

Article

Plans d'échantillonnage novateurs : discussion de trois communications présentées au U.S. Census Bureau

par Jean Opsomer

Décembre 2011



Plans d'échantillonnage novateurs : discussion de trois communications présentées au *U.S. Census Bureau*

Jean Opsomer¹

1. Introduction

Outre son rôle dans la collecte des données du recensement décennal des États-Unis, le U.S. Census Bureau est l'un des plus grands organismes de collecte de données d'enquête au monde. Les deux outils statistiques qu'il utilise principalement pour concevoir ses enquêtes sont la stratification et l'échantillonnage à plusieurs degrés. Mis en œuvre avec succès durant les années 1940, ces outils ont continué d'être adaptés et perfectionnés depuis.

Bien que cette approche générale d'échantillonnage ait été très fructueuse, la hausse des coûts d'enquête, la diminution des taux de réponse et l'existence de nouveaux problèmes de couverture des bases de sondage (surtout dans le cas des enquêtes téléphoniques) suscitent de plus en plus d'inquiétudes. Parallèlement, les progrès en ce qui concerne les méthodes de collecte des données, les nouvelles sources de données et les outils informatiques permettent d'envisager des plans d'enquête qui n'auraient pas été possibles auparavant. Dans le cadre du programme de remaniement entrepris en 2010, le Census Bureau a demandé à des chercheurs universitaires éminents de donner leur avis sur des méthodes d'échantillonnage novatrices, en vue de commencer à explorer de nouvelles approches possibles de conception de ses enquêtes. Ainsi, les professeurs Steve Thompson (Simon Fraser University), Sharon Lohr (Arizona State University) et Yves Tillé (Université de Neuchâtel) ont été invités à donner des exposés d'ensemble sur certains plans d'échantillonnage qu'ils ont élaborés. J'ai été invité à offrir une discussion sur chacun de ces exposés.

Dans les trois sections qui suivent, je résumerai mes commentaires sur chacun des exposés. Mes objectifs, en formulant ces commentaires, étaient de mettre en relief les aspects les plus importants des méthodes d'échantillonnage présentées, de discuter de quelques possibilités importantes de les utiliser dans le contexte de l'échantillonnage des ménages et de cerner les difficultés éventuelles de mise en œuvre.

2. Sondage par réseaux, échantillonnage adaptatif et échantillonnage dans l'espace

L'exposé du professeur Thompson portait sur une catégorie générale de plans d'échantillonnage qui englobent

l'échantillonnage adaptatif en grappes, le sondage par réseaux et l'échantillonnage en ligne adaptatif. Dans la suite du présent exposé, à moins de faire référence à un plan particulier dans cette classe, je donnerai à ces plans le nom d'« échantillonnage adaptatif ». Un avantage important de l'échantillonnage adaptatif tient au fait qu'il intègre certaines caractéristiques des approches d'échantillonnage « de commodité », telles que l'échantillonnage boule de neige, y compris le fait de s'appuyer moins sur une base de sondage et la capacité de cibler l'échantillonnage sur des parties de la population présentant un intérêt particulier. Cependant, contrairement à l'échantillonnage de commodité, l'échantillonnage adaptatif demeure fermement fondé sur un plan d'échantillonnage, au sens qu'il permet l'estimation et l'inférence de la population finie selon la randomisation.

Dans les procédures d'échantillonnage adaptatif, un échantillon initial est tiré conformément à un plan d'échantillonnage probabiliste $p_0(s_0)$. En fonction des caractéristiques des éléments compris dans s_0 (par exemple, présence/absence des caractéristiques d'intérêt ou énumération des « liens » avec d'autres éléments de la population), un échantillon de suivi s_1 est sélectionné parmi la population restante, en utilisant un plan de sondage conditionnel $p_1(s_1 | s_0)$. Ce processus est répété pour des échantillons additionnels successifs s_2, s_3, \dots jusqu'à ce que soit satisfait un critère cible, tel que la taille totale de l'échantillon ou le nombre de « vagues » d'échantillonnage. L'échantillon final correspond à l'union de tous les échantillons successifs. La façon détaillée dont les échantillons successifs sont tirés varie selon le plan d'échantillonnage adaptatif. La section 2.2 de l'article de Thompson publié dans le présent numéro et Thompson (2006) contiennent d'autres renseignements sur l'échantillonnage en ligne adaptatif, un type très souple d'échantillonnage adaptatif qui englobe un grand nombre des autres plans de sondage en tant que cas particuliers.

