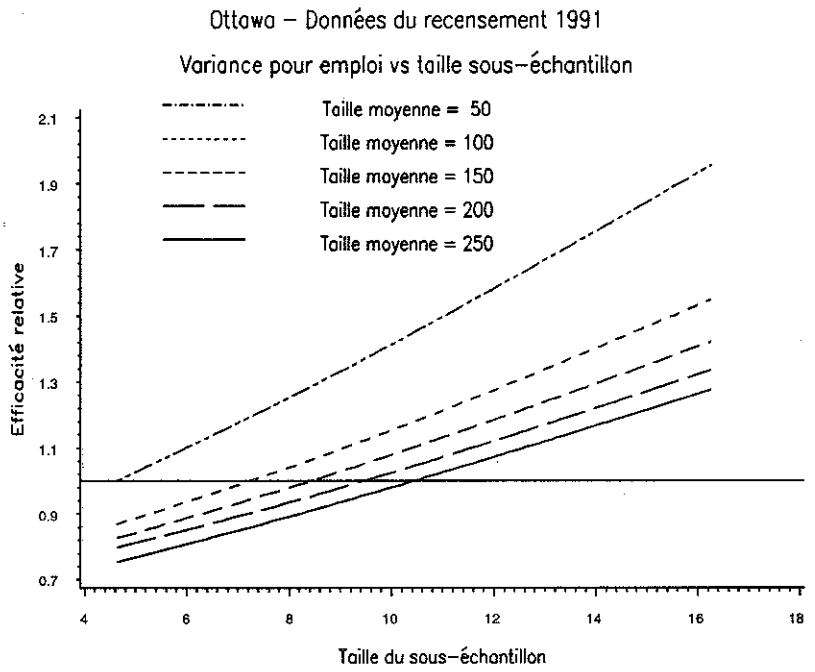

Figure 1

3.3.2 Coût

Pour produire les résultats pour le coût total d'opération sur le terrain, on a sommé les trois composantes : coûts de dénombrement, de listage initial et de mise à jour. Le total, relatif au plan de la région d'Ottawa, est représenté à la figure 3 pour les cinq tailles de grappes en fonction de la taille de sous-échantillon. On observe que, pour une taille de sous-échantillon donnée, le coût total augmente lorsque la taille de la grappe augmente. Aussi le coût diminue, pour une taille de grappes donnée, lorsque la taille du sous-échantillon augmente. En reprenant l'exemple d'une taille moyenne de grappes de 220 logements avec une taille de sous-échantillon de 8 logements, utilisé dans l'analyse des résultats de variance, on obtient que l'augmentation du coût de l'enquête sur le terrain pourrait augmenter de 5 %. Une analyse par composante révèle que cette augmentation est essentiellement due à une augmentation du coût de

mise à jour des listes de grappe. Ce résultat était prévisible puisque l'utilisation de plus grandes grappes implique nécessairement des aires géographiques plus grandes à mettre à jour.

