Techniques d'enquête, juin 1994
Vol. 20, n° 1, pp. 3–23
Statistique Canada

Les petites régions: problèmes et solutions

M.P. SINGH, J. GAMBINO et H.J. MANTEL[1]

RÉSUMÉ

Dans cet article, nous discutons de questions techniques relatives à la fourniture de données régionales tirées de recensements, de fichiers administratifs et d'enquêtes. Bien qu'il s'agisse de questions d'ordre général, nous les analysons en relation avec des programmes de Statistique Canada. Nous faisons ressortir la nécessité d'élaborer une approche globale pour les estimations d'enquête et nous soulignons, dans la perspective du remaniement d'une enquête- ménages, les aspects de la conception de plans de sondage qui ont une incidence sur les données régionales. Enfin, nous faisons un survol des méthodes d'estimation en faisant ressortir leurs avantages et leurs inconvénients.

MOTS CLÉS: Stratégie de plan d'échantillonnage; estimateurs de plan; estimateurs de modèle.

1. INTRODUCTION

Pendant des décennies, les fichiers administratifs et les recensements étaient les principales sources de données destinées à l'élaboration des politiques et à la planification pour les grandes et les petites régions. Ils constituent encore aujourd'hui la source de données la plus précieuse pour les petites régions dans la plupart des pays. Dans les années quarante et cinquante toutefois, comme les enquêtes par sondage prenaient de l'importance, les estimations d'enquête sont venues compléter les sources habituelles car elles constituent des données plus à jour et d'un bon rapport coût-efficacité pour un grand nombre de domaines spécialisés. Bien qu'elles soient conçues tout d'abord pour produire des estimations fiables aux niveaux d'agrégation supérieurs, comme les niveaux national et provincial, les enquêtes par sondage visent de plus en plus à répondre à la demande croissante d'estimations plus actuelles pour des domaines de types et de tailles variés. Aucun problème technique ne se pose tant que les domaines sont suffisamment grands (ex.: groupes d'âge-sexe, grandes villes et régions infra-provinciales) pour produire des estimations raisonnablement fiables. Cependant, si on a besoin de données pour de petits domaines, particulièrement des domaines dont les éléments sont répartis dans plusieurs strates de plan, des problèmes d'estimation particuliers se posent et plusieurs méthodes, proposées récemment, existent pour les résoudre.

Cet article a principalement pour but de souligner la nécessité d'envisager le problème des données régionales dans sa globalité. On devrait prendre conscience de la question des petites régions dès le début de la conception des plans de sondage pour les grandes enquêtes. Les plans d'échantillonnage devraient être conçus de manière que l'on puisse obtenir des données régionales fiables à l'aide d'estimateurs de plan ou de modèle. La solution voulant que les organismes statistiques règlent cette difficulté grandissante à l'étape de l'estimation ne devrait être envisagée qu'en dernier recours.

Dans la section 2, nous examinons les besoins en information et les trois principales sources de données socio-économiques au Canada, c'est-à-dire le recensement, les fichiers administratifs et les enquêtes. Dans la section suivante, nous traitons quelques questions techniques concernant les trois sources de données et nous mettons en lumière le problème des mesures de qualité et de leur interprétation. Ensuite, nous faisons ressortir la nécessité d'élaborer une approche globale qui recouvre les trois grandes étapes de la conception d'une enquête: planification, élaboration du plan de sondage et estimation. Nous examinons particulièrement deux aspects du plan de sondage, à savoir la formation de grappes dans un plan à plusieurs degrés et la répartition de l'échantillon. Dans la section 5, nous présentons quelques caractéristiques de plan de sondage qui ont été introduites récemment dans l'Enquête sur la population active du Canada la plus grande enquête mensuelle auprès des ménages de Statistique Canada à l'occasion de son remaniement afin qu'elle produise des données régionales de meilleure qualité. Enfin, la section 6 a pour but d'examiner les nombreuses méthodes d'estimation qui s'appliquent aux petites régions. En même temps, nous proposons de nouveaux estimateurs et nous faisons des commentaires sur les avantages et les faiblesses de divers estimateurs pour domaine. Nous conseillons la prudence dans l'utilisation d'estimateurs de modèle.

2. BESOINS EN INFORMATION ET SOURCES DE DONNÉES

En tant qu'organisme statistique national, Statistique Canada joue un rôle intégral dans la marche de la société canadienne. Tandis que le caractère confidentiel des données fournies par les répondants est préservé, l'information produite par l'organisme décrit les conditions économiques et sociales du pays et de sa population. Les divers programmes de l'organisme (statistique économique, statistique démographique, statistique sociale et statistique

[1] M.P. Singh, J. Gambino et H.J. Mantel, Statistique Canada, 16ième étage, Immeuble R.H. Coats, Parc Tunney, Ottawa (Ontario), Canada K1A 0T6.

des établissements) produisent des données fiables de tous genres aux niveaux national, provincial et infra-provincial à l'intention des administrations publiques fédérale et provinciales, des institutions privées, des universités et des médias. Comme la planification, l'administration et la surveillance des programmes sociaux et financiers tendent de plus en plus à se faire au niveau régional, on cherche à obtenir un plus grand nombre de données et des données de meilleure qualité à ce niveau. Nous examinons brièvement ci-dessous trois grandes sources de données socio-économiques et démographiques en mettant l'accent sur les statistiques régionales.

Recensement de la population: Le recensement quinquennal de la population produit des données-repères et constitue, à tous les cinq ans, la source d'information la plus précieuse ayant trait aux petites régions et à divers caractères, domaines ou groupes cibles tels que les minorités ethniques, les personnes en état d'incapacité, les jeunes et les autochtones.

Fichiers administratifs: Les fichiers administratifs sont une source de données statistiques de plus en plus importante. Les organismes statistiques puisent largement dans ces fichiers en démographie pour produire des estimations régionales (Schmidt 1952, Verma et Basavarajappa 1987). Pour certains domaines, comme la statistique de l'état civil, les fichiers administratifs sont la seule source d'information qui permet de produire des statistiques à des niveaux d'agrégation variés. Pour d'autres domaines, l'usage des fichiers administratifs dépend de leur capacité d'offrir des données à jour et des données de qualité comparativement aux recensements ou aux enquêtes. Les fichiers administratifs ne servent pas uniquement à la production de totalisations directes; ils sont utilisés dans un certain nombre de programmes comme source d'information supplémentaire dans le but d'améliorer la qualité d'estimations d'enquête. Ils servent aussi à la construction de bases de sondage pour des enquêtes, par exemple le Registre des entreprises et le Registre des adresses de logements résidentiels à Statistique Canada.

Comme le recensement de la population, les fichiers administratifs sont caractérisés par des niveaux d'agrégation géographique très détaillés, ce qui en fait une source d'information précieuse pour l'établissement de statistiques régionales. Leur accessibilité est maintenant plus grande et grâce aux progrès technologiques récents, ils deviennent une source de données ayant un meilleur rapport coût-efficacité. Sur le plan du contenu toutefois, les fichiers administratifs ne sont pas aussi complets et les concepts sont définis pour des programmes plutôt qu'à des fins statistiques. Brackstone (1987a, 1987b) donne les détails d'un programme de Statistique Canada qui a pour objet l'élaboration et l'intégration d'un système de fichiers administratifs destiné à produire des données statistiques. Le recueil rédigé par Coombs et Singh (1987) décrit l'expérience d'autres pays en ce qui concerne l'utilisation de fichiers administratifs.

Programme des enquêtes-ménages: Les enquêtes-ménages sont depuis longtemps une source majeure de données économiques et sociales à Statistique Canada. Les enquêtes de ce programme peuvent être classées en trois catégories: i) l'Enquête sur la population active, ii) les enquêtes spéciales et les enquêtes supplémentaires et iii) les enquêtes longitudinales et cycliques. Nous décrivons brièvement ci-dessous ces enquêtes en indiquant leur rapport avec la question des données régionales.

Ayant été créée comme une enquête trimestrielle en 1945, l'Enquête sur la population active du Canada (EPA) est devenue une enquête mensuelle en 1952. L'information recueillie au moyen de cette enquête s'est diversifiée considérablement au fil des années et aujourd'hui, elle donne un portrait complet et détaillé du marché du travail au Canada. Outre les estimations nationales et provinciales, l'EPA produit régulièrement des estimations pour les régions infra-provinciales. Il existe aussi une forte demande pour des estimations d'indicateurs courants du marché du travail relatives à de petites régions comme les circonscriptions électorales fédérales, les divisions de recensement et les territoires des centres d'emploi du Canada. Ces estimations servent aux administrations fédérale et provinciales pour contrôler les programmes et répartir les ressources, monétaires ou autres, entre les divers champs de compétence politique et administrative.

Pour des considérations de coût, l'appareil de l'EPA est souvent utilisé pour effectuer des enquêtes ponctuelles et des enquêtes périodiques à l'échelle nationale et provinciale; ce peut être sous la forme d'enquêtes supplémentaires ou d'enquêtes spéciales. Dans le cas des enquêtes supplémentaires, on demande aux participants à l'EPA de répondre à des questions additionnelles tandis que dans le cas des enquêtes spéciales, on demande à un autre échantillon de ménages, tiré de la base de l'EPA, de répondre aux questions. En règle générale, les enquêtes spéciales et les enquêtes supplémentaires sont parrainées par d'autres ministères et réalisées selon le principe de recouvrement des coûts. Dans le cas de ces enquêtes, la demande de statistiques régionales varie beaucoup selon l'enquête et, en règle générale, elle semble moins pressante que celle provenant de l'EPA.

Statistique Canada effectue annuellement une Enquête sociale générale (ESG) pour répondre, modestement, aux besoins grandissants en données sur des sujets d'actualité sociale. Le programme de l'ESG (Norris et Paton 1991) comporte cinq cycles d'enquête chacun ayant une thématique principale qui reviennent tous les cinq ans. À cause de la taille limitée de l'échantillon (10,000 ménages à l'échelle nationale), l'ESG met l'accent sur les estimations nationales et les données analytiques.

Les enquêtes longitudinales (ou enquêtes par panel) sont nouvelles dans le contexte canadien. Statistique Canada a mis sur pied deux enquêtes longitudinales qui viendront enrichir largement le programme des enquêtes-ménages; ce sont l'Enquête sur la dynamique du travail et du revenu et l'Enquête nationale sur la santé de la population. Il s'agit de deux grandes enquêtes par panel dont on s'attend déjà qu'elles produisent des données au niveau infra-provincial et régional.

3. QUESTIONS RELATIVES À L'ESTIMATION POUR DOMAINE

La fourniture de données régionales soulève de nombreuses questions d'ordre stratégique ou technique. L'importance de ces questions peut varier d'un organisme à l'autre et d'une application à l'autre au sein du même organisme, selon la qualité des données et les politiques de diffusion. Ces questions intéressent les estimations nationales et provinciales, mais elles supposent une plus grande importance pour les données régionales. Selon les termes mêmes de Brackstone (1987a), "(...) en ce qui concerne l'évaluation de données régionales, il convient de noter que les utilisateurs peuvent observer plus facilement une erreur dans des estimations régionales que dans des agrégats nationaux (...) il se trouvera des détracteurs prompts à souligner les lacunes des données régionales (...) il est vrai qu'en ce qui concerne les petites régions, pour lesquelles l'estimation est plus difficile, les valeurs estimées font l'objet d'un examen plus intense." Deux ouvrages en particulier Platek et coll. (1987) et Platek et Singh (1986) présentent plusieurs études de recherche et d'élaboration sur l'estimation pour petites régions. Pour avoir un aperçu des méthodes d'estimation pour petites régions utilisées actuellement dans les programmes statistiques fédéraux aux É.-U., prière de se référer au document du U.S. Statistical Policy Office (1993).

Utilisation de fichiers administratifs: Dans la plupart des cas, l'offre et la demande de données régionales sont influencées en tout premier lieu par les politiques des gouvernements fédéral et provinciaux. Du côté de l'offre, les fichiers administratifs basés sur des programmes gouvernementaux renferment une masse de renseignements qui peuvent servir à produire des données régionales. Des exemples de ces fichiers au Canada: allocations familiales, assurance-chômage, impôt sur le revenu, santé, éducation, sécurité de la vieillesse. On produit déjà de façon régulière des données régionales sur le revenu. Tout **changement** dans la politique du gouvernement et les programmes qui s'y rattachent peut avoir un effet immédiat, positif ou négatif, sur le champ d'application, la disponibilité, l'actualité ou la qualité des données tirées des fichiers administratifs correspondants. En ce qui concerne la demande, les administrations publiques ont besoin de données régionales pour la planification, l'administration et la surveillance de leurs programmes.

Problèmes conceptuels: Dans des séries de données, on confond très souvent les problèmes conceptuels ou définitionnels et les problèmes d'échantillonnage et d'estimation. Prenons par exemple le Régime d'assurance-chômage (AC) du Canada. Le Règlement sur l'assurance-chômage prévoit différentes périodes d'admissibilité selon le taux de chômage qui sévit dans une région donnée, de telle sorte que les périodes ouvrant droit aux prestations sont plus courtes dans les régions qui affichent des taux de chômage relativement élevés. Les estimations du taux de chômage régional établies à l'aide des données de l'EPA servent à déterminer l'admissibilité d'une personne aux prestations. C'est pourquoi les médias et la population en général prêtent une attention soutenue à ces estimations régionales. Mais cet intérêt a trait plus souvent à des questions **conceptuelles** qu'à des questions d'estimation; on s'interroge, par exemple, sur la place des travailleurs découragés, des mises à pied ou des licenciements et des techniques de recherche d'emploi dans le questionnaire d'enquête.