Comme les plans d'échantillonnage pour chacune des vagues d'échantillonnage sont des plans probabilistes, il est possible d'obtenir des estimateurs valides sous le plan. Un estimateur simple de la moyenne de population finie $\mu_N = N^{-1} \sum_U y_i$ est construit de la façon suivante. Partant du plan de sondage initial p_0 avec les probabilités d'inclusion connexes π_{0i} , un estimateur sans biais de la moyenne de population est donné par $\hat{\mu}_0 = N^{-1} \sum_{s_0} y_i / \pi_{0i}$. Pour chacune des vagues subséquentes d'échantillonnage $k = 1, \dots, K$,

1. Jean Opsomer, Département de la statistique, Université de l'État du Colorado, Fort Collins, CO 80523-1877. Courriel : jopsomer@stat.colostate.edu.

un estimateur sans biais de μ_N est donné par $z_k = \sum_{s_{k-1}} y_i + \sum_{s_k} y_i / q_{ki}$, où les q_{ki} sont les probabilités d'inclusion conditionnelles pour la vague k (voir Thompson (2006) pour des précisions sur la construction des q_{ki} , et la section 2.4 de l'article de Thompson dans le présent numéro pour des exemples spécifiques). En posant que $\hat{\mu}_r = K^{-1} \sum_{k=1}^K z_k$, un estimateur sans biais de μ_N s'obtient sous la forme $\hat{\mu} = w\hat{\mu}_0 + (1-w)\hat{\mu}_r$, qui est une combinaison linéaire de l'estimateur initial et de la moyenne des estimateurs subséquents.

L'estimateur $\hat{\mu}$ est sans biais sous le plan, mais il dépend de l'ordre des vagues de sélection de l'échantillon. Un estimateur plus précis peut être obtenu en calculant la moyenne sur les divers ordres dans lesquels un échantillon aurait pu être obtenu. Pour les petites tailles d'échantillon, une expression explicite existe pour cet estimateur plus efficace, mais en général, il doit être approximé par échantillonnage répété à partir d'une chaîne de Markov définie de manière appropriée, puis par calcul de la moyenne des échantillons. Les méthodes exactes d'établissement de la chaîne et de tirage des échantillons sont décrites dans Thompson (2006), qui discute également de l'estimation de la variance de l'estimateur résultant.

L'un des principaux avantages des plans d'échantillonnage adaptatif tient au fait qu'ils permettent à l'organisme chargé des enquêtes de concentrer l'échantillon sur les parties de la population présentant un intérêt. Cela est particulièrement utile dans les situations où certains éléments d'intérêt sont relativement rares et qu'ils ne peuvent pas être identifiés a priori dans une base de sondage. Les enquêtes sur la chasse et la pêche, les nouveaux immigrants, les enfants scolarisés à domicile et les propriétaires d'entreprises familiales en sont des exemples. Dans chacun de ces cas, les éléments sont assez « dispersés » dans la population et aucune base de sondage complète n'est généralement disponible. Cependant, il est probable que les individus qui font partie de cette population seront capables de fournir des renseignements sur d'autres individus, de sorte que des liens peuvent être identifiés et échantillonnés au cours de différentes vagues d'échantillonnage adaptatif. Notons que l'échantillonnage adaptatif peut également être utilisé quand ces types d'éléments rares font partie d'une sous-population d'intérêt dans une enquête auprès d'une population plus grande et non rare. Par exemple, une enquête sur les écoliers pourrait inclure une strate d'enfants scolarisés à domicile.