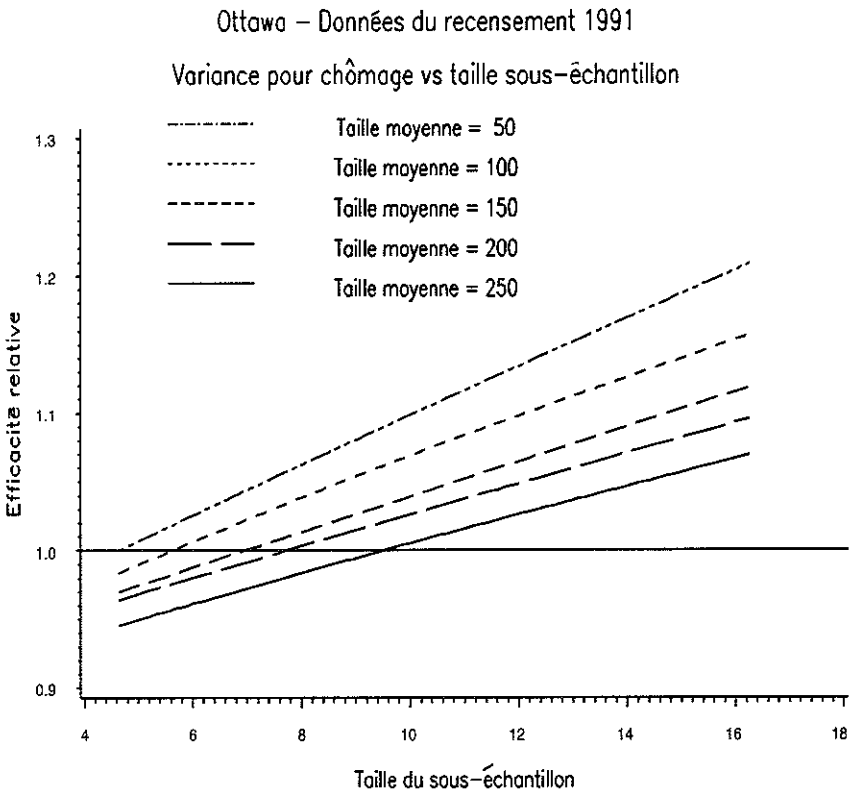

Figure 2

3.4 Conclusion

Basé sur les résultats ci-haut, il a été décidé d'augmenter la taille des grappes pour le nouveau plan urbain de l'EPA afin d'en améliorer la robustesse vis-à-vis de la croissance inégale de la population et donc l'efficacité par rapport aux estimations mensuelles et celles du changement de mois à mois. Les tailles des grappes sont maintenant d'environ 220 logements plutôt que 50. Cependant, on a aussi augmenté la taille des sous-échantillons afin de limiter l'effet d'un accroissement potentiel du coût de mise à jour des listes de grappes. La taille des sous-échantillons de grappe est d'environ 8 logements pour le nouveau plan. Le coût d'opération n'a pas été un facteur important dans la décision vu le grand gain d'efficacité apporté par la nouvelle répartition, 19 % pour le chômage, qui peut à tout le moins en partie se monnayer. Cependant, comme noté ci-haut, le coût de mise à jour va possiblement

augmenter. Aussi, comme plusieurs approximations ont été faites dans l'analyse de coût, on planifie d'entreprendre une étude plus exacte de l'impact du nouveau plan. Si effectivement, ce coût s'accroît substantiellement, on pourrait, par exemple, envisager de réduire la fréquence de mise à jour dans les régions où la croissance de la population est connue pour être faible.

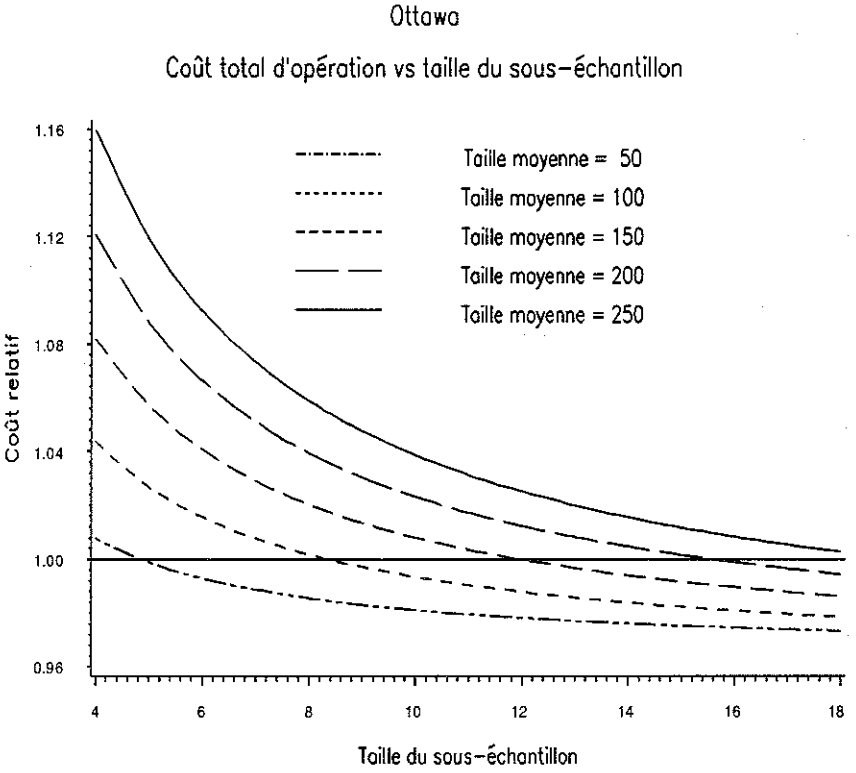

Figure 3

4. Plan rural

Comme mentionné au début de ce document, lors du remaniement des années 80 on a adopté, pour les régions rurales et les petits centres urbains, un plan à trois degrés avec des unités primaires d'échantillonnage compactes et suffisamment grandes pour qu'on puisse en tirer des échantillons d'environ 60 logements, ce qui correspond à la taille d'une tâche d'interviewer. Le but de former de telles UPÉ était de minimiser les coûts de déplacement. Cependant, ces dernières années, on a observé que la plupart du temps il n'y avait pas de correspondance un-à-un entre une tâche et une UPÉ. Pour cette raison, on a choisi de comparer le plan à trois degrés de 1981 à un plan à deux degrés. Les bénéfices espérés sont d'améliorer l'efficacité du plan rural, de le rendre plus utile aux enquêtes utilisant l'échantillon de l'EPA comme base de