Utilisation de modèles et des mesures de qualité correspondantes: Presque toutes les grandes enquêtes ont leurs estimations pour domaines et aucun problème ne se pose tant que les estimateurs de plan, c.-à-d. des estimateurs approximativement non biaisés selon le plan, sont de qualité acceptable. On considère deux catégories d'estimateurs de plan. D'après Schaible (1992), les estimateurs **directs** sont des estimateurs qui utilisent des valeurs de la variable étudiée qui se rapportent uniquement à des unités du domaine pour la période visée (ex.: l'estimateur par régression dont la pente est estimée au moyen de l'échantillon dans le domaine seulement). Les estimateurs de ce genre peuvent aussi utiliser – et cela est fréquent – de l'information relative à une ou à plusieurs variables auxiliaires pour d'autres domaines ou d'autres périodes, et ces estimateurs sont (approximativement) non biaisés selon le plan. Il existe une autre catégorie d'estimateurs, que nous appelons estimateurs **directs modifiés** et qui peuvent utiliser de l'information relative à la variable auxiliaire et à la variable étudiée pour d'autres domaines mais qui conservent leur propriété d'être (approximativement) sans biais selon le plan (ex.: l'estimateur par régression dont la pente est estimée au moyen de tout l'échantillon). Il existe de plus en plus d'ouvrages sur les **estimateurs indirects (ou estimateurs de modèle)**, c'est-à-dire des estimateurs qui utilisent de l'information relative à la variable étudiée et aux variables auxiliaires pour d'autres domaines ou d'autres périodes mais qui n'ont pas nécessairement la propriété d'être non biaisés selon le plan.

La plupart des producteurs et des utilisateurs de données d'enquête sont familiarisés avec les estimateurs de plan et les inférences correspondantes. Ils interprètent les données sur la base d'échantillons répétés formés à l'aide d'un plan d'échantillonnage probabiliste donné et se servent des **c.v.** de plan estimés (c.v. = coefficients de variation: racine carrée de la variance de plan divisée par la valeur estimée selon le plan) comme mesure de la qualité des données. Dans les cas où les domaines sont trop petits ou bien le plan de sondage ne prévoit pas la production d'estimations régionales, les estimations de plan peuvent donner des **c.v.** élevés et les estimations de modèle peuvent alors s'imposer comme la seule solution s'il faut produire des estimations d'enquête pour domaines. La tâche qui attend les statisticiens n'est pas simple: comment estimer et comparer la précision relative des estimations d'une enquête qui produit un grand nombre d'estimations nationales et infra-nationales et de nombreuses estimations pour de grands et de petits domaines en utilisant surtout des estimateurs de plan mais aussi, dans une moindre mesure, des estimateurs de modèle, et comment présenter cela aux utilisateurs. Les **c.v.** de modèle (racine carrée de la variance théorique de la valeur estimée selon le modèle, divisée par la valeur estimée selon le modèle) peuvent traduire une situation totalement

différente et être beaucoup moins élevés que les **c.v.** de plan correspondants pour la même petite région et, dans beaucoup de cas, moins élevés que les **c.v.** de plan pour des régions beaucoup plus grandes.

Il est généralement simple d'établir l'expression de l'erreur quadratique moyenne (c.-à-d., variance théorique + carré du biais théorique) pour les estimateurs de modèle. Il n'en va pas de même lorsqu'il s'agit d'estimer cette expression avec un degré de précision raisonnable. S'il est vrai que les données (ex.: taille d'échantillon) pour de petits domaines ne se prêtent pas à la production d'estimations de plan, il est peu probable qu'elles puissent servir à produire une estimation de la variance et du biais correspondants. Comme il est relativement plus difficile d'estimer le biais, certains auteurs recherchent chez les estimateurs de modèle la propriété de convergence selon le plan, ce qui laisse supposer peut-être qu'on peut faire abstraction du biais. Or, si la taille de l'échantillon à l'intérieur d'un domaine est suffisamment grande pour qu'un estimateur de modèle soit convergent, un estimateur de plan devrait produire des estimations fiables pour ce domaine. En ce qui concerne les estimateurs de modèle, on a proposé d'utiliser l'estimation de l'erreur quadratique moyenne calculé pour l'ensemble des domaines. Comme, en règle générale, on cherche à obtenir des estimations pour chacun des domaines pris individuellement parce qu'on croit que ces domaines sont différents les uns des autres, on doit s'efforcer d'expliquer pourquoi les estimations de ces divers domaines ont toutes le même degré de fiabilité. Une autre possibilité serait de calculer des estimations indirectes (basées sur un modèle) de la variance et du biais des estimateurs de modèle pour des domaines **pris individuellement**. De fait, l'élaboration de méthodes qui permettraient d'estimer l'erreur quadratique moyenne pour des domaines pris individuellement devrait figurer parmi les priorités de recherche. Une autre préoccupation majeure des expérimentateurs est de savoir comment parer aux défaillances de modèle. Peut-être faudrait-il effectuer des recherches sur la validation de modèle dans des enquêtes complexes. En outre, dans le cas des estimateurs de modèle qui utilisent des données relatives à la variable étudiée pour d'autres périodes que la période pertinente, les estimations de la **variation** d'une période à l'autre seraient de qualité douteuse; voir Schaible (1992). Par surcroît, les estimateurs de modèle qui empruntent de l'information à d'autres domaines de la région présenteraient la même lacune si on comparait des différences entre deux domaines de la région.

Protection de la vie privée: Afin de créer des bases de données complètes desquelles on peut tirer des statistiques régionales, on doit parfois combiner des données de recensement et d'enquête et des données de fichiers administratifs. Cette opération nécessite le couplage d'enregistrements de diverses sources. Or, compte tenu de l'intérêt qu'attache la population à la protection de la vie privée, on ne devrait effectuer des couplages d'enregistrements qu'après en avoir analysé minutieusement toutes les conséquences. La Loi sur la statistique permet à Statistique Canada d'avoir accès aux fichiers administratifs d'autres ministères ou organismes pour des raisons statistiques. Même dans ce cas, comme le souligne Fellegi (1987), "(. . .) nous devrions disposer de procédures de révision rigoureuses et vérifiables afin de nous assurer que nous ne combinons ces données que lorsque le bien public qui résulte de ces nouvelles informations statistiques l'emporte manifestement sur l'intrusion dans la vie privée que leur création entraîne."

4. NÉCESSITÉ D'UNE APPROCHE GLOBALE

Bien que les grandes enquêtes visent surtout à produire des estimations nationales et provinciales, il est rare que les estimations issues de ces enquêtes concernent uniquement les populations nationale ou provinciales dans leur entier. Autrement dit, les grandes enquêtes servent immanquablement à produire des estimations pour divers domaines **recoupés** et, parfois, pour des domaines **spatiaux** (ex.: régions infra-provinciales). Dans beaucoup de cas, on ne cherche pas particulièrement à obtenir un niveau de précision voulu pour le domaine, que ce soit à l'étape de la conception du plan de sondage ou à celle de l'estimation, pourvu que le degré de fiabilité soit (jugé) raisonnable. Les problèmes surgissent lorsqu'un domaine recoupé correspond à une sous-population rare ou qu'un domaine spatial correspond à une petite région, auxquels cas il n'existe aucune estimation ou, si elles existent, les estimations sont de qualité douteuse. Dans un certain nombre de cas, ces problèmes peuvent survenir pour la simple raison qu'on n'a pas été assez attentif à ces besoins au début du processus de planification. Si on doit répondre aux besoins en données régionales à l'aide de données d'enquête, il est nécessaire d'élaborer une approche globale qui visera particulièrement à répondre à ces besoins à chacune des étapes d'une enquête: planification, sondage et estimation. En ce qui regarde le plan de sondage et l'estimation, nous classons les domaines en deux catégories:

Domaines planifiés: Du point de vue de l'échantillonnage, les domaines planifiés sont des strates ou des groupes de strates qui ont été formés en vue de la création d'échantillons. Des exemples de ces domaines au Canada sont les régions infra-provinciales, par exemple les régions économiques, les régions de l'Assurance-chômage et les régions créées aux fins de la planification des services de santé. D'autres exemples de ces domaines seraient les grands comtés, les districts ou les régions infra-provinciales de ce type.

Domaines non planifiés: Ce sont des domaines qui n'ont pas été définis à l'étape de la conception du plan de sondage et qui, par conséquent, peuvent recouper des strates. Ils peuvent avoir n'importe quelle taille et créer des problèmes d'estimation particuliers.

Planification: Comme nous l'avons mentionné plus tôt, les données d'enquêtes périodiques comme l'EPA sont beaucoup plus en demande que les données d'enquêtes ponctuelles. En ce qui concerne les enquêtes périodiques qui sont remaniées à tous les cinq ou dix ans, on peut élaborer une stratégie adéquate au moment du remaniement puisque dans ces cas précis, les organismes statistiques sont généralement beaucoup plus en mesure de prévoir les

besoins en données régionales en se fondant sur la demande antérieure. Quant aux enquêtes ponctuelles, les concepteurs devraient inclure la définition des besoins dans le processus d'établissement des objectifs de l'enquête. Par conséquent, dans l'un et l'autre cas, les concepteurs devraient fixer un degré de précision non seulement pour les estimations nationales ou provinciales, mais aussi pour les domaines étudiés.

La première étape dans la fourniture de données régionales dépendra dans quelle mesure les domaines auront été définis à l'avance, de sorte qu'ils puissent être considérés comme des domaines planifiés au moment de la conception (ou du remaniement) de l'enquête. Si, pour des considérations budgétaires, on ne peut obtenir d'estimations fiables pour certains domaines particulièrement petits, les concepteurs d'une enquête devraient envisager sérieusement, de concert avec les promoteurs, la possibilité de regrouper des domaines, de réunir des estimations d'enquêtes différentes ou encore de ne pas produire d'estimations. Certains domaines ne peuvent être définis à l'avance. Ces domaines non planifiés appellent des méthodes d'estimation spéciales.

Plan de sondage: Il est rare dans la pratique de trouver un plan "optimal", que ce soit au niveau national ou provincial ou pour un domaine en particulier. Le plus souvent, on introduira des compromis à différentes étapes de l'échantillonnage et de la collecte des données pour répondre aux contraintes théoriques et opérationnelles. Selon les besoins en données, les estimations pour domaines devraient aussi faire partie intégrante de ces compromis. Nous allons étudier deux façons de tenir compte des besoins en données régionales à l'étape de l'élaboration du plan de sondage, notamment la répartition de l'échantillon et le degré de mise en grappes de l'échantillon.

Méthode de répartition: De façon générale, une méthode de répartition optimale pour des estimations nationales permet de répartir les échantillons dans les provinces au prorata de leur population. La fiabilité des estimations pour les petites provinces en souffre. Il est donc préférable d'effectuer une répartition intermédiaire. Il y a diverses façons d'effectuer cette répartition, selon l'importance que l'on accorde aux estimations provinciales ou aux estimations des États (dans le cas de É.-U.). Dans le cas des grandes provinces, une faible réduction de la taille de l'échantillon aura généralement peu d'effet sur la fiabilité des données pour ces provinces (ou des données nationales), mais une augmentation équivalente de la taille d'échantillon dans les petites provinces aura un effet positif appréciable sur la fiabilité des données relatives à ces provinces.

Le même raisonnement s'applique aux domaines planifiés à l'intérieur des provinces car dans la plupart des cas, les répartitions optimales autorisent un certain "jeu" et les concepteurs peuvent exploiter cette caractéristique en redistribuant l'échantillon par un transfert d'unités des grandes régions vers des domaines planifiés de taille moindre.

Mise en grappes: Les enquêtes-ménages de grande envergure comportent habituellement des plans à plusieurs degrés stratifiés avec des unités primaires d'échantillonnage relativement grandes afin d'assurer un bon rapport coût-efficacité pour les estimations nationales et provinciales. Ce genre de plans impliquent donc de nombreuses grappes et nuisent par conséquent à la production de statistiques pour des domaines spatiaux non planifiés en ce sens que, sous l'effet du hasard, certains domaines peuvent contenir de nombreuses unités d'échantillon alors que d'autres peuvent n'en contenir aucune. Étant donné l'importance des estimations pour domaines, on devrait s'efforcer de limiter le plus possible la formation de grappes dans l'échantillon. À cet égard, les facteurs suivants jouent un rôle important: choix de la base de sondage, choix des unités d'échantillonnage et de la taille de ces unités, nombre de strates et leur taille et degrés d'échantillonnage. L'objectif est de réduire au maximum les effets de plan étant donnée les contraintes opérationnelles.

Estimation: Quel que soit le degré d'attention que l'on prête aux estimations pour domaine dans les premières étapes de la planification et de la conception d'une enquête, il existera toujours de petits domaines pour lesquels il faudra appliquer des méthodes d'estimation spéciales afin d'obtenir des valeurs suffisamment justes. Depuis quelque temps, on s'intéresse d'une manière particulière aux estimateurs synthétiques; ceux-ci empruntent de l'information à des domaines qui ressemblent au domaine étudié. Or, comme les estimateurs synthétiques sont très sensibles à l'hypothèse de la similarité des domaines, il en faudrait peu pour que le biais théorique soit élevé et que l'utilisation de ces estimateurs soit remise en question. Préoccupés du biais théorique, les concepteurs de plan de sondage ont proposé de combiner des estimateurs directs et des estimateurs synthétiques dans le but de résoudre le problème du biais théorique tout en essayant de conserver les qualités de l'estimateur synthétique. On s'est servi de méthodes empiriques de Bayes et de méthodes similaires pour pondérer chaque composante des estimateurs combinés. Nous traiterons brièvement ce sujet dans la section 6, qui porte sur l'estimation.

5. CONSIDÉRATIONS RELATIVES AU PLAN D'ÉCHANTILLONNAGE

5.1 Introduction

En règle générale, on pense résoudre le problème des petits domaines (ou des petites régions) par l'estimation. Or, comme nous l'avons mentionné dans la section précédente, il est possible d'envisager la question dès l'étape de la conception du plan de sondage. Pour illustrer nos propos, nous allons nous servir de l'Enquête sur la population active (EPA) du Canada.

Voici les caractéristiques du plan d'échantillonnage actuel de l'EPA: l'Enquête sur la population active du Canada est une enquête mensuelle menée auprès de 59,000 ménages qui sont échantillonnés en plusieurs étapes par diverses méthodes. L'unité d'échantillonnage ultime, c.-à-d. le ménage, demeure dans l'échantillon pendant six mois, puis est remplacée. Les unités des degrés supérieurs (unités primaires d'échantillonnage (UPÉ), grappes) sont

aussi renouvelées périodiquement. Chacune des dix provinces du Canada est divisée en régions économiques (RÉ), lesquelles, en vertu de l'EPA, sont subdivisées en régions autoreprésentatives (grandes villes et villes de taille moyenne) et en régions non autoreprésentatives (le reste de la RÉ). La stratification et l'échantillonnage s'effectuent à l'intérieur de ces régions, et le nombre de degrés d'échantillonnage de même que les unités d'échantillonnage varient selon le type de région. Par exemple, le nombre de degrés d'échantillonnage est de trois dans les régions non autoreprésentatives mais seulement deux dans les régions autoreprésentatives. Pour une description détaillée du plan d'échantillonnage de l'EPA, prière de se référer à Singh et coll. (1990).