Rejoindre des (sous-)populations relativement rares est un défi fréquent dans les enquêtes, et un certain nombre de méthodes sont régulièrement déployées pour résoudre ce problème. Dans le contexte des enquêtes-ménages, le plan d'échantillonnage sans doute le plus fréquent est l'échantillonnage stratifié à plusieurs degrés. Dans la mesure où de l'information auxiliaire pertinente au niveau de l'UPE est disponible, l'organisme chargé de l'enquête peut

suréchantillonner les UPE que l'on s'attend à contenir une fraction importante des groupes d'intérêt. Une enquête sur les hommes afro-américains courant le risque d'avoir la maladie de Parkinson dans laquelle pourraient être suréchantillonnés les secteurs de recensement comptant une fraction élevée de la population afro-américaine en est un exemple. Un autre plan d'échantillonnage qui peut être utile dans ce contexte est l'échantillonnage à plusieurs phases. Dans ce cas, la première phase d'échantillonnage est utilisée comme échantillon de sélection ou comme un moyen de recueillir de l'information auxiliaire pertinente, tandis que les phases subséquentes visent à obtenir les données d'enquête d'intérêt. L'Agricultural Resource Management Survey (ARMS) (menée par le USDA) suit ce genre de plan. Un échantillon de toutes les fermes est sélectionné à la phase 1, en vue de recueillir des données sur les caractéristiques des fermes pour l'année de référence de l'enquête. Durant les phases ultérieures, on procède à la sélection d'échantillons ciblés en se basant sur les produits d'intérêt (par exemple, produits laitiers, blé, etc.). Une troisième approche d'échantillonnage parfois utile pour obtenir des échantillons de (sous-)populations rares est l'échantillonnage à bases multiples. Le principe qui sous-tend l'échantillonnage à bases multiples est la combinaison de plusieurs bases de sondage ayant différentes caractéristiques de couverture, par exemple une « bonne » base de sondage contenant une grande proportion des éléments d'intérêt, mais pouvant éventuellement être incomplète, et une « mauvaise » base de sondage qui est complète, mais ne contient qu'une faible proportion des éléments d'intérêt. Par exemple, une enquête auprès des sociétés d'une industrie particulière pourrait être réalisée en se servant d'une liste des membres de groupes d'industries comme « bonne » base de sondage et d'une liste générale des sociétés comme « mauvaise » base de sondage. Pour un examen plus approfondi de l'échantillonnage à bases de sondage multiples, voir la section 3 qui suit.

Comparativement à ces trois plans d'échantillonnage, l'échantillonnage adaptatif est plus souple et permet un contrôle plus fin du nombre et des caractéristiques des éléments qui sont inclus dans l'échantillon, ce qui, souvent, accroît l'efficacité et/ou réduit le coût. Un inconvénient de l'échantillonnage adaptatif est qu'il faut recueillir l'information sur les liens entre les éléments, ce qui peut augmenter le fardeau de réponse et le coût de la collecte, et éventuellement poser des problèmes de confidentialité.

Comme l'échantillonnage adaptatif s'appuie fréquemment sur des « liens » entre éléments afin de définir les probabilités de sélection conditionnelles dans les vagues d'échantillonnage, il convient aussi particulièrement bien pour les enquêtes qui visent à étudier les liens entre les

éléments d'une population. Il pourrait s'agir, par exemple, d'enquêtes portant sur les transactions ou les relations entre entreprises, d'enquêtes sur le comportement de troc/d'échanges des ménages, ou d'enquêtes sur les relations ou les caractéristiques des réseaux familiaux.

Un organisme d'enquête qui envisage d'adopter l'échantillonnage adaptatif doit prendre en considération un certain nombre de questions concernant l'estimation et la diffusion des données. Dans de nombreux cas, les données d'enquête sont diffusées sous forme d'un ensemble de données pondérées et les estimations de variance sont fournies sous forme d'une description simplifiée du plan de sondage (par exemple, strates et UPE), de poids de rééchantillonnage ou de fonctions généralisées de variance. Il est également très fréquent que les poids soient calés et/ou ajustés pour tenir compte de la non-réponse. Les estimateurs pour les plans adaptatifs peuvent effectivement être exprimés sous forme de somme d'échantillons pondérés, de sorte qu'un ensemble de données pondérées pourrait être facilement créé même pour la version chaîne de Markov des estimateurs susmentionnés. Le choix du meilleur moyen de fournir les estimations de variance avec l'ensemble de données est une question qu'il convient encore d'étudier et qui pourrait dépendre des particularités de l'enquête. De même, la façon d'intégrer le calage et les corrections pour la non-réponse dans l'estimation sous échantillonnage adaptatif est un domaine où les travaux de recherche doivent se poursuivre.