sondage et de faciliter l'estimation pour les petites régions. Cependant, il est possible qu'un plan à deux degrés augmente les coûts de déplacement reliés au dénombrement. Dans ce qui suit, on décrit les plans d'échantillonnage comparés, la méthodologie pour calculer variance et coût, et on présente les résultats et la conclusion de l'étude.

4.1 Plans étudiés

Suite au remaniement post-censitaire de 1981, on a choisi un plan d'échantillonnage stratifié à trois degrés pour les régions rurales et les petits centres urbains. Pour ce plan rural, on distingue trois niveaux de stratification : (i) les régions économiques de l'EPA (i.e., les RÉ), (ii) à l'intérieur de chaque RÉ, une partition en trois groupes de régions (uniquement rurales, uniquement urbaines et mixtes, (i.e., un mélange de régions rurales et urbaines)) et (iii) finalement, une sous-stratification des groupes. Les parties rurale et mixte sont sous-stratifiées en strates compactes et composées d'éléments géographiquement contigus, à l'aide de la version modifiée de l'algorithme de stratification optimale de Friedman et Rubin, alors que la partie urbaine est sous-stratifiée géographiquement. Ces strates sont suffisamment grandes pour que l'on puisse tirer, comme échantillon ultime de logements, l'équivalent de deux à trois tâches d'interviewers, soit de 120 à 180 logements. Une fois les strates formées, on construit les UPÉ rurales et mixtes de façon à maximiser, d'un point de vue socio-économique, la compacité et la similarité des UPÉ d'une même strate. Les UPÉ urbaines sont des centres urbains ou parfois une combinaison de centres urbains voisins. La sélection des UPÉ se fait avec la méthode d'échantillonnage systématique avec probabilité proportionnelle à la taille et randomisation (SPPTR) et deux à trois UPÉ sont tirées par strate. Au deuxième degré, les unités d'échantillonnage sont, dans les strates rurales, les secteurs de dénombrement du recensement (SD). Dans les strates urbaines, ce sont de petites grappes, souvent des îlots urbains. Dans les strates mixtes, on stratifie d'abord chaque UPÉ en parties rurale et urbaine puis, on utilise respectivement les SD et les îlots comme unités secondaires d'échantillonnage (USÉ). Toutes les USÉ sont sélectionnées avec la méthode SPPTR. Au troisième degré, on sélectionne des logements avec la méthode systématique, et on en tire environ dix par SD et trois par grappe urbaine.

L'alternative, au plan à trois degrés ci-dessus, est un plan stratifié à deux degrés. Ce plan se distingue d'une part par la stratification, qui se veut plus utile pour ceux utilisant l'EPA comme base de sondage, et par l'élimination des grandes UPÉ équivalentes à une tâche d'interviewer. Comme dans l'ancien plan, il y a, en général, trois niveaux de stratification. Le premier niveau est constitué des intersections des RÉ avec les RAC, dans le but de faciliter la production d'estimations fiables pour ces deux ensembles de régions. Au deuxième niveau, les intersections sont divisées en strates en utilisant la division du recensement (DR) (i.e., une aire géographique ayant une population d'environ 80 000 habitants) comme unité de stratification. Si une DR a une taille suffisamment grande, elle sert alors de strate au deuxième niveau sinon, on la regroupe

avec d'autres DR. Au troisième niveau, on utilise la stratification optimale pour sous-stratifier les DR ou groupes de DR. Le respect des limites géographiques des DR dans la stratification rend cette dernière plus utile pour les autres enquêtes car certaines d'entre elles utilisent les DR pour définir leurs populations à l'étude. Un autre point à noter au sujet de la stratification est que les strates au dernier niveau sont formées de telle sorte qu'on tire de chacune environ 60 logements. Elles sont donc de deux à trois fois plus petites que celles de l'ancien plan, ce qui veut dire que l'échantillon est géographiquement réparti plus uniformément. Cela rend plus aisé la production d'estimations pour de petites régions. À l'intérieur de chaque strate de troisième niveau, l'objectif est de sélectionner six SD qui sont ensuite sous-échantillonnés pour donner dix logements chacun. La méthode de sélection est SPPTR au premier degré et celle au deuxième est l'échantillonnage systématique. On s'attend à une réduction de la variance par l'élimination d'un degré d'échantillonnage. Pour plus de détails sur l'étude on peut lire Mantel et coll. (1994) et Marion et coll. (1995).