5.2 Degrés d'échantillonnage et unités d'échantillonnage

Les bases aréolaires ont rapport généralement à l'échantillonnage en grappes, c'est-à-dire que les unités d'échantillonnage du premier degré sont habituellement des superficies qui contiennent un certain nombre d'unités du deuxième degré. Si on dispose d'une liste des unités du deuxième degré, on peut alors effectuer un échantillonnage à même la liste, ce qui donne un échantillon comportant un moins grand nombre de grappes. Nous aurons ainsi non seulement des estimations de meilleure qualité (grâce à des effets de plan moindres), mais aussi de meilleures estimations régionales pour les domaines non planifiés. La seconde conséquence est vraie car une répartition plus uniforme de l'échantillon a de fortes chances de "placer" des unités échantillonnées dans un domaine spatial non planifié. Inversement, dans le cas d'un plan avec formation de grappes, il arrive souvent qu'un domaine renferme un nombre suffisant d'unités de l'échantillon du fait qu'il contient des grappes échantillonnées, tandis qu'un autre domaine semblable ne contiendra pas suffisamment de grappes pour produire de bonnes estimations.

Nous avons examiné deux solutions pour réduire la formation de grappes dans l'EPA: i) remplacer la base aréolaire (avec plan à deux degrés) pour les grandes villes par une nomenclature fondée sur le Registre des adresses ou ii) réduire le nombre de degrés d'échantillonnage pour les régions rurales et les petits centres urbains. Le Registre des adresses, qui a été créé dans le but d'améliorer le taux de couverture du recensement du Canada de 1991 (Swain, Drew, Lafrance et Lance 1992), contient des adresses et des numéros de téléphone ainsi que de l'information géographique se rapportant à des logements pour chaque secteur de dénombrement (SD). La première solution examinée consistait à tirer un échantillon aléatoire simple stratifié de logements dans le Registre des adresses. À cet échantillon pouvait s'ajouter un second tiré d'une base de croissance qui contient des logements qui ne figurent pas dans le registre des adresses post-censitaire. Seulement, la manière de traiter la croissance est devenue le principal obstacle dans l'application de la première solution car il n'a pas été possible d'élaborer et de tester une méthode d'un bon rapport coût-efficacité dans les délais requis par le projet de remaniement de l'EPA. Néanmoins, on travaille actuellement à l'élaboration d'une méthode de mise à jour du registre des adresses post-censitaire en vue des enquêtes et des recensements futurs.

En ce qui concerne la deuxième solution, on a envisagé la modification d'unités et la réduction du nombre de degrés d'échantillonnage pour les régions non urbaines en gardant en mémoire l'idée qu'une réduction de la formation en grappes se traduit par de meilleures estimations régionales. Étant donné l'évolution qu'ont connue les méthodes de collecte de données (de l'interview directe à l'interview téléphonique à l'interview assistée par ordinateur), les analyses qui visaient auparavant à trouver un équilibre entre le coût d'une enquête et la variance désirée sont révolues. Plus de 80 pour cent des interviews de l'EPA sont maintenant effectuées par téléphone. Avec l'accroissement du nombre des interviews téléphoniques et la diminution connexe du nombre des déplacements, il est presque toujours possible maintenant de supprimer le premier degré d'échantillonnage et de prélever directement des SD.

5.3 Stratification

Une méthode de stratification conforme à l'esprit de l'analyse qui s'est faite précédemment sur la taille de l'UPÉ consiste à remplacer les grandes strates par de nombreuses petites strates dans l'espoir qu'un domaine redéfini ou qu'un domaine non planifié renferment des strates presque complètes. La taille des domaines sera ainsi plus stable.

Il peut y avoir plusieurs régions en chevauchement pour lesquelles on a besoin d'estimations. Par exemple, chaque province canadienne est divisée en régions économiques (RÉ) et en régions de l'Assurance-chômage (RAC). Une solution consiste à considérer tous les secteurs créés par l'intersection des régions comme des strates. Ainsi, au Canada, le recoupement des 71 RÉ et des 61 RAC produit en tout 133 régions en chevauchement, ce qui est raisonnable. Dans certains cas toutefois, le nombre de régions chevauchantes pourrait être trop élevé pour que les opérations soient efficaces. En outre, certains des secteurs pourraient avoir un effectif minime, ce qui en ferait des strates inutilisables.

En combinant la réduction de la formation en grappes et la diminution de la taille des strates, nous espérons avoir un plan de sondage qui pourra mieux répondre aux exigences des petites régions. Par exemple, les plans de ce genre devraient offrir toute la souplesse voulue pour répondre efficacement aux exigences des RÉ et des RAC et traiter les changements de définition des régions.

5.4 Répartition

Si nous connaissons d'avance les définitions des petites régions, nous pourrons peut-être considérer ces régions comme des domaines planifiés et en tenir compte au moment de la conception de l'enquête. Le concepteur peut s'appliquer à répartir l'échantillon dans les petites régions de manière à rendre possible la production d'estimations fiables. Pour de grandes enquêtes comme l'Enquête sur la population active du Canada, cette approche peut, à

tout le moins en théorie, favoriser la production d'un grand nombre d'estimations régionales. Si nous avons un échantillon mensuel de 59,000 ménages et si nous supposons qu'il faut, par exemple, 100 ménages par mois pour produire des estimations trimestrielles fiables, on peut diviser le pays en quelque 600 régions non chevauchantes, chacune contenant avec certitude un nombre suffisant d'unités de l'échantillon. De la même manière, en groupant des régions, on obtiendra des territoires qui contiendront un nombre suffisant d'unités pour produire des estimations mensuelles fiables.

Il existe diverses méthodes de répartition de l'échantillon. Dans une perspective du haut vers le bas, on détermine tout d'abord la taille de l'échantillon provincial; on répartit ensuite l'échantillon entre les régions infra-provinciales. Il se peut toutefois que la taille de l'échantillon provincial empêche que le degré de fiabilité requis pour les estimations infra-provinciales soit respecté. Dans une perspective du bas vers le haut, on répartit tout d'abord l'échantillon entre les régions infra-provinciales de manière à répondre aux objectifs de fiabilité pour chaque région. En conséquence, ces régions devraient normalement avoir des tailles d'échantillon comparables. Avec cette méthode, il se peut que la taille de l'échantillon provincial soit plus grande que celle déterminée par la méthode précédente. Quelle que soit la méthode utilisée, il faudra normalement apporter des corrections aux chiffres initiaux. La répartition finale sera une sorte de compromis entre la répartition proportionnelle et la répartition uniforme. En pratique, le concepteur de plan de sondage doit se livrer à un exercice complexe où il doit tenir compte à la fois des conditions de fiabilité au niveau provincial, des exigences au niveau infra-provincial pour un ou plusieurs groupes de régions, du coût total de l'enquête et des conditions sur le terrain.

La méthode de répartition qui a été choisie pour la version remaniée de l'EPA peut être utile pour d'autres enquêtes. La répartition de l'échantillon s'est faite en deux étapes: on a tout d'abord réparti un échantillon de base de 42,000 ménages de manière à obtenir de bonnes estimations aux niveaux national et provincial; on a ensuite réparti le reste de l'échantillon de manière à obtenir des estimations infra-provinciales le plus justes possible. Cette formule de répartition permettra d'obtenir des estimations fiables pour presque tous les domaines planifiés. Cette solution de compromis s'est traduite par des pertes mineures au niveau provincial mais par des gains appréciables au niveau infra-provincial. Par exemple, le c.v. théorique pour les "personnes en chômage" en Ontario et au Québec est 3.2 et 3.0% respectivement, au lieu de 2.8 et 2.6%. Les pourcentages correspondants pour le Canada sont 1.51 et 1.36. Si on vise l'optimisation au niveau provincial, on obtient des c.v. pouvant atteindre 17.7% pour les régions de l'A.-C. Grâce à la formule de répartition décrite ci-dessus, ces mêmes c.v. ne dépassent pas 9.4%.

Redistribution de l'échantillon: Le concepteur jouit normalement d'une certaine marge de manoeuvre pour déplacer des unités de l'échantillon d'une région à une autre. Par exemple, si on réduit de 1,000 ménages la taille de l'échantillon d'une grande province et qu'on augmente d'autant la taille de l'échantillon d'une petite province, la qualité des estimations pour la grande province s'en trouvera peu affectée alors que celle des estimations pour la petite province en sera grandement améliorée. On peut effectuer des transferts semblables à l'intérieur d'une province.

5.5 Autres considérations

Modification de la définition des petites régions: Les concepteurs de plan de sondage doivent se faire à l'idée que les définitions de domaines planifiés peuvent changer au cours de l'existence d'une enquête et qu'ils peuvent devoir considérer de nouveaux domaines comme des domaines non planifiés. Par exemple, il se peut très bien que la définition des régions de l'Assurance-chômage soit modifiée dans les deux ou trois années qui vont suivre la mise en application du nouveau plan de sondage de l'EPA en 1995. La meilleure chose que le concepteur pourra faire pour parer à cette difficulté à l'étape de l'élaboration sera de choisir comme unités de base des régions dites normalisées (par ex., des régions propres au recensement dont la définition varie peu) et d'espérer que les régions redéfinies correspondent à un groupement de régions normalisées. C'est la solution qui est préconisée dans la version remaniée de l'EPA.

Une autre solution consisterait à adopter une méthode de mise à jour. Cela nécessiterait le tirage de nouvelles unités; ce tirage serait effectué de manière à maximiser le chevauchement des unités tirées initialement et des nouvelles unités. De cette manière, on réduirait au maximum le nombre de nouvelles unités à inscrire sur la liste. On réduirait aussi au maximum les autres contraintes opérationnelles, comme la nécessité d'embaucher d'autres intervieweurs.

6. ESTIMATION

Cette section a pour but d'exposer quelques-unes des méthodes d'estimation de totaux pour petites régions. Nous ne cherchons pas ici à faire une analyse exhaustive de ces méthodes; nous décrivons plutôt l'évolution des recherches sur l'estimation pour petites régions. Pour une analyse détaillée de la question, le lecteur est prié de consulter le tout récent article de Ghosh et Rao (1993). Pour faciliter notre exposé, nous allons classer les méthodes d'estimation pour petites régions en deux groupes. (Il existe beaucoup d'autres possibilités de classement). Le premier groupe comprend les estimateurs de plan, c.-à-d. des estimateurs (approximativement) non biaisés selon le plan; dans ce groupe on trouve les estimateurs directs et les estimateurs directs modifiés. Comme nous l'avons mentionné plus haut, les estimateurs de plan laissent souvent à désirer à cause de leur variance élevée, qui s'explique par le faible effectif des petites régions (ou même un effectif nul). Le second groupe est désigné comme le groupe des estimateurs indirects (ou estimateurs de modèle) et il comprend les estimateurs synthétiques et les estimateurs combinés. Certains de ces estimateurs font l'objet d'une comparaison empirique dans une version antérieure de cet article; voir Singh, Gambino et Mantel (1992).

6.1 Estimateurs de plan

Estimateurs directs: Les estimateurs directs pour petite région reposent sur des données d'enquête qui proviennent uniquement de la petite région; à ces données peuvent s'ajouter parfois des données auxiliaires tirées du recensement ou de fichiers administratifs. La forme la plus simple de l'estimateur direct d'un total est l'estimateur par facteur d'extension,

$$\hat{Y}_{e,a} = \sum_{i \in s_a} w_i y_i, \tag{6.1}$$

où s_a est la portion de l'échantillon contenue dans la petite région a et w_i est le poids de sondage de l'unité i. Cet estimateur est non biaisé; cependant, il peut être caractérisé par une forte variabilité du fait que la taille d'échantillon dans la région a est aléatoire.

Si la taille de la population, N_a, est connue, on peut se servir d'un estimateur de stratification a posteriori,

$$\hat{Y}_{pst,a} = N_a \sum_{i \in s_a} w_i y_i \Big/ \sum_{i \in s_a} w_i = N_a \hat{Y}_{e,a} / \hat{N}_{e,a} = N_a \bar{y}_{e,a}. \tag{6.2}$$

Cet estimateur est plus stable que l'estimateur par facteur d'extension; cependant, il peut être entaché d'un biais d'estimation par quotient dans les enquêtes complexes.

Si l'échantillonnage est stratifié et que $N_{h,a}$, l'effectif de la population contenu dans la strate h et la petite région a, est connu, on peut utiliser un autre estimateur de stratification a posteriori: $\hat{Y}_{st,pst,a} = \sum_h (N_{h,a} \sum_{i \in s_{h,a}} w_i y_i / \sum_{i \in s_{h,a}} w_i) = \sum_h N_{h,a} \hat{Y}_{h,e,a} / \hat{N}_{h,e,a} = \sum_h N_{h,a} \bar{y}_{h,a}$. Il peut aussi s'agir de strates formées a posteriori au lieu de strates initiales.

L'estimation par quotient ressemble à l'estimation de stratification a posteriori avec cette différence qu'on utilise une variable auxiliaire au lieu des effectifs N_a et $N_{h,a}$. Par exemple, si x est une covariable pour laquelle on connaît le total de petite région, X_a, ou le total de strate de petite région, $X_{h,a}$, nous pouvons définir les estimateurs par quotient

$$\hat{Y}_{r,a} = X_a \hat{R}_a \quad \text{et} \quad \hat{Y}_{st,r,a} = \sum_h X_{h,a} \hat{R}_{h,a}, \tag{6.3}$$

où $\hat{R}_a = \hat{Y}_{e,a} / \hat{X}_{e,a}$ est une estimation du rapport Y_a / X_a et $\hat{R}_{h,a} = \hat{Y}_{h,e,a} / \hat{X}_{h,e,a}$.