3. Échantillonnage avec bases de sondage multiples chevauchantes

La professeure Lohr a donné un aperçu complet des plans d'échantillonnage généraux et des méthodes d'estimation quand l'échantillonnage repose sur plusieurs bases de sondage. Les approches classiques de réalisation des enquêtes sont de plus en plus souvent remises en question aujourd'hui, parce que les coûts augmentent, que les niveaux de réponse diminuent pour les modes de collecte classiques et que les préoccupations se multiplient quant à la couverture incomplète des bases de sondage existantes (par exemple, numéros de téléphone fixe rejoints par la méthode de composition aléatoire). En tirant des échantillons de plusieurs bases de sondage au lieu d'une seule, il est possible de réduire les coûts de l'enquête, d'améliorer la couverture de l'échantillon global et même, éventuellement, d'accroître les taux de réponse selon l'enquête particulière qui est réalisée (par exemple, grâce à de meilleurs renseignements d'identification des répondants dans l'une des bases de sondage).

L'échantillonnage fondé sur plusieurs bases de sondage représente une approche de tirage d'échantillons purement

fondée sur la randomisation et l'échantillonnage dans les bases de sondage individuelles se fait selon la même méthodologie que l'échantillonnage « classique » dans une seule base de sondage. Des méthodes d'estimation entièrement fondées sur le plan de sondage existent pour l'échantillonnage à bases de sondage multiples et plusieurs d'entre elles peuvent être déployées facilement dans le contexte des enquêtes à grande échelle dans lesquelles un ensemble de données pondérées est le principal produit (voir plus bas). La caractéristique principale de toutes les méthodes d'estimation est l'estimation du chevauchement des bases de sondage, qui est habituellement inconnu, mais qui doit être pris en compte. Pour l'estimer, on construit, pour chaque base de sondage, des estimateurs fondés sur le plan pour la ou les sous-populations d'éléments qui se retrouvent aussi dans la ou les autres bases de sondage. Les estimateurs des caractéristiques de la ou des intersections entre les bases de sondage doivent alors être combinés sur l'ensemble des bases de sondage. Les méthodes existantes diffèrent quant à la façon de combiner ces estimateurs, les plus simples utilisant des moyennes pondérées par la taille d'échantillon et les plus complexes, des estimateurs pondérés par des estimations de la précision des estimateurs individuels.

L'échantillonnage à partir de bases de sondage multiples s'applique particulièrement aux cas pour lesquels il n'existe aucune base de sondage unique couvrant l'ensemble de la population. Des exemples types de ce genre de situations sont l'échantillonnage par composition aléatoire, où une fraction croissante de la population ne peut pas être rejointe au moyen d'un numéro de téléphone fixe, les enquêtes auprès de professionnels ou d'entreprises pour lesquelles des listes partielles peuvent être obtenues auprès de fournisseurs ou d'organismes professionnels. Les enquêtes sur des populations rares qui existent au sein de la population plus générale sont d'autres situations dans lesquelles l'échantillonnage à bases de sondage multiples pourrait être appliqué. Une base de sondage globale existe pour la population, mais la sélection des répondants pour savoir s'ils appartiennent à la sous-population d'intérêt est longue et coûteuse. Une autre base de sondage contenant une proportion nettement plus élevée d'éléments de la sous-population d'intérêt est parfois disponible, mais si la couverture de cette base de sondage est incomplète, l'organisme chargé de l'enquête pourrait ne pas vouloir s'en servir de crainte de ne pas obtenir un échantillon valide. La combinaison de cette base de sondage de rechange, mais incomplète, de la sous-population avec la base de sondage complète, mais inefficace, de l'ensemble de la population pourrait être à la fois rentable et défendable du point de vue statistique. Les enquêtes sur la chasse et la pêche, pour lesquelles il existe souvent une liste des permis octroyés qui peut être incomplète ou non à jour, sont des exemples