4.2 Méthodologie de l'étude

Pour comparer la variance et les coûts de dénombrement des deux plans décrits ci-dessus, quatorze des 66 RÉ (telles que définies au moment du recensement de 1981) provenant des dix provinces canadiennes ont été choisies. Les tailles d'échantillon utilisées sont celles d'après la réduction de 1986. Comme on le verra dans le calcul de la variance, on s'est servi d'une approche similaire à celle de l'étude sur les plans urbains. Cependant, ce sont les données du recensement de 1986 qui furent utilisées et non celles de 1991, ceci à cause de la difficulté d'apparier les SD ruraux de 1991 avec ceux de 1981. En effet, il y a moins de changement dans les limites géographiques des SD à cinq ans d'intervalle qu'à dix. De plus, les changements dans les SD ruraux sont plus difficilement identifiables que ceux des SD dans les grands centres urbains pour lesquels on a de l'information à un niveau plus fin que le SD, soit les côtés d'îlot. Ainsi, la comparaison de variance reflétera la situation au début de la vie du plan d'échantillonnage.

Dans l'analyse des coûts pour le plan rural, on a tenu compte seulement des coûts de dénombrement car, lorsqu'on a spécifié les fractions de sondage du plan alternatif, on s'est assuré que les coûts de listage initial et de mise à jour seraient les mêmes pour les deux plans. Des modèles ont été formulés pour représenter les composantes du coût total de dénombrement et les paramètres ont été estimés aux moyens des résultats de l'étude de temps et coût de Mantel et coll. (1994) et d'une simulation de type Monte-Carlo. Le tout est discuté à la section 4.2.2.

4.2.1 Calcul de la variance

Tout comme dans le cas de la région d'Ottawa, l'estimateur de l'EPA se ramène pour les RÉ à l'estimateur par quotient combiné (voir section 3.2.1). De même, pour calculer la variance de cet estimateur, on a utilisé une approximation de la variance

de cet estimateur non-linéaire et enfin estimé l'expression obtenue avec les données de l'échantillon du recensement de 1986.

Pour le plan d'échantillonnage à trois degrés, les trois composantes qui ont été sommées par strate sont (le tilde dénote une estimation basée sur l'échantillon 20 % du recensement de 1986):

$$\tilde{V}_1(\hat{U}_h) = \sum_{i=1}^{N_h} \sum_{i'>i}^{N_h} (p_{hi} p_{hi'} - p_{hi,i'}) \left(\left(\frac{u_{hi}}{p_{hi}} - \frac{u_{hi'}}{p_{hi'}} \right)^2 - \frac{\tilde{V}(u_{hi})}{p_{hi}^2} - \frac{\tilde{V}(u_{hi'})}{p_{hi'}^2} \right)$$

$$\tilde{V}_2(\hat{U}_h) = \sum_{i=1}^{N_h} \frac{1}{p_{hi}} \sum_{j=1}^{M_{hi}} \sum_{j'>j}^{M_{hi}} (p_{hij} p_{hij'} - p_{hijj'}) \left(\left(\frac{u_{hij}}{p_{hij}} - \frac{u_{hij'}}{p_{hij'}} \right)^2 - \frac{\tilde{V}(u_{hij})}{p_{hij}^2} - \frac{\tilde{V}(u_{hij'})}{p_{hij'}^2} \right)$$

$$\tilde{V}_3(\hat{U}_h) = \sum_{i=1}^{N_h} \frac{1}{p_{hi}} \sum_{j=1}^{M_{hi}} \frac{1}{p_{hij}} \frac{L_{hij}}{\ell_{hij}} (L_{hij} - \ell_{hij}) \tilde{S}_{hij}^2$$