Par ailleurs, l'estimateur par régression tente d'expliquer la différence entre la valeur de covariables pour l'effectif d'une petite région et celle pour un sous-échantillon au moyen d'une équation de régression estimée qui met en relation la variable étudiée y et les covariables x. Un des avantages de l'estimation par régression est qu'elle s'étend facilement à des covariables vectorielles. L'estimateur est défini

$$\hat{Y}_{reg,a} = \hat{Y}_a + \hat{\beta}_a (X_a - \hat{X}_a), \tag{6.4}$$

où $\hat{Y}_a$ peut être un estimateur par facteur d'extension ou un estimateur de stratification a posteriori, $\hat{X}_a$ doit être calculé de la même manière que $\hat{Y}_a$, et $\hat{\beta}_a = \sum_{i \in s_a} v_i^{-1} w_i y_i x_i' \{ \sum_{i \in s_a} v_i^{-1} w_i x_i x_i' \}^{-1}$, où v_i désigne les poids donnés pour la régression. Notons que $\hat{\beta}_a = \hat{R}_a$ lorsque x est une grandeur scalaire et $v_i = x_i$. Lorsque $\hat{Y}_a$ et $\hat{X}_a$ sont des estimateurs par facteur d'extension, l'estimateur (6.4) est aussi appelé estimateur par régression généralisé. Il est approximativement non biaisé selon le plan à la condition qu'il en soit de même pour $\hat{Y}_a$ et $\hat{X}_a$.

Comme l'estimateur par quotient, l'estimateur par régression peut être appliqué à des strates initiales ou à des strates formées a posteriori.

Estimateurs directs modifiés: Les estimateurs directs modifiés peuvent utiliser des données d'enquête qui ne proviennent pas du domaine étudié; cependant, ils demeurent des estimateurs approximativement non biaisés selon le plan. Selon notre définition, l'estimateur direct modifié est un estimateur direct qui comporte un facteur de correction synthétique pour biais de modèle; comme l'espérance du facteur de correction serait approximativement nulle par rapport au plan, l'estimateur modifié est approximativement non biaisé selon le plan si l'estimateur direct est aussi approximativement non biaisé selon le plan. On a un exemple d'estimateur direct modifié si on remplace $\hat{\beta}_a$ dans (6.4) par un estimateur synthétique, $\hat{\beta} = \sum_{i \in s} v_i^{-1} w_i y_i x_i' \{ \sum_{i \in s} v_i^{-1} w_i x_i x_i' \}^{-1}$; nous désignerons cet estimateur par $\hat{Y}_{sreg,a}$. $\hat{\beta}$ sera généralement plus stable que $\hat{\beta}_a$; le choix entre l'un et l'autre dépendra du rapport entre la grandeur de la variance de $\hat{\beta}_a$ et la variation des valeurs $\beta_a s$ pour les régions a. Une solution intermédiaire serait d'opter pour une moyenne pondérée, $\lambda_a \hat{\beta}_a + (1 - \lambda_a) \hat{\beta}$, où la valeur de λ_a est choisie judicieusement; le choix d'une valeur de λ_a est examiné dans la partie de la section 6.2 qui porte sur les estimateurs combinés. On a un autre exemple d'estimateur direct modifié si on remplace $\hat{\beta}_a$ dans (6.4) par $\hat{R} = \hat{Y}_e / \hat{X}_e$; notons que $\hat{R}$ est un cas particulier de $\hat{\beta}$ lorsque x est une grandeur scalaire et $v_i = x_i$.

6.2 Estimateurs indirects

Estimateurs synthétiques: Les méthodes d'estimation synthétique reposent sur l'hypothèse selon laquelle la petite région ressemble d'une certaine manière à une autre région, souvent plus grande, dans laquelle elle est contenue. Les estimations relatives à cette autre région seraient généralement plus précises que les estimations relatives à la petite région. L'estimateur synthétique serait alors un estimateur à faible variance, qui pourrait toutefois être fortement biaisé si l'hypothèse mentionnée plus haut ne se vérifiait pas.

L'une des formes les plus simples d'estimateur synthétique repose sur l'hypothèse que la moyenne pour petite région est égale à la moyenne globale. On a ainsi l'estimateur synthétique de moyenne

$$\hat{Y}_{syn,m,a} = N_a \sum_{i \in s} w_i y_i \Big/ \sum_{i \in s} w_i = N_a \bar{y}. \tag{6.5}$$

Une forme plus courante d'estimateur synthétique est celle qui repose sur la stratification ou la stratification a posteriori, soit

$$\hat{Y}_{syn,st,m,a} = \sum_h N_{h,a} \sum_{i \in s_h} w_i y_i \Big/ \sum_{i \in s_h} w_i = \sum_h N_{h,a} \bar{y}_h .$$

Comme dans le cas des estimateurs directs, les estimateurs synthétiques par quotient peuvent utiliser des données auxiliaires en plus des effectifs N_a et $N_{h,a}$. Par exemple, les estimateurs synthétiques par quotient basés sur une covariable x sont définis le plus souvent sous la forme

$$\hat{Y}_{syn,r,a} = X_a \hat{Y}_e / \hat{X}_e \quad \text{et} \quad \hat{Y}_{syn,st,r,a} = \sum_h X_{h,a} \hat{Y}_{h,e} / \hat{X}_{h,e}, \tag{6.6}$$

où $\hat{Y}_e = \sum_{i \in s} w_i y_i$ est l'estimateur par facteur d'extension du total de population pour y et $\hat{Y}_{h,e} = \sum_{i \in s_h} w_i y_i$. $\hat{X}_e$ et $\hat{X}_{h,e}$ sont définis de la même manière. Ces estimateurs ont été étudiés par Gonzalez (1973), Gonzalez et Waksburg (1973) et Ghangurde et Singh (1977,1978), pour n'en nommer que quelques-uns.

Singh et Tessier (1976) ont proposé une autre forme d'estimateur synthétique par quotient, où X remplace $\hat{X}_e$:

$$\tilde{Y}_{syn,r,a} = X_a \hat{Y}_e / X. \tag{6.7}$$

Les deux estimateurs, $\hat{Y}_{syn,r,a}$ et $\tilde{Y}_{syn,r,a}$, ont le même biais synthétique et le biais de $\hat{Y}_{syn,r,a}$ est négligeable pour de grands échantillons. Le choix entre l'un et l'autre de ces estimateurs dépend de ρ, le coefficient de corrélation de $\hat{Y}_e$ et $\hat{X}_e$. On peut montrer que pour de grands échantillons, $V(\hat{Y}_{syn,r,a}) \leq V(\tilde{Y}_{syn,r,a})$ si $\rho \geq 0.5 c_x / c_y$, où c_x et c_y sont les coefficients de variation de $\hat{X}_e$ et de $\hat{Y}_e$ respectivement. Dans la plupart des cas, lorsque ρ est élevé ou que la population est asymétrique, on préférera $\hat{Y}_{syn,r,a}$; cependant, si c_x est élevé et que la corrélation n'est pas trop forte, $\tilde{Y}_{syn,r,a}$ pourra être un meilleur choix.

Dans certains cas, il peut exister de l'information sur une autre variable auxiliaire (z) en plus de x. On peut alors construire un estimateur synthétique par quotient à deux variables:

$$\hat{Y}^{(2)}_{syn,r,a} = \gamma_a X_a \hat{Y}_e / \hat{X}_e + (1 - \gamma_a) Z_a \hat{Y}_e / \hat{Z}_e, \tag{6.8}$$

où la valeur de γ_a est choisie judicieusement. On peut aussi envisager un estimateur synthétique par quotient à plusieurs variables si l'on se fonde sur l'article d'Olkin (1958).

L'estimation synthétique par régression ressemble à l'estimation synthétique par quotient,

$$\hat{Y}_{syn,reg,a} = \hat{\beta} X_a ,$$

$$\hat{\beta} = \sum_{i \in s} v_i^{-1} w_i y_i x_i' \left\{ \sum_{i \in s} v_i^{-1} w_i x_i x_i' \right\}^{-1}. \tag{6.9}$$

Comme dans le cas des estimateurs directs, l'estimateur synthétique par régression peut être appliqué à des strates initiales ou à des strates formées a posteriori. Royall (1979) a proposé un variante de cet estimateur, soit $\hat{Y}_{syn,Roy,a} = \sum_{i \in s_a} y_i + \hat{\beta}(X_a - \sum_{i \in s_a} x_i)$, où la somme des valeurs y qui se rapportent uniquement aux unités non incluses dans l'échantillon est estimée de façon synthétique.

Remarque: Les estimateurs directs modifiés présentés dans la section 6.1 peuvent aussi être exprimés comme des estimateurs synthétiques par quotient ou des estimateurs synthétiques par régression comportant un facteur de compensation du biais basé sur le plan. Nous pouvons écrire par exemple l'expression $\hat{Y}_{sreg,a} = \hat{Y}_{syn,reg,a} + (\hat{Y}_a - \hat{\beta}\hat{X}_a)$, où $\hat{Y}_a - \hat{\beta}\hat{X}_a$ est la valeur estimée du biais de $\hat{Y}_{syn,reg,a}$. De même, nous pouvons exprimer $\hat{Y}_{sreg,a}$ comme l'estimateur de Royall, $\hat{Y}_{syn,Roy,a}$, accompagné d'un facteur de compensation du biais basé sur le plan.

Purcell et Kish (1980) étudient une autre forme d'estimation synthétique qu'ils appellent SPREE (pour "structure preserving estimation") et qui s'applique à l'estimation d'effectifs pour petites régions. Selon cette méthode, on combine des données antérieures détaillées, provenant peut-être d'un recensement, avec des estimations d'enquête courantes moins détaillées dans le but d'obtenir des estimations détaillés de valeurs courantes. Cette méthode suppose que certains rapports parmi les données détaillées ne varient pas dans le temps.

Estimateurs combinés: L'estimateur combiné est défini comme la moyenne pondérée d'un estimateur de plan et d'un estimateur synthétique, c'est-à-dire

$$\hat{Y}_{com,a} = \lambda_a \hat{Y}_{des,a} + (1 - \lambda_a) \hat{Y}_{syn,a}, \tag{6.10}$$

où la valeur de λ_a est choisie judicieusement. L'estimateur combiné vise à faire l'équilibre entre la possibilité pour l'estimateur synthétique d'être biaisé et l'instabilité de l'estimateur de plan. Il existe trois grandes méthodes pour déterminer les poids λ_a en (6.10): ces poids peuvent être établis à l'avance, ils peuvent dépendre de la taille de l'échantillon ou encore, ils peuvent dépendre des données.

La première méthode de pondération, et la plus simple, consiste à fixer des poids à l'avance (par exemple, utiliser une moyenne simple). Or, cette méthode ne tient aucunement compte de la précision réelle de l'estimateur de plan. L'estimateur de plan pour la petite région a est plus précis pour certains échantillons réels que pour d'autres. Le poids attribué à l'estimateur de plan devrait refléter cette différence.

La deuxième grande méthode de pondération des composantes de l'estimateur combiné a un rapport de dépendance avec la taille de l'échantillon; en effet, les poids sont fonction du rapport $\hat{N}_{e,a}/N_a$. Une variante de cette méthode, que nous n'examinons pas ici, a plutôt un rapport de dépendance avec les valeurs empiriques d'une covariable x; par exemple, les poids peuvent être fonction de $\hat{X}_{des,a}/X_a$ ou de $S^2_{x,a}/\sigma^2_{x,a}$, où $S^2_{x,a}$ est la variance empirique de $\hat{X}_{des,a}$, étant donné $\hat{N}_{e,a}$ ou une autre caractéristique pertinente de l'échantillon réel, et $\sigma^2_{x,a}$ est la variance inconditionnelle de $\hat{X}_{des,a}$.

Des estimateurs de ce genre ont été proposés antérieurement. Drew, Singh et Choudhry (1982) ont proposé, par exemple, l'estimateur suivant (tributaire de la taille de l'échantillon):

$$\hat{Y}_{ssd,r,a} = \lambda_a \hat{Y}_{r,a} + (1 - \lambda_a)\hat{Y}_{syn,r,a}, \qquad (6.11a)$$

où

$$\lambda_a = \begin{cases} 1 & \text{si} \quad \hat{N}_{e,a} \geq \delta N_a \\ \hat{N}_{e,a}/\delta N_a & \text{dans le cas contraire} \end{cases} \qquad (6.11b)$$

et la valeur δ est choisie subjectivement pour régler la contribution de la composante synthétique. Särndal (1984) a proposé l'estimateur suivant

$$\hat{Y}_{ssd,reg,a} = \lambda_a \hat{Y}_{sreg,a} + (1 - \lambda_a)\hat{Y}_{syn,reg,a}, \qquad (6.12)$$

où $\lambda_a = \hat{N}_{e,a}/N_a$. Rao (1986) a proposé de modifier cet estimateur dans le sens suivant: λ_a aurait la valeur 1 si $\hat{N}_{e,a} \geq N_a$. Särndal et Hidiroglou (1989) ont amélioré la proposition de Rao en ajoutant la relation suivante: $\lambda_a = (\hat{N}_{e,a}/N_a)^{h-1}$ si $\hat{N}_{e,a} < N_a$, la valeur h étant choisie de façon rationnelle pour régler la contribution de la composante synthétique.

Dans la pratique, ce qui compte le plus lorsqu'on utilise des estimateurs de ce genre est le biais de la composante synthétique. Cette composante devrait être affectée d'un poids tel que le biais se situe dans des limites raisonnables. Par exemple, on se sert actuellement de l'estimateur tributaire de la taille d'échantillon de Drew, Singh et Choudhry (1982) dans l'Enquête sur la population active du Canada pour produire des estimations pour domaine, compte tenu de ce que l'estimation par régression généralisée remplace l'estimation par quotient et que $\delta = 2/3$. Pour une majorité de domaines, le poids attribué à la composante synthétique est nul puisque l'estimateur direct satisfait à lui seul les exigences de fiabilité. Pour les autres domaines, le poids de la composante synthétique est d'environ 10% en moyenne et ne dépasse jamais 20%. Selon le biais qu'on est prêt à accepter, la valeur de δ peut se situer dans l'intervalle [2/3,3/2] pour la plupart des cas.