d'enquêtes auprès de ce genre de sous-populations. L'approche à bases de sondage multiples pourrait également être utile pour une enquête auprès de la population générale comme moyen d'accroître la taille de l'échantillon dans certaines sous-populations présentant un intérêt particulier. Par exemple, dans une enquête générale sur les fermes, on pourrait souhaiter produire des estimations pour les fermes de type biologique, qui ne représentent qu'une faible fraction des fermes, mais dont bon nombres figurent dans les répertoires d'entreprises de type biologique. La section 1 de l'article de Lohr publié dans le présent numéro donne plusieurs autres exemples du vaste champ d'application des enquêtes à bases de sondage multiples.

Comme je l'ai mentionné plus haut, les méthodes d'estimation comprennent la construction d'estimateurs pour la sous-population contenue dans l'intersection des bases de sondage, ce qui requiert le choix d'une méthode de pondération pour les estimateurs obtenus d'après les différentes bases de sondage. Les méthodes de pondération qui s'appuient sur l'estimation de la précision de ces estimateurs pourraient être privilégiées dans une perspective d'efficacité. Cependant, leur mise en œuvre en pratique pose quelques problèmes, parce que les poids résultants peuvent varier pour diverses variables de l'enquête. Des approches plus commodes consistent à renoncer à une certaine efficacité afin de pouvoir utiliser les mêmes poids pour toutes les variables de l'enquête, une caractéristique soulignée à plusieurs reprises dans l'article de Lohr publié dans le présent numéro. La méthode du *pseudo-maximum de vraisemblance* (PMV) de Skinner et Rao (1996), qui produit un ensemble unique de poids, est recommandée par Lohr comme méthode privilégiée pour les enquêtes uniques, tandis qu'une approche à poids fixe plus simple est préférable pour les enquêtes longitudinales.

Bien que la méthodologie de base pour la construction d'estimateurs fondés sur le plan pour l'échantillonnage à bases de sondage multiples soit établie aujourd'hui, il est nécessaire de poursuivre l'étude d'approches pour appliquer le calage et la correction de la non-réponse dans ce contexte. Comme il est possible d'appliquer ces corrections au niveau de la base de sondage individuelle, au niveau de la population ou aux deux niveaux (selon l'information auxiliaire disponible), une étude des propriétés des estimateurs sous ces divers scénarios serait fort utile et devrait servir à élaborer des lignes directrices à l'intention des praticiens des sondages. La section 3 de l'article de Lohr dans le présent numéro offre une discussion de quelques-uns des premiers résultats dans ce domaine.

La section 4.2 de l'article de Lohr est consacrée à l'examen des méthodes d'estimation de la variance des estimateurs pour bases de sondage multiples qui ont été élaborées, y compris les méthodes de linéarisation et les

méthodes de rééchantillonnage. Dans le cas de l'approche par linéarisation, une question pratique importante tient au fait qu'il faut pouvoir déterminer à quelle base de sondage appartient chaque élément de l'échantillon, car la variance doit être estimée séparément dans chaque base de sondage. L'organisme d'enquête qui produit les données pourrait juger cette situation indésirable, pour des raisons de confidentialité des données. Dans le cas des méthodes de rééchantillonnage, telles que le *jackknife* et le *bootstrap*, l'organisme d'enquête peut créer des ensembles de poids de rééchantillonnage qui ne requièrent pas que l'on divulgue aux utilisateurs des données à quelle base de sondage appartiennent les divers éléments de l'échantillon. Lohr (2007) recommande l'approche du *bootstrap combiné* pour l'inférence sous échantillonnage à bases de sondage multiples. La méthode du *jackknife groupé* de Kott (2001) pourrait également être considérée comme une autre solution.