où N_h est le nombre d'UPÉ dans la strate h ; π_{hi} et $\pi_{hi'}$ sont les probabilités de sélection des UPÉ i et i' de la strate h; $\pi_{hi,i'}$ est la probabilité de sélection conjointe pour les UPÉ i et i' de la strate h; u_{hi} et $u_{hi'}$ sont les valeurs de la variable transformée pour les UPÉ i et i' de la strate h; M_{hi} est le nombre d'USÉ dans l'UPÉ i de la strate h; π_{hij} et $\pi_{hij'}$ sont les probabilités de sélection des USÉ j et j' de l'UPÉ i de la strate h; $\pi_{hij,j'}$ est la probabilité conjointe de sélection pour les USÉ j et j' de l'UPÉ i de la strate h; u_{hij} et $u_{hij'}$ sont les valeurs de la variable transformée pour les USÉ j et j' de l'UPÉ i de la strate h; L_{hij} est le nombre de logements dans l'USÉ j, selon le recensement de 1986, de l'UPÉ i de la strate h; ℓ_{hij} est le nombre de logements sélectionnés de l'USÉ j de l'UPÉ i de la strate h; et, $\tilde{S}_{hij}^2$ est la variance estimée de la population dans l'USÉ j, au recensement de 1986, de l'UPÉ i de la strate h.

Pour ce qui est de la formule pour calculer la variance du plan à deux degrés, la première composante est la même que ci-haut et la deuxième est donnée par

$$\tilde{V}_2(\hat{U}_h) = \sum_{i=1}^{N_h} \frac{1}{p_{hi}} \frac{L_{hi}}{\ell_{hi}} (L_{hi} - \ell_{hi}) \tilde{S}_{hi}^2 .$$

Notons ici que, encore une fois, les variances estimées des totaux d'unités d'échantillonnage estimés, à l'aide de l'échantillon 20 % du recensement de 1986, servent à corriger la surestimation obtenue en prenant le carré de l'estimation d'un total comme estimation du total au carré.

Dans les formules ci-haut, le calcul des probabilités conjointes de sélection de façon exacte, avec l'algorithme de Hidiroglou et Gray (1980), demande beaucoup de temps d'ordinateur. Pour réduire le temps de calcul, on a utilisé l'approximation de Hartley et Rao (1962) lorsque la taille de la population d'unités d'échantillonnage était supérieure à 15.

4.2.2 Calcul du coût

Pour la comparaison des coûts de collecte, on a considéré deux composantes : les coûts fixes et ceux dépendant du plan. Les coûts fixes comprennent le temps de préparation de la tâche, le temps pour faire les interviews ainsi que les déplacements de logement-à-logement dans les SD échantillonnés (au deuxième degré dans le plan à trois degrés et au premier degré dans le plan à deux). Les coûts dépendant du plan incluent les déplacements entre la maison de l'interviewer et son aire de travail et les déplacements entre SD échantillonnés.

Pour arriver aux modèles de coûts, on a supposé que les nombres moyens de déplacements et les vitesses moyennes de ceux-ci, pour chacun des deux types de déplacement, étaient les mêmes pour les deux plans. On a aussi supposé que les deux plans pouvaient différer en ce qui concerne les distances moyennes des déplacements. Ces différences entre les deux plans seront vraisemblablement causées par les différences entre les étendues géographiques moyennes des tâches d'interviewer. Cependant, on a fait l'hypothèse que les deux types de déplacement seraient affectés de façons différentes, ceci en se basant sur les résultats de l'étude de temps et coût. Cette étude a montré que les distances et vitesses moyennes de la maison à l'aire de travail sont plus grandes que les distances et vitesses moyennes d'un SD à un autre pour le plan à trois degrés (en fait les déplacements de la maison à l'aire de travail constituent les deux tiers du total des frais de déplacement), suggérant ainsi que la maison d'un interviewer est souvent située à l'extérieur du périmètre décrit par sa tâche. Avec des tâches géographiquement plus étendues, pour le plan à deux degrés, cette situation pourrait changer à une où la maison serait plus fréquemment située à l'intérieur du périmètre de la tâche.

Basé sur les considérations ci-dessus, le modèle suivant fut suggéré pour le plan à trois degrés

$C3 = C_F + C_D + C_H$,

et pour le plan à deux degrés

$C2 = C_F + f_D C_D + f_H C_H$,

où C_F est le total des coûts fixes; C_D est le total des coûts de déplacement de SD-à-SD; C_H est le total des coûts de déplacement de la maison à l'aire de travail; f_D est le ratio des distances moyennes de SD-à-SD du plan à deux degrés sur le plan à trois; et f_H est le ratio des distances moyennes de la maison à l'aire de travail du plan à deux degrés sur le plan à trois.