La troisième méthode de pondération met les poids dans un rapport de dépendance avec les données. Les poids optimaux nécessaires à la combinaison de deux estimateurs dépendent généralement de l'erreur quadratique moyenne des estimateurs et de leur covariance. Ces valeurs sont le plus souvent inconnues mais elles peuvent être estimées à partir des données. Dans le cas des estimateurs combinés, il faudrait normalement modéliser le biais de la composante synthétique. On trouve dans Fay et Herriot (1979) un exemple très connu d'application de cette méthode, qui est un des premiers du genre dans la littérature statistique. Fay et Herriot modélisent les biais des estimateurs synthétiques pour les petites régions comme des effets aléatoires indépendants ayant une variance inconnue mais fixe. En termes plus précis, si $\hat{Y}_{des,a}$ est l'estimateur de plan, ils considèrent alors le modèle $Y_a = X_a\beta + \alpha_a$ et l'estimateur $\hat{Y}_{des,a} = Y_a + \epsilon_a$, où $\alpha_a \sim (0,\sigma^2)$, $\epsilon_a \sim (0,\nu_a^2)$, et α_a et ϵ_a sont indépendants et non corrélés pour a, σ^2 est inconnu et les ν_a^2 sont supposés connus (en pratique, ils devraient être estimés). Pour une valeur donnée de σ^2, on peut calculer les poids optimaux nécessaires à la combinaison de $\hat{Y}_{des,a}$ et de $X_a\hat{\beta}$. On calcule une estimation de σ^2 à l'aide de la méthode des constantes d'ajustement, puis on introduit cette estimation dans la formule des poids optimaux. Pour parer à la défaillance du modèle, on prévoit une troncation de l'estimation calculée si l'écart entre celle-ci et l'estimation directe est supérieur à un multiple déterminé de ν_a. Schaible (1979) et Battese et Fuller (1981) considèrent aussi des poids optimaux λ_a (en 6.12) estimés empiriquement selon des modèles à effets aléatoires semblables pour les totaux de petites régions.

Prasad et Rao (1990) définissent un estimateur de l'erreur quadratique moyenne de l'estimateur de Fay-Herriot qui tient compte de l'estimation des composantes de la variance. Kott (1989) propose un estimateur convergent selon le plan de l'erreur quadratique moyenne mais constate que cet estimateur est très instable.

Une autre méthode consiste à calculer les poids au moyen de données historiques; l'avantage de cette solution est que les poids calculés de cette manière peuvent être plus stables que des poids estimés à l'aide des données d'enquête courantes; néanmoins, cela suppose l'hypothèse que les poids optimaux sont stables.

Remarque: Dans l'estimation tributaire de la taille de l'échantillon, les poids peuvent dépendre de la taille observée du sous-échantillon s_a mais non des valeurs de la variable étudiée. Cette absence de relation de dépendance entre les poids et la variable étudiée présente des avantages et des inconvénients. Un des avantages est que les mêmes poids servent à l'estimation de totaux pour toutes les variables étudiées. On n'a à les calculer qu'une seule fois. Un avantage plus important encore est que l'estimation de la somme de deux variables est égale à la somme des estimations de chaque variable. Du côté des inconvénients, les poids ne tiennent pas compte directement de la précision de l'estimateur de plan pour la variable étudiée ou de la grandeur probable du biais de l'estimateur synthétique.

Combinaison de données de diverses périodes: Dans les enquêtes à passages répétés, il est courant de grouper les données de plusieurs passages dans le but d'accroître la précision des estimations. Selon le plan de renouvellement utilisé dans ces enquêtes, on peut réussir à accroître sensiblement la fiabilité des estimations. Cette technique de groupement ou de mise en moyenne sur plusieurs périodes suscite donc un intérêt particulier en ce qui concerne l'estimation pour domaine, caractérisée par un degré de précision modeste. En ce qui regarde l'estimation pour domaine dans l'Enquête sur la population active du Canada, il est pratique courante d'utiliser un estimateur tributaire de la taille de l'échantillon basé sur une moyenne trimestrielle des estimations du nombre de personnes occupées et du nombre de personnes en chômage. Grâce au plan de renouvellement de six mois utilisé dans cette enquête, la mise en moyenne sur trois mois permet d'accroître du tiers

la taille de l'échantillon. Si les échantillons des différentes périodes se chevauchent parfaitement, la mise en moyenne ne se traduit par aucun gain d'efficacité. Cependant, pour d'autres plans de renouvellement, ce procédé pourrait plus que doubler la taille de l'échantillon pour des estimations pour domaine. Mais les estimations groupées posent un problème théorique en ceci qu'elles représentent une valeur moyenne du paramètre étudié (par ex., niveau de chômage) pour une période de trois mois par exemple.

Dans l'estimation composite, on combine l'estimateur de plan de la période courante avec l'estimateur composite de la période précédente, ce dernier étant redressé au moyen d'une estimation de la variation basée sur l'échantillon commun. Ce procédé a été appliqué pour autre chose que l'estimation pour petites régions par Jessen (1942) et Patterson (1950), pour ne nommer que ceux-là. Binder et Hidiroglou (1988) font un examen de la question. Les poids nécessaires à la combinaison sont ordinairement des estimations des poids optimaux dans l'hypothèse que ceux-ci ne varient pas dans le temps. Ces poids tributaires des données ont pour inconvénient de produire des estimations incohérentes pour diverses caractéristiques et les sommes de ces caractéristiques.

Dans l'estimation pour petites régions, on a recours depuis peu à des méthodes basées sur des séries temporelles pour les enquêtes périodiques. On modélise la relation entre les paramètres étudiés pour différentes périodes, puis on se sert du modèle afin d'accroître l'efficience des estimations pour le passage courant. Dans la plupart des cas, il faut aussi tenir compte, par la modélisation ou autrement, de la non-indépendance des échantillons des divers passages, condition attribuable au plan de renouvellement de l'échantillon. Cette approche fondée sur les séries chronologiques est traitée dans les ouvrages suivants: Choudhry et Rao (1989); Pfeffermann et Burck (1990); Singh, Mantel et Thomas (1994) et Singh et Mantel (1991). Tous ces ouvrages présentent des versions généralisées du modèle de Fay-Herriot selon lesquelles les paramètres de régression, les effets de petite région et les erreurs d'enquête varient selon divers modèles chronologiques. Le vecteur d'estimations régionales calculé par cette méthode peut être exprimé comme la moyenne pondérée du vecteur des estimations de plan et du vecteur des estimations synthétiques, celles-ci étant fondées sur des données de périodes antérieures et les valeurs courantes des covariables; cependant, la matrice des poids sera très rarement une matrice diagonale, de sorte que l'estimateur pour une petite région quelconque dépendra généralement aussi bien des estimations de plan et des estimations synthétiques d'autres petites régions.

7. CONCLUSION

Les concepteurs de plan de sondage doivent surmonter beaucoup d'obstacles pour produire des estimations pour domaine qui soient précises et actuelles. Ils doivent tout d'abord convaincre les promoteurs d'enquêtes et les chargés de programme qu'un système conçu spécialement pour produire des estimations nationales et infra-nationales ne peut répondre à tous les besoins en données régionales. On peut réaliser des gains d'efficience appréciables (selon les enquêtes) au niveau du domaine avec très peu de perte aux niveaux supérieurs. Nous avons souligné la nécessité d'élaborer une approche globale, par laquelle on pourrait définir un degré de fiabilité pour les domaines planifiés comme pour les régions de niveau supérieur par des répartitions intermédiaires et par laquelle on pourrait réduire la formation de grappes dans le but d'améliorer les estimations relatives aux domaines non planifiés. Il convient de noter que bon nombre de domaines dits planifiés à l'étape de la conception du plan de sondage peuvent se transformer en domaines non planifiés (révisés) au fil du temps lorsqu'il s'agit d'enquêtes permanentes.

L'approche globale devrait aussi permettre de considérer les estimateurs de plan pour de grands domaines aussi bien que les estimateurs de modèle pour de petits domaines. L'estimateur de modèle devrait être préféré à l'estimateur de plan seulement si l'erreur quadratique moyenne du premier (variance de plan + biais^2) peut être estimée et qu'elle est inférieure à la variance correspondante de l'estimateur de plan **par une marge suffisante**. Il devrait y avoir des estimations de l'erreur quadratique moyenne pour chacun des domaines **pris individuellement**. Les organismes statistiques peuvent choisir de grouper des domaines semblables ou de grouper des estimations de périodes différentes qui se rapportent au même domaine. Ils peuvent même supprimer les estimations touchant certains domaines à cause de leur faible degré de fiabilité ou pour des raisons de protection du secret statistique.

La deuxième tâche qui attend les statisticiens est d'expliquer aux utilisateurs les divers types de mesures de fiabilité qui peuvent s'appliquer à différentes séries d'estimations provenant de la même enquête. Espérons qu'avec la poursuite des recherches sur la validation de modèle et de meilleurs estimations de l'erreur quadratique moyenne, les concepteurs hésiteront moins à utiliser des estimateurs de modèle pour les petits domaines. D'ici là, on devra se servir des estimateurs de modèle avec circonspection même si leurs coefficients de variation sont beaucoup plus petits.

Les recensements, avec les fichiers administratifs en appui, demeureront vraisemblablement la principale source de données socio-économiques pour les petites régions, surtout dans les pays où il se fait un recensement de la population et du logement à tous les cinq ans. En outre, on continuera de se préoccuper des problèmes conceptuels que posent les données des fichiers administratifs tant que les organismes statistiques n'auront pas la chance de participer à l'élaboration des formulaires qui servent à recueillir ces données. En attendant, on ne peut exploiter pleinement cette source de données extrêmement riche que constituent les fichiers administratifs; ce qui est dommage pour l'analyse statistique et surtout pour l'estimation pour domaine.

REMERCIEMENTS

Nous sommes reconnaissants à Jon Rao d'avoir agi comme rédacteur associé pour cet article et nous remercions aussi les rapporteurs pour leurs nombreux conseils utiles.

BIBLIOGRAPHIE

BATTESE, G.E., et FULLER, W.A. (1981). Prediction of county crop areas using survey and satellite data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 500-505.

BINDER, D., et HIDIROGLOU, M.A. (1988). Sampling in time. *Handbook of Statistics*, (Éds. P.R. Krishnaiah et C.R. Rao). New York: Elsevier Science, 187-211.

BRACKSTONE, G.J. (1987a). Small area data: Policy issues and technical challenges. *Small Area Statistics*, (Éds. R. Platek, J.N.K. Rao, C.-E. Särndal et M.P. Singh), New York: Wiley and Sons, 3-20.

BRACKSTONE, G.J. (1987b). Utilisations statistiques des données administratives: questions et défis. *Recueil: Symposium sur les utilisations statistiques des données administratives*, (Éds. J.W. Coombs et M.P. Singh), Statistique Canada, 5-17.

CHOUDHRY, G.H., et RAO, J.N.K. (1989). Estimation de données régionales à l'aide de modèles qui combinent des séries chronologiques et des données transversales. *Recueil: Symposium sur l'analyse des données dans le temps*, (Éds. A.C. Singh et P. Whitridge), Statistique Canada, 71-80.

COOMBS, J.W., et SINGH, M.P. (Éds.) (1987). *Recueil: Symposium sur les utilisations statistiques des données administratives*, Statistique Canada.

DREW, J.D., SINGH, M.P., et CHOUDHRY, G.H. (1982). Evaluation of small area estimation techniques for the Canadian Labor Force Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 545-550.

FAY, R.E., et HERRIOT, R.A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.

FELLEGI, I.P. (1987). Allocution d'ouverture. *Recueil: Symposium sur les utilisations statistiques des données administratives*, (Éds. J.W. Coombs et M.P. Singh), Statistique Canada, 1-2.

GHOSH, M., et RAO, J.N.K. (1993). Small area estimation: An appraisal. À paraître dans *Statistical Science*.

GHANGURDE, P.D., et SINGH, M.P. (1977). Synthetic estimates in periodic household surveys. *Techniques d'enquête*, 3, 152-181.

GHANGURDE, P.D., et SINGH, M.P. (1978). Evaluation of efficiency of synthetic estimates. *Proceedings of the Social Statistics Section, American Statistical Association*, 53-61.

GONZALEZ, M.E. (1973). Use and evaluation of synthetic estimators. *Proceedings of the Social Statistics Section, American Statistical Association*, 33-36.

GONZALEZ, M.E., et WAKSBERG, J. (1973). Estimation of the error of synthetic estimates. Présenté à la première réunion de l'Association Internationale des Statisticiens d'Enquête, Vienna, Austria.

JESSEN, R.J. (1942). Statistical investigation of a farm survey for obtaining farm facts. *Iowa Agricultural Station Research Bulletin*, 304, 54-59.

NORRIS, D., et PATON, D. (1991). L'enquête sociale générale canadienne: bilan des cinq premières années. *Techniques d'enquête*, 17, 245-260.

OLKIN, I. (1958). Multivariate ratio estimation for finite populations. *Biometrika*, 45, 154-165.

PATTERSON, H.D. (1950). Sampling on successive occasions with partial replacement of units. *Journal of the Royal Statistical Society*, Séries B, 12, 241-255.

PFEFFERMANN, D., et BURCK, L. (1990). Estimation robuste pour petits domaines par la combinaison de données chronologiques et transversales. *Techniques d'enquête*, 16, 229-249.

PLATEK, R., RAO, J.N.K., SÄRNDAL, C.-E., et SINGH, M.P. (1987). *Small Area Statistics*. Présentations invitées. New York: Wiley.

PLATEK, R., et SINGH, M.P. (1986). Small Area Statistics, An International Symposium' 85 (articles contribués), Technical Report Series of the Laboratory for Research in Statistics and Probability, Carleton University, University of Ottawa, Canada.

PRASAD, N.G.N., et RAO, J.N.K. (1990). The estimation of mean squared errors of small-area estimators. *Journal of the American Statistical Association*, 85, 163-171.

PURCELL, N.J., et KISH, L. (1980). Postcensal estimates for local areas (or Domains). *Revue Internationale de Statistique*, 48, 3-18.

RAO, J.N.K. (1986). Synthetic estimates, SPREE and best model based predictors. *Proceedings of the Conference on Survey Research Methodology in Agriculture*, American Statistical Association and National Agricultural Statistics Service, USDA, 1-6.

ROYALL, R.M. (1979). Prediction models in small area estimation. Dans *Synthetic Estimates for Small Area*, NIDA Research Monograph Series 24, U.S. Department of Health, Education and Welfare.