La mise en œuvre d'enquêtes à bases de sondage multiples est parfois plus difficile que celle d'enquêtes à base de sondage unique. Comme il est mentionné à la section 5 de l'article de Lohr, il faut être conscient du risque accru d'erreurs non dues à l'échantillonnage, surtout si les modes ou protocoles de collecte des données varient d'une base de sondage à l'autre. Par exemple, les éléments échantillonnés dans une base de sondage reçoivent une lettre de présentation envoyée à l'avance, tandis que ceux d'une autre base de sondage reçoivent un « appel direct » à cause du manque de renseignements sur l'adresse. Il se peut aussi que les caractéristiques de la non-réponse diffèrent selon la base de sondage, de sorte que des corrections distinctes sont nécessaires. Enfin, dans de nombreux cas, les éléments présents dans les diverses bases de sondage pourraient avoir des caractéristiques différentes (par exemple, fermes biologiques membres d'une association nationale d'entreprises de type biologique vs celles qui n'en sont pas membres). Dans tous ces cas, il convient de faire attention aux effets propres à la base de sondage et de construire prudemment les pondérations afin d'obtenir des estimateurs pour données d'enquête valides. Par ailleurs, l'existence de multiples bases de sondage offre la possibilité de mesurer les erreurs non dues à l'échantillonnage, parce qu'elles fournissent des échantillons multiples d'une même population. Par exemple, il pourrait être utile d'effectuer des « appels directs » auprès d'une partie des éléments sélectionnés dans la base de sondage contenant des adresses pour évaluer les effets de mode.

4. Échantillonnage équilibré par la méthode du cube

Dans son exposé, le professeur Tillé a énoncé les fondements de l'échantillonnage équilibré et décrit la *méthode du*

cube, qu'il a élaborée en tant qu'algorithme pratique pour le tirage d'échantillons équilibrés. Les objectifs des plans d'échantillonnage équilibré consistent à maintenir la représentation de la structure de la population dans l'échantillon (d'où le terme « équilibré ») et à améliorer l'efficacité des estimateurs d'après des données d'enquête. Aujourd'hui, la stratification est le principal outil qu'utilisent la plupart des statisticiens d'enquête en vue de réaliser ces deux objectifs. La stratification permet d'atteindre l'équilibre en forçant la composition de l'échantillon à concorder avec la répartition entre les strates et accroît l'efficacité des estimateurs en éliminant la composante de la variance due aux différences entre strates. L'échantillonnage systématique est également utilisé pour atteindre ces objectifs, le plus souvent dans le contexte des enquêtes sur les ressources naturelles. Dans ce cas, la composition de l'échantillon correspond exactement à la composition de la population pour les variables de tri et y correspond approximativement pour toute variable corrélée à la variable de tri. Un gain d'efficacité est réalisé parce que les moments des variables d'intérêt dans l'échantillon concordent (approximativement) avec les moments dans la population. Bien que les deux approches soient utilisées à grande échelle et donnent de bons résultats, elles manquent de souplesse. La stratification requiert souvent de diviser la population en « cellules » définies par l'intersection des variables de stratification, ce qui peut donner lieu à une prolifération de petites cellules correspondant à de petites tailles d'échantillon. L'échantillonnage systématique est une forme hautement contrainte d'échantillonnage qui offre une souplesse limitée en ce qui concerne la construction de l'échantillon et qui pose aussi le problème de l'absence d'un estimateur de variance fondé sur le plan de sondage.

L'échantillonnage équilibré peut être considéré comme une généralisation de la stratification. Sous cette interprétation, les échantillons stratifiés sont tirés avec des probabilités d'inclusion données pour tous les éléments de la population, mais sous la contrainte de la taille d'échantillon dans chaque strate. Dans l'échantillonnage équilibré, les contraintes de stratification sont remplacées par des contraintes de la forme $\sum_s x_i / \pi_i = \sum_U x_i$, où x_i est un vecteur de *variables d'équilibrage*. Quand les x_i sont des indicateurs de strate, l'échantillonnage équilibré coïncide avec la stratification, mais toute variable catégorique ou continue (ou une combinaison de celles-ci) peut être utilisée, ce qui donne une grande souplesse pour la construction de l'échantillon.