Les coûts totaux ont été estimés à l'aide des données de l'étude de temps et coût exécutée pour le plan à trois degrés. Dans cette étude spéciale, on a demandé à tous les interviewers de l'EPA, pendant deux mois successifs, d'inscrire sur le questionnaire de l'information sur leurs déplacements. Les interviewers devaient, pour chaque visite en personne, prendre note du jour, des heures de départ et d'arrivée, et la lecture de l'odomètre de leur véhicule.

Les ratios f_D et f_H ont, quant à eux, été estimés à partir d'une simulation Monte-Carlo. Pour cette simulation, les étapes suivantes furent répétées plusieurs fois pour chacun des deux plans. D'abord, un échantillon de SD est sélectionné (pour le plan à trois degrés, on sélectionne en premier des UPÉ). Puis, on assigne aléatoirement un numéro de rotation, de un à six, à chaque SD sélectionné. Pour le plan à trois degrés, chaque UPÉ a généralement un sous-échantillon de six SD avec chacun un numéro de rotation différent. Pour le plan à deux degrés, on retrouve la même situation mais au niveau de la strate cette fois. Ensuite, avec les SD sélectionnés, on forme des tâches d'interviewers géographiquement compactes, équilibrées en termes de numéros de rotation et grosso modo de même taille. Dans la formation des tâches, on a tenu compte des échantillons provenant des grands centres urbains de sorte que certaines tâches sont mixtes. Finalement, on détermine un emplacement pour la maison de l'interviewer assigné à une tâche. Si la tâche est près d'un centre urbain important, la maison est localisée au centroïde de la partie urbaine de la tâche. Si la tâche est purement rurale, la maison est localisée aléatoirement avec une loi normale bivariée circulaire. La loi est standardisée de sorte que la distance espérée du centroïde de la tâche à la maison a pour valeur la distance moyenne de la maison à l'aire de travail observée dans l'étude spéciale de temps et coût. Pour améliorer l'efficacité de la simulation, cette dernière étape est répétée dix fois pour chaque échantillon simulé. Les estimations de f_D et f_H sont obtenues en calculant les ratios des moyennes Monte-Carlo des distances appropriées.

4.3 Résultats

Les résultats du calcul de variance et de coût sont présentés au tableau 2. Pour chacune des 14 régions économiques dans l'étude, on y trouve le code de la région, la variance du plan à deux degrés relativement à celle du plan à trois degrés pour les variables emploi et chômage, ainsi que le coût du plan à deux degrés relativement à

celui du plan à trois degrés. Notons que lorsque la variance relative est inférieure à l'unité cela signifie que le plan à deux degrés est plus efficace que le plan à trois degrés. De même lorsque le coût relatif est inférieur à un cela veut dire que le plan à deux degrés est moins coûteux que celui à trois degrés.

4.3.1 Variance

Tableau 2 Variance relative pour emploi et chômage et coût relatif selon la RÉ.

RÉ	$V_{rel,emploi}$	$V_{rel,chômage}$	C_{rel}
020	1,35	0,86	1,03
220	0,86	0,68	1,11
310	0,81	0,81	1,05
320	0,77	0,66	1,05
330	1,07	1,02	1,09
340	0,77	0,84	1,05
350	1,02	0,93	1,12
411	0,75	0,83	1,21
510	0,94	0,96	1,09
560	1,11	1,06	1,08
630	1,03	0,95	1,10
720	1,09	0,97	1,11
820	0,98	0,93	1,06
960	0,69	0,95	1,13
Moy.	0,94	0,90	1,09

Pour la variable emploi, on observe que, dans 7 cas sur 14, le plan à deux degrés a un gain d'efficacité supérieur à 5 % sur celui à trois. Dans 3 cas, les plans sont d'efficacité similaire (i.e. la différence est inférieure à 0,05). Et, dans 4 cas, le plan à trois degrés est plus efficace d'au moins 5 %. La moyenne de la variance relative des 14 cas est de 0,94.

Pour la variable chômage, dans 10 cas, le plan à deux degrés est plus efficace par au moins 5 %. Dans 3 cas, les plans ont, à toutes fins pratiques, une efficacité semblable (i.e. différent par moins de 5%). Et, dans 1 cas, le plan à trois degrés est plus efficace par au moins 5 %. Aussi, la moyenne de la variance relative est de 0,90.