SÄRNDAL, C.-E. (1984). Design-consistent versus model-dependent estimators for small domains. *Journal of the American Statistical Association*, 79, 624-631.

SÄRNDAL, C.-E., et HIDIROGLOU, M.A. (1989). Small domain estimation: A conditional analysis. *Journal of the American Statistical Association*, 84, 266-275.

SCHAIBLE, W.L. (1979). A composite estimator for small area statistics. *Synthetic Estimates for Small Area*, NIDA Research Monograph Series 24, U.S. Department of Health, Education and Welfare, Library of Congress catalogue number 79-600067, 36-53.

SCHAIBLE, W.L. (1992). Use of small area estimators in U.S. federal programs. *Small Area Statistics and Survey Designs*, (Vol. 1), (Éds. G. Kalton, J. Kordos et R. Platek). Warsaw: Central Statistical Office, 95-114.

SCHMIDT, R.C. (1952). Short-cut methods for estimating county populations. *Journal of the American Statistical Association*, 47, 232-238.

SINGH, A.C., et MANTEL, H.J. (1991). Estimation composite par espace d'états pour les petites régions. *Recueil: Symposium 91, Questions spatiales liées aux statistiques*, Statistique Canada, 21-30.

SINGH, A.C., MANTEL, H.J., et THOMAS, B.W. (1994). MPLSE à données chronologiques pour petites régions évalués d'après des données d'enquête. *Techniques d'enquête*, 20, 35-46.

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

SINGH, M.P., GAMBINO, J.G., et MANTEL, H. (1992). Issues and options in the provision of small area statistics. *Small Area Statistics and Survey Designs*, (Vol. 1), (Éds. G. Kalton, J. Kordos et R. Platek), Warsaw: Central Statistical Office, 37-75.

SINGH, M.P., et TESSIER, R. (1976). Some estimators for domain totals. *Journal of the American Statistical Association*, 71, 322-325.

SWAIN, L., DREW, J.D., LAFRANCE, B., et LANCE, K. (1992). La création d'un registre des adresses résidentielles pour améliorer la couverture du recensement du Canada de 1991. *Techniques d'enquête*, 18, 139-155.

U.S. STATISTICAL POLICY OFFICE (1993). Indirect Estimators in Federal Programs. Statistical Policy Working Paper 21. Préparé par le sous-comité sur Small Area Estimation, Federal Committee on Statistical Methodology.

VERMA, R.B.P., et BASAVARAJAPPA, J.G. (1987). Recent developments in the regression method for estimation of population for small areas in Canada. *Small Area Statistics*, (Éds. R. Platek, J.N.K. Rao, C.-E. Särndal et M.P. Singh), New York: Wiley and Sons, 46-61.

COMMENTAIRES

W.A. FULLER[1]

Les auteurs méritent des félicitations pour avoir exposé avec brio les problèmes de conception de plan et d'estimation qui ont rapport aux domaines. Les auteurs examinent l'estimation pour domaines planifiés en particulier les cas où les éléments d'un domaine peuvent être identifiés dans la base de sondage de même que l'estimation pour domaines non planifiés, notamment les domaines dont les éléments ne peuvent être identifiés à l'aide de la base de sondage. Cet article est un élément positif supplémentaire dans la littérature de plus en plus abondante qui existe sur la question de l'estimation pour domaine.

Les auteurs font une description particulièrement bonne des opérations de planification, de collecte des données et de traitement qui sont exécutées dans le cadre des enquêtes de Statistique Canada. Ils mentionnent les problèmes qui sont liés depuis toujours à la conception des plans de sondage: arbitrage entre la nécessité de recourir à l'estimation pour domaine et le désir d'obtenir une certaine efficience aux niveaux (d'agrégation) supérieurs, importance de la confidentialité dans l'utilisation de dossiers administratifs en vue d'établir des estimations pour domaine, et importance de la compatibilité des définitions lorsqu'on tente de combiner des données de diverses sources.

Les auteurs font ressortir très bien l'importance de prendre en considération l'estimation pour domaine à l'étape de la conception du plan de sondage, ce que font rarement d'autres auteurs qui s'intéressent particulièrement à l'estimation pour petite région. Comme le soulignent Singh, Gambino et Mantel, le statisticien pourra souvent construire des estimateurs pour domaine qui sont directs et convergents selon le plan s'il exécute avec soin l'étape de la conception. Je suis sûr que les concepteurs de plan de sondage tiennent compte de l'importance de la formation de grappes lorsqu'ils élaborent des enquêtes qui doivent servir à l'estimation pour domaine, mais il est réjouissant de voir qu'on en fait mention explicitement.

Les auteurs décrivent plusieurs types d'estimateur pour domaine. La classification de ces estimateurs fait ressortir le nombre de possibilités qui s'offrent au praticien. Les erreurs quadratiques moyennes théoriques peuvent nous renseigner sur les avantages relatifs des estimateurs. Afin d'illustrer ce genre de comparaison, supposons un échantillon aléatoire simple de taille n prélevé dans une population qui est divisée en k domaines. Supposons qu'on connaît la taille des domaines et la valeur moyenne d'une variable auxiliaire, X, pour le domaine. Considérons les trois estimateurs par régression de la moyenne de domaine

$$\hat{\mu}_{(1)yi} = \bar{y}_{i.} + (\mu_{xi} - \bar{x}_{i.})b_i,$$

$$\hat{\mu}_{(2)yi} = \bar{y}_{i.} + (\mu_{xi} - \bar{x}_{i.})b_.$$

et

$$\hat{\mu}_{(3)yi} = \bar{y}_{..} + (\mu_{xi} - \bar{x}_{..})b_.,$$

où

$$(\bar{x}_{..},\bar{y}_{..}) = \sum_{i=1}^{k} N^{-1}N_i(\bar{x}_{i.},\bar{y}_{i.}),$$

$$(\bar{x}_{i.},\bar{y}_{i.}) = n_i^{-1}\sum_{j=1}^{n_i}(X_{ij},Y_{ij}),$$

$$b_i = \left[\sum_{j=1}^{n_i}(X_{ij} - \bar{x}_{i.})^2\right]^{-1} \times \sum_{j=1}^{n_i}(X_{ij} - \bar{x}_{i.})(Y_{ij} - \bar{y}_{i.}),$$

$$b_. = \left[\sum_{i=1}^{k} N^{-1}N_i\, n_i^{-1}\sum_{j=1}^{n_i}(X_{ij} - \bar{x}_{i.})^2\right]^{-1} \times \sum_{i=1}^{k} N^{-1}N_i\, n_i^{-1}\sum_{j=1}^{n_i}(X_{ij} - \bar{x}_{i.})(Y_{ij} - \bar{y}_{i.}),$$

n_i est le nombre d'observations dans le domaine i, N_i est le nombre d'éléments de la population contenus dans le domaine i, μ_{xi} est la moyenne de population de X pour le domaine i et $\mu_{x.}$ est la moyenne de X pour l'ensemble de la population. Selon la terminologie utilisée par les auteurs, le premier estimateur est un estimateur par régression direct, le deuxième, un estimateur direct modifié et le troisième, un estimateur synthétique. Nous avons

$$\mathrm{EQM}\{\hat{\mu}_{(1)yi} \mid n_i\} = n_i^{-1}(1 + n_i^{-1})V\{Y_{\ell j} - \beta_\ell X_{\ell j} \mid \ell = i\} + O(n_i^{-2}),$$

$$\mathrm{EQM}\{\hat{\mu}_{(2)yi} \mid n_i\} = n_i^{-1}(1 + n^{-1})V\{Y_{\ell j} - \beta X_{\ell j} \mid \ell = i\} + O(n^{-2}),$$

[1] W.A. Fuller, Distinguished Professor, Statistical Laboratory and Department of Statistics, Iowa State University, Snedecor Hall, Ames, Iowa.

$$\mathrm{EQM}\{\hat{\mu}_{(3)yi} \mid n_i\} = (1 + n^{-1})$$

$$\times \sum_{i=1}^{k} N^{-2} N_i^2 n_i^{-1} V\{Y_{\ell j} - \beta X_{\ell j} \mid \ell = i\}$$

$$+ (\mu_{xi} - \mu_{x.})^2 V\{b_.\}$$

$$+ [\mu_{yi} - \mu_{y.} - \beta(\mu_{xi} - \mu_{x.})]^2 + O(n^{-2}),$$

où $V\{b_.\} = E\{(b_. - \beta)^2\}$, $V\{a_\ell \mid \ell = i\}$ est la variance de la variable a pour le domaine i,

$$\beta_i = [V\{X_{\ell j} \mid \ell = i\}]^{-1} C\{Y_{\ell j}, X_{\ell j} \mid \ell = i\}$$

et

$$\beta = \left[\sum_{i=1}^{k} N^{-1} N_i V\{X_{\ell j} \mid \ell = i\} \right]^{-1}$$

$$\times \sum_{i=1}^{k} N^{-1} N_i C\{Y_{\ell j}, X_{\ell j} \mid \ell = i\}.$$

L'estimateur $\hat{\mu}_{(1)yi}$ utilise uniquement l'information contenue dans l'échantillon de n_i observations. Par conséquent, toutes les propriétés de l'estimateur sont fonction de n_i et des paramètres du domaine. Le biais de régression est d'ordre n_i^{-1}, comme la variance. L'estimateur $\hat{\mu}_{(2)yi}$ utilise les moyennes de domaine, mais utilise aussi l'échantillon complet pour l'estimation du coefficient de régression. En conséquence, la variance fondamentale est aussi d'ordre n_i^{-1} et sera plus élevée que la variance fondamentale de $\hat{\mu}_{(1)yi}$ dans les cas où $\beta_i \neq \beta$. Or, la contribution de deuxième ordre à la variance est d'ordre $n_i^{-1}\, n^{-1}$ pour $\hat{\mu}_{(2)yi}$ et d'ordre n_i^{-2} pour $\hat{\mu}_{(1)yi}$. De plus, le biais de régression pour $\hat{\mu}_{(2)yi}$ est d'ordre n^{-1}. Si les domaines étaient des strates, on pourrait désigner $\hat{\mu}_{(1)yi}$ comme l'estimateur par régression simple et $\hat{\mu}_{(2)yi}$ comme l'estimateur par régression combiné.

L'estimateur $\hat{\mu}_{(3)yi}$ est un estimateur synthétique qui a une variance d'ordre n^{-1}, contrairement aux deux premiers estimateurs, qui ont une variance d'ordre n_i^{-1}. Cependant, cette réduction de la variance a un inconvénient: le biais est d'ordre un. On aura un biais nul seulement si la droite de régression est la même pour le domaine et pour la population entière.

On peut estimer la moyenne des erreurs quadratiques moyennes des trois estimateurs pour n'importe quel sous-groupe de petites régions. Si les n_i sont faibles, les variances estimées ne donneront qu'une information restreinte pour faire la différence entre les estimateurs. En outre, il n'existe qu'un degré de liberté pour le carré du biais pour un domaine particulier. Cependant, si l'écart pour un domaine devait être élevé par rapport à l'erreur-type, il faudrait reconsidérer l'estimateur synthétique.

Dans leur analyse de modèles, les auteurs soulignent l'importance de produire des estimateurs de la fiabilité des estimateurs pour petites régions. Ils laissent entendre que les principaux estimateurs de l'erreur quadratique moyenne pour des méthodes fondées sur un modèle sont des estimateurs de la valeur moyenne de l'erreur quadratique moyenne. Bien que ce soit vrai, il convient de mentionner que la méthode des "composantes de la variance" ne suppose pas que l'erreur quadratique moyenne est la même dans chaque domaine. En outre, dans une enquête classique, les estimateurs de l'erreur quadratique moyenne ne sont pas nécessairement les mêmes dans tous les domaines. Par exemple, un des termes de l'estimateur de l'erreur quadratique moyenne utilisé dans la méthode des "composantes de la variance" est l'estimateur de la variance de l'estimateur direct. La variance estimée de l'estimateur direct sera une fonction de l'effectif de l'échantillon contenu dans le domaine et pourra aussi être une fonction de la variance estimée (directement) de l'estimateur direct pour ce domaine. Voir Battese, Harter et Fuller (1988), Harville (1976), Prasad et Rao (1990) et Ghosh et Rao (1993).

Dans leur analyse de plans de sondage, les auteurs expliquent que la courbe de la fonction de variance est souvent "aplatie" dans le voisinage de la répartition optimale des unités entre les strates. Une légère redistribution de l'échantillon peut avoir pour effet d'accroître sensiblement l'efficience des estimations pour domaine sans réduire de façon notable l'efficience des estimations globales. Il en va de même de la combinaison d'un estimateur direct et d'un estimateur synthétique. Par conséquent, si quelqu'un a une idée assez juste de la composante de variance liée aux petites régions, soit grâce à une étude antérieure qui portait sur la même population ou grâce à une étude qui portait sur une population similaire, et s'il doit produire des estimations dans un délai très limité, il est raisonnable, dans les circonstances, d'attribuer des poids fixes pour former la combinaison linéaire. Il y aura vraisemblablement peu de perte d'efficience et les exigences de programmation pour la construction d'estimateurs seront beaucoup moins grandes. L'estimateur synthétique est un estimateur indiqué dans les circonstances, et beaucoup de praticiens l'utilisent.

Les auteurs évoquent la question de la propriété de fermeture liée à la construction d'estimateurs pour petites régions. Comme ils l'affirment, si on utilise une méthode dépendante des données comme la méthode des composantes de la variance, on obtient, pour chaque variable dépendante, des estimations qui n'ont pas la propriété de fermeture. Dans les circonstances, on peut toujours recourir à des méthodes multidimensionnelles. Voir, par exemple, Fuller et Harter (1987) et Fay (1987). Une autre méthode proposée par Fuller (1990) consiste à construire des estimateurs de composantes de la variance pour un sous-ensemble limité de variables, puis de se servir des estimations correspondantes comme variables de contrôle dans une méthode de régression. Celle-ci produit des poids pour les observations individuelles. Une fois que les poids sont déterminés, on peut construire un nombre indéfini de tableaux de données et toutes les estimations ont la propriété de fermeture.