Le fait d'obtenir un gain d'efficacité plus grand pour le chômage s'explique par la méthodologie employée pour la formation des grandes UPÉ à trois degrés. Cette méthodologie multivariée cherche à optimiser l'efficacité du plan avec, entre autres, huit variables reliées à l'emploi mais aucune reliée au chômage, ainsi les grandes UPÉ sont, en général, moins efficaces pour estimer le chômage.

4.3.2 Coût

En ce qui concerne le coût relatif, dans quatre cas l'augmentation est inférieure à 5 % et dans les onze autres cas elle est de plus de 5 %. La valeur moyenne du coût relatif est 1,09, ce qui veut dire une augmentation de 9 %. Les deux composantes, déplacement de SD-à-SD et entre maison et aire de travail, contribuent environ dans la même proportion à cette augmentation.

4.4 Conclusion

Comme on vient de le constater, le plan à deux degrés est en général plus efficace que le plan à trois degrés. Le plan à deux degrés est aussi plus simple, possède une stratification basée sur les divisions du recensement (donc plus utile aux enquêtes utilisant l'EPA comme base de sondage) et donne lieu à un échantillon géographiquement plus uniformément réparti (plus approprié pour l'estimation des petits domaines). Pour ces raisons, c'est le plan à deux degrés, avec six SD comme UPÉ, qui a été adopté dans la plupart des régions rurales. Cependant, pour limiter autant que possible les augmentations de coût, on a utilisé un plan différent dans les régions où la densité de population est plutôt faible. Ce plan, économique, est aussi à deux degrés mais les UPÉ sont des groupes de six SD, tous inclus dans l'échantillon. La sélection des logements se fait de façon indépendante dans chacun des SD. Évidemment, ce plan n'a pas les avantages de la stratification et de l'estimation pour les petites régions du plan alternatif étudié. Par l'utilisation de ce plan on espère réduire, au moins en partie, l'augmentation de coût. Cependant, notons encore une fois que toute augmentation n'est pas critique avec le grand gain d'efficacité obtenu avec la nouvelle répartition.

5. Sommaire et plans futurs

En résumé, l'efficacité du plan d'échantillonnage de l'EPA a été très améliorée par l'usage d'une répartition presque proportionnelle à la population pour l'échantillon de base, ce qui va donner lieu à des estimations provinciales et nationales plus

fiables. Aussi, cela rendra le plan plus propre à subir les contrecoups de nouvelles réductions de la taille de l'échantillon. Avec l'utilisation de grappes plus grandes dans le plan urbain, on a amélioré la robustesse du plan face à la croissance inégale de la population et, donc, amélioré l'efficacité des estimations mensuelles ainsi que celles du changement de mois à mois. Quant au plan rural, il est plus efficace avec l'usage d'un plan à deux degrés, plus utile avec des strates basées sur les divisions du recensement et plus apte à la production d'estimations pour les petites régions avec ses strates plus petites. Cependant, avec le nouveau plan d'échantillonnage, il faudra surveiller de près les coûts afin de s'assurer qu'ils ne dépassent pas les budgets alloués à l'EPA. Si l'on observe une augmentation significative, on pourra, par exemple, réduire la taille de l'échantillon de base afin de compenser cette augmentation. Cela pourra être fait tout en maintenant un gain d'efficacité global et, ce, grâce à la nouvelle répartition.

En termes de projets futurs reliés au plan d'échantillonnage de l'EPA, mentionnons tout d'abord un projet visant à optimiser l'utilisation des ressources pour la mise à jour des listes de grappes. Comme l'ont montré les résultats de l'étude sur les coûts de mise à jour pour le plan urbain, il est possible qu'avec l'usage de grappes de taille plus grande ces coûts augmentent. Il faudra donc, à l'intérieur du projet, étudier des stratégies de mise à jour qui permettraient d'économiser. Par exemple, on pourra considérer de mettre à jour moins souvent les listes dans les régions où il est peu probable qu'il y ait une croissance. Aussi, il faudra regarder du côté des sources d'information sur la croissance, telles que les nouveaux permis de construction ou les fichiers téléphoniques. Avec une source indicatrice de régions en croissance, on pourrait alors limiter les mises à jour à ces secteurs.

Un autre projet, qui sera prochainement initié, est la possibilité d'utiliser le Registre des adresses de Statistique Canada comme base de sondage pour l'EPA. Ce registre a tout d'abord été construit pour aider à l'énumération des logements lors des recensements de la population. Il couvre toutes les régions urbaines de 50 000 habitants ou plus. Un grand avantage d'utiliser le Registre des adresses est l'efficacité du plan à un degré qu'il est alors possible de concevoir. L'inconvénient majeur est que, présentement, le registre n'est mis à jour que pour les recensements, donc seulement tous les cinq ans. Ce n'est évidemment pas suffisant pour une enquête mensuelle comme l'EPA. Il faut donc étudier des approches pour s'assurer que l'échantillon couvre les nouveaux logements. Une approche possible serait d'utiliser une base aréolaire avec un échantillon relativement petit. Cela rendrait le plan moins efficace et plus coûteux car il faudrait alors encore mettre des listes de grappe à jour. Une autre approche, plus économique, consisterait à tirer un échantillon de dossiers administratifs, tels les nouveaux permis de construction ou les nouveaux codes postaux. Il faudrait cependant mesurer la qualité de ces sources de données et déterminer dans quelle mesure elles peuvent bien couvrir les nouveaux logements.

Remerciements

L'auteur tient à remercier tous ceux qui ont contribué de près ou de loin au contenu du présent document, en particulier M.P. Singh, Jack Gambino, Ijaz Mian, Chris Mohl, Jocelyne Marion et Harold Mantel. Il remercie aussi Johane Dufour ainsi que Francine Hardy pour avoir gracieusement révisé le texte.

BIBLIOGRAPHIE

BANKIER M.D. (1988), Power allocation: Determining sample sizes for subnational areas, *The American Statistician*, 42, 174-177.

CHOUDHRY G.H., LEE H., et DREW J.D. (1985), Optimisation du coût et de la variance dans le cadre de l'enquête sur la population active du Canada, *Techniques d'enquête*, 11, 37-56.

CHOUDHRY G.H. et LEE H. (1987), Estimation de la variance pour l'enquête sur la population active du Canada, *Techniques d'enquête*, 13, 157-172.

DREW J.D., BELANGER Y. et FOY P. (1985), La stratification dans l'enquête sur la population active du Canada, *Techniques d'enquête*, 11, 109-124.

DREW D., GAMBINO J., AKYEAMPONG E. et WILLIAMS B. (1991), Plans for the 1991 redesign of the Canadian Labour Force Survey, *Proceedings of the Survey Research Methods Section, American Statistical Association.*

FELLEGI I.P., GRAY G.B. et PLATEK R. (1967), The New Design of the Canadian Labour Force Survey, *Journal of the American Statistical Association*, 62, 421-453.

FRIEDMAN H.P. et RUBIN J. (1967), On some invariant criteria for grouping data, *Journal of the American Statistical Association*, 62, 1159-1178.

HARTLEY H.O. et RAO J.N.K. (1962), Sampling with unequal probabilities and without replacement, *Annals of Mathematical Statistics*, 33, 350-374.

HIDIROGLOU M.A. et GRAY G.B. (1980), Construction of joint probability of selection for systematic PPS sampling, *Journal of the Royal Statistical Society*, C29, 107-112.

KEYFITZ N. (1951), Sampling with probability proportional to size, *Journal of the American Statistical Association*, 58, 183-201

LANIEL N. et MOHL C. (1994), Analysis of Urban Cluster Size in the Canadian Labour Force Survey, *Proceedings of the Survey Research Methods Section, American Statistical Association.*

MANTEL H., LANIEL N., DUVAL M.-C. et MARION J. (1994), Cost Modelling of Alternative Sample Designs for Rural Areas in the Canadian Labour Force Survey. *Proceedings of the Survey Research Methods Section, American Statistical Association.*

MARION J., MANTEL H. et LANIEL N. (1995), Sample Design in Non-Self Representing Areas, Rapport interne de la Division des méthodes d'enquêtes-ménages, Statistique Canada.

MIAN I.U.H. et LANIEL N. (1994), Sample allocation for the Canadian Labour Force Survey, *Proceedings of the Survey Research Methods Section, American Statistical Association.*

RAO, J.N.K., HARTLEY H.O., et COCHRAN W.G.(1962), On a simple procedure of unequal probability sampling without replacement, *Journal of the Royal Statistical Society*, Series B, 24, 482-490.

SINGH M.P., DREW J.D., GAMBINO J.G. et MAYDA F. (1990), *Méthodologie de l'enquête sur la population active du Canada, 1984-1990.* Ottawa: Statistique Canada, Catalogue 71-526.

SINGH M.P., GAMBINO J. et LANIEL N. (1993), Research studies for the Labour Force Survey Redesign, *Proceedings of the Survey Research Methods Section, American Statistical Association.*