J'estime que dans la plupart des cas d'estimation pour domaine, les gains d'efficience découlent surtout de l'utilisation judicieuse de l'information supplémentaire. Par conséquent, les efforts consacrés à l'obtention d'information supplémentaire de qualité sont justifiés. Si nous pouvons trouver une variable x qui soit étroitement corrélée avec la variable y, il n'y aura plus autant de variabilité à répartir entre la variance d'une région à l'autre et la variance d'échantillonnage.

REMERCIEMENTS

Je remercie Jay Breidt pour ses commentaires.

BIBLIOGRAPHIE

BATTESE, G.E., HARTER, R.M., et FULLER, W.A. (1988). An error components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83, 28-36.

FAY, R.E. (1987). Application of multivariate regression to small domain estimation. Dans *Small Area Statistics*. (Éds. R. Platek, J.N.K. Rao, C.-E. Särndal et M.P. Singh). New York: Wiley.

FULLER, W.A. (1990). Analyse d'enquêtes à passages répétés. *Techniques d'enquête*, 16, 177-190.

FULLER, W.A., et HARTER, R.M. (1987). The multivariate components of variance model for small area estimation. Dans *Small Area Statistics*. (Éds. R. Platek, J.N.K. Rao, C.-E. Särndal et M.P. Singh). New York: Wiley.

HARVILLE, D.A. (1976). Extension of the Gauss-Markov Theorem to include the estimation of random effects. *The Annals of Statistics*, 4, 384-395.

GHOSH, M., et RAO, J.N.K. (1993). Small area estimation: An appraisal. Manuscrit non-publié. Carleton University, Ottawa, Ontario, Canada.

PRASAD, N.G.N., et RAO, J.N.K. (1990). The estimation of mean squared errors of small-area estimators. *Journal of the American Statistical Association*, 85, 163-171.

COMMENTAIRES

GRAHAM KALTON[1]

Comme le mentionnent Singh, Gambino et Mantel (SGM), on cherche de plus en plus à réaliser des enquêtes qui vont produire des estimations pour des domaines de tous genres et de toutes tailles. Cette tendance est observée dans de nombreux pays dans le monde. D'une part, elle peut être simplement le résultat d'une évolution naturelle des exigences des analystes d'enquête, qui auparavant se contentaient d'estimations nationales et d'estimations touchant un petit nombre de domaines majeurs mais qui, aujourd'hui, veulent comparer des estimations relatives à de nombreux types de domaines différents. D'autre part, cette tendance vient des besoins des responsables de l'action gouvernementale, qui recherchent de l'information pour domaine afin d'analyser les effets des programmes courants sur différents domaines, de prévoir les conséquences d'une révision des politiques et de suivre l'exécution des programmes. Les données relatives à des divisions administratives (par ex., province ou État, comté, arrondissement scolaire) présentent un intérêt particulier pour l'élaboration de programmes (par ex., lorsqu'il s'agit de définir les régions à faible revenu aux fins de l'élaboration des programmes de soutien public).

Dans certains cas, on peut réussir à obtenir des estimations pour domaine suffisamment précises à l'aide des méthodes d'inférence fondée sur un plan qui sont utilisées habituellement dans l'analyse de données d'enquête. C'est le cas notamment pour les grands domaines, où la taille des échantillons est suffisamment grande pour qu'on ait le degré de précision voulu. Cela est possible aussi pour de petits domaines, pourvu qu'ils soient définis à l'avance et que le plan de sondage soit élaboré de manière à produire des échantillons de taille acceptable. Ainsi, aux États-Unis par exemple, la National Health and Nutrition Examination Survey et la Continuing Survey of Food Intakes by Individuals utilisent des fractions de sondage qui varient selon l'âge, le sexe ou l'origine ethnique dans le premier cas et selon l'âge et le sexe ou le niveau de faible revenu dans le second cas, afin de produire des échantillons acceptables pour les domaines créés par le recoupement de ces variables. La Current Population Survey des É.-U. utilise des fractions de sondage différentes selon les États afin de produire des estimations de l'emploi pour l'État. La faiblesse de cette méthode ressort clairement lorsqu'il y a un grand nombre de petits domaines, auquel cas la somme des effectifs de chaque domaine dans l'échantillon donne une taille d'échantillon global très élevée. On observe souvent cela pour les petites divisions administratives comme les comtés, les arrondissements scolaires et les territoires de bureaux de placement. Dans de telles circonstances, on peut devoir remplacer les méthodes classiques d'inférence fondée sur un plan par des méthodes à modèles, qui utilisent pour l'estimation un modèle statistique par lequel on renforce les estimations d'une petite région par des données recueillies dans l'enquête. On peut devoir recourir aussi aux méthodes fondées sur un modèle pour de petits domaines non planifiés, pour lesquels on n'avait pas prévu la nécessité d'un suréchantillonnage à l'étape de la conception.

Compte tenu de l'intérêt pour les estimations régionales, un nombre appréciable d'ouvrages sont parus sur les méthodes d'estimation pour petit domaine fondées sur un modèle. Cependant, peu d'ouvrages ont été écrits jusqu'à maintenant sur les questions que traite l'article de SGM, lesquelles méritent une plus grande attention. Comme les auteurs, je crois qu'il faut envisager la question de l'utilisation d'estimateurs pour petite région fondés sur un modèle avec circonspection. C'est pourquoi je suis heureux de voir que SGM examinent des méthodes qui permettraient d'établir des estimations régionales au moyen de l'inférence fondée sur un plan.

À mon point de vue, la première étape dans le calcul d'estimations régionales est de vérifier si on peut obtenir des estimations ayant un degré de précision acceptable avec des méthodes fondées sur un plan. Si les domaines sont définis à l'avance, on devrait envisager de concevoir le plan d'échantillonnage de manière à faciliter la production d'estimations régionales. Cela peut signifier de faire en sorte que les petites régions ne chevauchent pas de strates et qu'il y ait un échantillon de taille suffisamment grande pour chaque petite région. Une autre solution proposée par SGM consiste à réduire au maximum la formation de grappes. Plus la formation de grappes est limitée, moins la taille de l'échantillon dans chaque petite région est laissée au hasard. À cet égard, le principal avantage de la limitation de la formation de grappes est de permettre la production d'estimations pour de petites régions qui n'étaient pas définies à l'étape de la conception du plan de sondage. Lorsque de petites régions pour lesquelles des estimations sont prévues ne chevauchent pas des strates, la taille de l'échantillon dans chaque petite région devrait être contrôlée raisonnablement même lorsqu'il s'agit d'un échantillon en grappes (pourvu que les mesures de taille utilisées dans l'échantillonnage avec PPT soient raisonnables). Or, même dans le cas d'estimations prévues, il arrivera souvent qu'on se demande comment calculer les estimations de la variance pour une

[1] Graham Kalton, Westat, 1650 Research Blvd., Rockville, Maryland, U.S.A. 20850.

petite région étant donné un plan avec grappes, puisque les UPÉ tirées dans chaque petite région seront vraisemblablement peu nombreuses. Une estimation de la variance fondée sur les UPÉ d'une petite région sera donc imprécise, étant donné quelques degrés de liberté, et on préférera probablement une fonction de variance généralisée (en supposant, par exemple, que l'effet de plan au niveau national vaut pour chaque petite région). Autrement dit, même si l'estimation proprement dite est fondée sur un plan, l'estimation de la variance correspondante peut être une estimation indirecte, renforcée par les données d'autres régions. C'est pourquoi il est préférable d'avoir un plan de sondage qui prévoit le moins de grappes possible, même lorsqu'il s'agit de petits domaines planifiés. Toutefois, il est plus important de penser à modéliser les estimations proprement dites que de se concentrer sur la modélisation des variances.

Un des principes fondamentaux de l'approche fondée sur un plan est que l'information supplémentaire qui existe sur la population peut servir à l'étape de la conception du plan de sondage, à celle de l'analyse ou aux deux à la fois. Lorsqu'il existe de l'information sur des variables auxiliaires qui ont un rapport étroit avec la variable étudiée, on peut réaliser des gains appréciables en précision. L'utilisation d'information supplémentaire à l'étape de l'analyse, par l'intermédiaire de techniques comme la stratification a posteriori, l'estimation par quotient ou par régression et l'estimation de différences, a un attrait particulier en ce qui concerne l'estimation pour petite région. Il convient de souligner que les estimateurs par quotient et par régression peuvent trouver leur justification dans des hypothèses portant sur le modèle qui met en relation la variable étudiée (Y) et les variables auxiliaires (X), mais les estimateurs seront convergents selon le plan peu importe si le modèle convient ou non. L'utilisation d'un modèle approprié permet de réaliser les meilleurs gains en précision mais quel que soit le modèle choisi, les estimations sont approximativement non biaisées. On peut illustrer cela par un exemple simple, où la valeur des variables $X_1, X_2, \ldots, X_p$ est connue pour chaque élément de la population et la combinaison linéaire $\tilde{Y}_i = B_0 + B_1X_{1i} + \ldots + B_pX_{pi}$ est utilisée pour estimer Y_i, la valeur de la variable Y pour l'élément i de la population. Supposons pour des raisons de simplicité que les valeurs de B sont déterminées à l'aide de données externes, indépendantes de l'échantillon. Étant donné l'équation $Y_i = \tilde{Y}_i + e_i$, le total de domaine est $Y_a = \sum_{i\epsilon a}\tilde{Y}_i + \sum_{i\epsilon a} e_i = \tilde{Y}_a + E_a$. Comme $\tilde{Y}_a$ est connu, le problème revient à estimer E_a. À partir d'un échantillon d'éléments du domaine a, on peut estimer E_a au moyen de l'équation $\hat{E}_a = \sum_{j\epsilon s_a} e_j/\pi_j$, où π_j est la probabilité de sélection pour l'élément j de l'échantillon. L'estimateur $\hat{E}_a$ est non biaisé, peu importe si le modèle utilisé est valide ou non. En fait, la procédure est conçue de telle manière qu'au lieu d'estimer directement Y_a, on estime E_a et on ajoute à cette valeur une constante connue $\tilde{Y}_a$. Pour que la procédure soit efficace, la variance des e_i pour le domaine doit être inférieure à celle des Y_i. Il n'est pas nécessaire que $E_a = 0$. Le raisonnement est le même dans le cas plus habituel où les valeurs de B sont estimées au moyen de l'échantillon. Dans ces circonstances, l'estimation de Y_a est convergente selon le plan, peu importe le modèle utilisé (Särndal 1984). En outre, il est possible d'estimer les valeurs de B à l'aide des données de l'échantillon relatives au domaine étudié seulement, ce qui donne ce que SGM appellent un estimateur direct, ou à l'aide des données de l'échantillon global, ce qui donne un estimateur direct modifié. Un facteur déterminant dans le choix entre l'estimateur direct et l'estimateur direct modifié dans ces circonstances est de savoir si les valeurs de B pour l'échantillon global valent aussi pour le domaine. Si ce n'est pas le cas, il faut ajouter dans le modèle pour l'échantillon global des termes d'interaction entre X et les indicateurs de domaine. Avec un ensemble complet de termes d'interaction, l'estimateur direct modifié se ramène en fait à l'estimateur direct ordinaire.

On doit recourir à une approche dépendante d'un modèle lorsque les estimations fondées sur un plan ne sont pas assez précises, malgré qu'on ait utilisé le plus efficacement possible l'information supplémentaire qui était disponible. Il arrive en effet qu'on ne soit pas capable de calculer une estimation directe à cause de l'absence d'éléments de l'échantillon dans la petite région. Dans ces circonstances, on doit recourir à un modèle statistique afin d'''emprunter de l'information'' à d'autres régions. Ces modèles reposent sur des hypothèses (ex.: $E_a = 0$ dans l'exemple ci-dessus) et la qualité des estimations obtenues à l'aide de ces modèles dépend de la validité des hypothèses. Comme une hypothèse est toujours un peu inexacte, les estimations régionales sont entachées d'un biais, et comme il s'agit d'estimations indirectes biaisées, on se sert largement de l'erreur quadratique moyenne (EQM) de plan pour en mesurer la qualité (notons que EQM $= V' + B^2$ où V' est la variance et B est le biais de l'estimation).

La manière la plus courante de comparer la qualité d'une estimation directe avec celle d'une estimation indirecte est de comparer la variance (V) de la première avec l'EQM de la seconde. Or, en lisant l'article de SGM, j'en suis venu à me demander si l'EQM était la bonne mesure pour évaluer la qualité d'un estimateur indirect. Dans la pratique, il est possible d'estimer la variance V d'un estimateur direct, ce qui n'est pas le cas pour l'EQM de plan d'un estimateur indirect. Compte tenu de cette observation, on préférera nettement l'estimateur direct si $V =$ EQM. De fait, on optera le plus souvent pour l'estimateur direct si celui-ci offre un degré de précision acceptable, peu importe le rapport probable des valeurs V et EQM. Dans d'autres cas, si B est le biais prévu, on pourra préférer l'estimateur direct à l'estimateur indirect, à moins que $V > V' + kB^2$, où k est un facteur plus grand que 1 qui tient compte de la possibilité que le biais inconnu soit plus élevé que prévu.

On peut appliquer le même raisonnement aux estimateurs combinés (ou composites), qui sont la moyenne pondérée d'un estimateur direct et d'un estimateur indirect. On considère souvent que la meilleure manière de déterminer les poids est de minimiser l'erreur quadratique moyenne de l'estimateur combiné, ce qui donne des poids inversement proportionnels à V et à EQM, selon qu'il s'agit de

l'estimateur direct ou de l'estimateur indirect. Or, d'après le raisonnement ci-dessus, on pourrait tout aussi bien minimiser le poids de l'estimateur indirect, à la condition que l'estimateur combiné soit suffisamment précis. On pourrait aussi déterminer les poids selon une valeur probable maximum de l'EQM, plutôt que selon l'EQM prévue, afin de réduire la probabilité d'un biais important dans l'estimateur combiné.

Je ne comprends pas très bien l'explication des estimateurs tributaires de la taille d'échantillon définis par les équations (6.11) et (6.12) de SGM mais suivant certaines hypothèses, ces estimateurs pourraient cadrer dans l'argumentation présentée ci-dessus. Étant donné un plan d'échantillonnage avec probabilités égales et une valeur $\delta = 1$, ces estimateurs se ramènent à l'estimateur direct lorsque la taille effective de l'échantillon est égale ou supérieure à la taille prévue. Si on suppose que la taille prévue de l'échantillon implique un degré de précision acceptable pour une petite région, l'argumentation ci-dessus est respectée. Si la taille effective de l'échantillon est plus petite que prévu, l'estimateur tributaire de la taille d'échantillon devient la moyenne pondérée d'un estimateur direct et d'un estimateur indirect. Si on suppose que la taille prévue de l'échantillon représente l'effectif minimum nécessaire pour obtenir le degré de précision voulu, l'argumentation ci-dessus est de nouveau respectée. Si les observations précédentes constituent le fondement des estimateurs tributaires de la taille d'échantillon, il semblerait utile d'étendre ces estimateurs aux situations où la taille prévue de l'échantillon ne correspond pas à celle permettant tout juste d'atteindre le degré de précision voulue.

Comme je l'ai mentionné plus haut, l'information supplémentaire compte pour beaucoup dans la production d'estimations régionales précises. Cette information peut servir à accroître la précision des estimations fondées sur un plan ou elle peut être utilisée dans les modèles auxquels on recourt par l'approche dépendante de modèles. Idéalement, il faut de l'information supplémentaire qui soit étroitement liée aux variables étudiées dans l'enquête. En extrayant régulièrement de sources administratives ou autres des données auxiliaires récentes concernant les petites régions, on peut contribuer à améliorer les données régionales.

Bien que les auteurs évoquent la question plus générale des petits domaines, ils concentrent leur attention sur les petites régions. D'ailleurs, la question des petites régions est celle qui est le plus souvent traitée dans la littérature statistique et qui est en rapport avec l'application de méthodes d'estimation indirecte. Cela peut s'expliquer d'une part, par le fait que les domaines socio-économiques et les autres petits domaines pertinents (ex.: groupes d'âge-sexe) sont habituellement assez peu nombreux comparativement aux petites régions de sorte qu'on peut concevoir pour ces domaines un plan de sondage qui produise des estimations d'une précision acceptable pour chacun d'eux et d'autre part, par le fait que les domaines socio-économiques et démographiques sont souvent définis en fonction de la possibilité d'obtenir des estimations de plan d'une précision acceptable pour ces domaines (par ex., utiliser des groupes d'âge plus étendus pour certains domaines); toutefois, en ce qui concerne les domaines spatiaux, les territoires sont définis à l'avance et aucun regroupement n'est possible. Il en est ainsi d'une part, parce que les données auxiliaires sont en nombre insuffisant pour être utilisées dans les modèles statistiques s'appliquant aux domaines de ce genre et d'autre part, parce que l'analyse des domaines socio-économiques a souvent pour but d'établir des comparaisons entre les domaines. Or, ces comparaisons sont faussées si l'estimation relative à un domaine est "renforcée" par de l'information provenant d'autres domaines (voir, par exemple, Schaible 1992). Cela nous amène à dire qu'on ne devrait pas utiliser les estimations indirectes sans faire preuve d'esprit critique.

En conclusion, je suis d'accord avec l'approche générale présentée dans cet article. Lorsque c'est possible, les échantillons devraient être conçus de manière à produire des estimations directes pour petite région d'une précision acceptable, et les plans de sondage devraient être élaborés en ce sens. De même, lorsque c'est possible, on devrait se servir de données auxiliaires pour accroître la précision des estimations directes pour petite région. Lorsqu'il faut recourir à des estimations indirectes, on doit procéder avec prudence. Il faut élaborer soigneusement les modèles, chercher des estimateurs qui résistent bien aux défaillances d'un modèle et faire des analyses pour évaluer la pertinence des estimations indirectes. En l'absence de bons indices pour mesurer la qualité des estimations indirectes, il est essentiel de distinguer clairement ces estimations des estimations fondées sur un plan. Comme les estimations indirectes ne conviennent pas à tous les usages, l'utilisateur doit s'assurer que les estimations indirectes dont il veut se servir répondront à ses besoins.

BIBLIOGRAPHIE

SÄRNDAL, C.-E. (1984). Design-consistent versus model-dependent estimators for small domains. *Journal of the American Statistical Association*, 79, 624-631.

SCHAIBLE, W.L. (1992). Use of small area estimators in U.S. Federal programs. *Proceedings of the International Scientific Conference on Small Area Statistics and Survey Design*, (Vol. 1), 95-114, Central Statistical Office of Poland, Warsaw.

RÉPONSE DES AUTEURS

Nous remercions Wayne Fuller et Graham Kalton pour leurs commentaires stimulants qui complètent très bien les idées que nous avons exprimées dans l'article. Plusieurs de ces commentaires apportent des éclaircissements sur certains points et renforcent les arguments exposés. Encouragés par cet appui, nous aimerions approfondir certains points touchant le plan de sondage et, en même temps, répondre aux principaux commentaires des critiques.

Il est clair que le concepteur de plans de sondage s'efforce, malgré des contraintes opérationnelles, d'optimiser le plan dans le but de respecter les objectifs d'une enquête. Les grandes enquêtes poursuivent habituellement plusieurs objectifs, et il est fort probable que le concepteur a peu de prise sur l'établissement d'un ordre de priorité pour ces objectifs. Néanmoins, c'est justement à l'étape de l'établissement d'un ordre de priorité qu'il faut faire ressortir avec vigueur les besoins en données régionales, particulièrement en ce qui concerne les grandes enquêtes permanentes.

Dans les années soixante et soixante-dix, la plupart des pays ont mis l'accent sur les estimations infranationales (État, province) et modifié dans une certaine mesure les plans de sondage antérieurs qui, eux, visaient à optimiser les estimations nationales. Par exemple, on a utilisé des fractions de sondage différentes pour être sûr que les petits États ou les petites provinces contiendraient au moins tel nombre d'unités de l'échantillon. Vu la demande croissante de données au niveau infraprovincial ou infra-État (comté, district et municipalité), il faut modifier davantage les plans antérieurs, qui, eux, prévoient une répartition optimale de l'échantillon au niveau national. Cette modification peut consister à utiliser des fractions de sondage différentes pour les divisions administratives d'un État ou d'une province. Par exemple, si l'on souhaite produire des estimations infraprovinciales de qualité comparable, on attribuera sans doute à chaque province un échantillon qui sera à peu près proportionnel au nombre de régions infraprovinciales que contient la province. Cela n'équivaut peut-être pas à la taille relative de la population des provinces. Comme nous l'avons mentionné dans la section 5.4 de notre article, il faudrait privilégier une approche ascendante pour la répartition de l'échantillon. Les pertes enregistrées aux niveaux d'agrégation supérieurs, comme les gains aux niveaux inférieurs, varieraient selon les enquêtes, mais il est probable que, dans beaucoup de cas, une légère augmentation du coefficient de variation au niveau national se traduise par des gains appréciables au niveau des petites régions.

Kalton souligne l'importance de limiter la grosseur des grappes pour l'estimation de la variance; il est avantageux d'accroître le nombre de degrés de liberté en formant un grand nombre de petites grappes au lieu d'un petit nombre de grosses grappes. Précisons que l'agglomération d'unités d'échantillonnage présente un autre inconvénient du point de vue de l'estimation, en particulier de l'estimation pour petites régions; en effet, un plan qui prévoit la formation de grosses grappes produira des effets de plan prononcés, même pour de petits domaines planifiés. En règle générale, on recourt au plan avec formation de grappes pour réduire les coûts de l'enquête. Compte tenu de l'évolution actuelle des méthodes de collecte de données – par exemple l'usage moins fréquent de l'interview sur place au profit de l'interview assistée par ordinateur – il est devenu nécessaire de revoir périodiquement les modèles coût-variance qui conditionnent les décisions concernant la formation de grappes.

Une autre question que nous n'avons pas traitée dans l'article est celle de l'effet du renouvellement de l'échantillon dans les enquêtes permanentes. Pour une période donnée, il se peut que certains petits domaines n'aient pas un effectif suffisant pour qu'on puisse établir des estimations fiables. Cependant, à mesure que les unités de l'échantillon se succèdent par renouvellement, la taille cumulée ou effective de l'échantillon dans les domaines augmente et on peut ainsi calculer, par domaine, des estimations fiables, quoique biaisées chronologiquement. Par un choix judicieux de plans de renouvellement, les concepteurs d'enquêtes peuvent maximiser la taille cumulée de l'échantillon dans une période donnée. Par exemple, pour des estimations trimestrielles d'une enquête mensuelle, le plan de renouvellement optimal est $[1(2)]^k$, c'est-à-dire la répétition de k fois la séquence "un mois dans l'échantillon, deux mois hors de l'échantillon". Ce raisonnement s'apparente à celui de Leslie Kish sur le cumul d'échantillons dans le temps.

Kalton nuance la mise en garde faite sur l'emploi des estimateurs indirects en proposant une erreur quadratique moyenne pondérée, par laquelle on associe un poids plus grand que 1 au terme du biais pour tenir compte de ce que le biais de l'estimateur indirect peut être plus élevé que prévu. Deux raisons font que le biais peut être plus élevé que le laisse croire le modèle pour les effets de petite région: la variation aléatoire à l'intérieur du modèle et la défaillance du modèle. Il est utile de se rappeler ici la suggestion de Fay et Herriot (1979), qui proposent qu'il y ait un écart de moins d'une erreur type entre une estimation combinée et une estimation de plan; cette mesure permet de concevoir la possibilité que l'estimateur de modèle renferme un biais élevé quelqu'en soit la raison. Kalton affirme aussi, comme nous, qu'étant donné un estimateur direct de qualité acceptable, on pourra choisir en pratique d'utiliser cet estimateur même si l'erreur quadratique moyenne estimée correspondante est plus grande que celle d'estimateurs fondés sur un modèle. Compte tenu de la possibilité toujours réelle d'une défaillance du modèle, on favorisera cette approche du "mieux vaut prévenir que guérir", du moins tant qu'on n'aura pas expérimenté certains estimateurs indirects dans des conditions particulières. Et cela n'empêche pas qu'il y ait des situations où il faut faire preuve d'un peu d'audace.

Dans ses observations sur l'estimateur tributaire de la taille de l'échantillon, Kalton laisse entendre que l'attribution d'un poids nul à la composante synthétique peut poser un problème si la taille réelle de l'échantillon pour un petit domaine excède la taille prévue, car celle-ci peut n'être pas assez élevée pour qu'on puisse produire des estimations directes acceptables. Une solution possible serait d'utiliser

une valeur $n_{\min}$, soit l'effectif qui permet de produire des estimations directes tout juste acceptables. Notons toutefois que $n_{\min}$, tel qu'il est défini ici, est dépendant des caractéristiques.

Dans ses commentaires, Fuller décrit sommairement une méthode d'estimation pour petites régions qui exploite un modèle fondé sur les composantes de la variance et qui, pourtant, utilise des poids fixes pour garantir la cohérence interne des estimations relatives à diverses caractéristiques. Outre la cohérence interne des estimations régionales, il faut parfois que la somme des totaux estimés des petites régions qui forment une grande région corresponde à l'estimation directe officielle pour la grande région. Une façon d'assurer cette concordance est d'ajuster les estimations régionales à l'estimation directe pour la grande région au moyen, par exemple, d'un simple coefficient de correction; cependant, si les coefficients de correction dépendent des caractéristiques, la cohérence interne mentionnée précédemment n'existera plus. On peut garantir simultanément les deux types de cohérence à la condition que l'estimateur direct pour la grande région soit un estimateur par régression généralisé, $\hat{Y}_e + (X - \hat{X}_e)\hat{\beta}$, et qu'on utilise l'estimateur direct modifié $\hat{Y}_{sreg,a} = \hat{Y}_{e,a} + (X_a - \hat{X}_{e,a})\hat{\beta}$ (section 6.1 de notre article) pour les petites régions.

Comme le mentionne Fuller, il est possible d'estimer la valeur moyenne du biais au carré d'un estimateur pour un sous-ensemble quelconque de petites régions. Nous aimerions préciser de nouveau ici que la valeur moyenne du biais pour un ensemble de petites régions n'a pas de rapport direct avec une petite région en particulier. C'est pourquoi nous préférons utiliser, quand c'est possible, des estimateurs qui sont approximativement non biaisés selon le plan. Si nous n'avons d'autre choix que d'utiliser un estimateur fondé sur un modèle, nous devons nous efforcer de trouver des covariables pertinentes pour lesquelles il existe de l'information supplémentaire fiable afin de réduire au minimum le biais résiduel de l'estimateur.

Peut-être à cause des délais qu'implique inévitablement la production des données du recensement, aucun des deux critiques ne considère le recensement comme une source de données pour de petits domaines. À cet égard, il convient de mentionner qu'on envisage actuellement une formule qui prévoirait une sorte de grande enquête postcensitaire continue en remplacement du questionnaire détaillé utilisé dans le recensement décennal. Cette formule, appelée méthode des échantillons avec renouvellement complet, est décrite dans Kish (1990); une formule semblable, appelée méthode de dénombrement continu, est décrite dans Alexander (1994). Cette méthode comporte un certain nombre de fonctions qui méritent d'être examinées dans la mesure où elles peuvent devenir autant de moyens de produire à meilleur marché et dans de courts délais des données pour petits domaines.

Enfin, nous tenons à préciser qu'en soulignant l'importance de considérer l'estimation pour domaine à l'étape de la conception du plan de sondage, surtout en ce qui a trait aux domaines de taille moyenne, nous ne voulons aucunement minimiser le rôle des modèles dans l'estimation qui s'applique à de très petits domaines.

Nous espérons que l'approche générale présentée dans l'article, de même que les commentaires perspicaces des critiques, en particulier les conseils et les mises en garde formulés par Kalton dans le dernier paragraphe de ses commentaires, permettront aux concepteurs de plans de sondage et aux chercheurs de trouver les solutions qui conviennent pour les problèmes qui les intéressent.

SOURCES ADDITIONNELLES

ALEXANDER, C.H. (1994). A prototype continuous measurement system for the U.S. Census of Population and Housing. Document pour présentation à la réunion annuelle de la Population Association of America, Miami, Florida, 5 mai, 1994.

KISH, L. (1990). Recensement par étapes et échantillons avec renouvellement complet. *Techniques d'enquête*, 16, 67-77.