Comme il est mentionné plus haut, la méthode du cube est un algorithme qui permet de tirer des échantillons équilibrés étant donné un ensemble de probabilités d'inclusion et de contraintes. Si des échantillons exactement équilibrés existent dans la population, l'algorithme essaiera de sélectionner l'un d'eux. Si aucun échantillon ayant les

probabilités d'inclusion postulées et satisfaisant exactement les contraintes d'équilibrage ne peut être trouvé, l'algorithme essaiera de trouver une solution qui satisfait d'aussi près que possible les contraintes. La méthode du cube requiert que les variables d'équilibrage x_i soient connues pour tous les éléments de la population. Selon le contexte de l'enquête, cette exigence pourrait représenter une limite importante de l'applicabilité de l'échantillonnage équilibré.

Même si l'équilibrage sur les variables auxiliaires au niveau de la population est effectué à l'étape de l'élaboration du plan, il paraît probable qu'en pratique, le calage et d'autres corrections de la pondération, telles que celles de la non-réponse, soit souvent requis. En fait, selon Tillé, la combinaison de l'équilibrage et du calage constitue la stratégie la plus efficace (voir la section 7.4 de l'article de Tillé dans le présent numéro). Toutefois, les propriétés théoriques des estimateurs qui sont à la fois équilibrés et calés n'ont pas encore été établies complètement.

Bien que l'échantillonnage équilibré permette de maintenir les probabilités d'inclusion des éléments dans la population, il est clair que l'existence de contraintes d'équilibrage affecte les probabilités d'inclusion *conjointes* et donc la variance des estimateurs. Ce sujet est abordé à la section 6 de l'article de Tillé. Deville et Tillé (2005) ont montré que, dans certaines conditions, la variance des estimateurs sous échantillonnage équilibré peut être approximée par une variance de type linéarisation, qui dépend des résidus d'une régression linéaire des variables étudiées sur les variables d'équilibrage. Quoiqu'il s'agisse d'un résultat important et utile, il ne mène pas à une approche d'estimation de la variance convenant à toutes les applications d'enquête. L'un des problèmes est que l'estimation de la variance fondée sur cette méthode requiert l'accès aux variables d'équilibrage pour tous les participants à l'enquête et que ces variables pourraient ne pas être diffusées publiquement dans l'ensemble de données d'enquête. Dans ce contexte, une méthode de rééchantillonnage pourrait être particulièrement séduisante, parce qu'elle ne nécessiterait pas la diffusion de ces variables. Cependant, aucune méthode de ce genre n'est disponible à l'heure actuelle.

L'échantillonnage équilibré présente des liens étroits avec l'*échantillonnage réjectif*, dont les objectifs sont les mêmes. Dans l'échantillonnage réjectif, un échantillon est tiré avec des probabilités d'inclusion préspecifiées, et l'échantillon est accepté ou rejeté selon qu'il est compris ou non dans une fourchette de tolérance donnée d'une contrainte d'équilibrage. Si l'échantillon est rejeté, la procédure est répétée jusqu'à ce que soit trouvé un échantillon qui se trouve dans la fourchette de tolérance. L'échantillonnage réjectif existe de longue date, mais Fuller (2009) a décrit une certaine théorie asymptotique montrant qu'asymptotiquement, sa version de l'échantillonnage réjectif était

approximativement équivalente à l'échantillonnage équilibré.

5. Conclusion

Les méthodes décrites dans les trois exposés sont remarquablement complémentaires. Les plans d'échantillonnage adaptatif permettent d'obtenir des échantillons aléatoires, statistiquement valides, pour des populations habituellement difficiles à échantillonner efficacement. Très peu d'information est nécessaire pour tirer ce genre d'échantillon, mais de nombreux efforts doivent être faits durant la collecte des données afin de découvrir et de suivre les « liens » entre les éléments de la population et de tirer des échantillons successifs. En revanche, l'échantillonnage équilibré est utile quand des renseignements très détaillés sont disponibles dans la base de sondage et, dans cette situation, il permet d'obtenir des plans d'échantillonnage hautement personnalisés et efficaces. Une fois qu'un échantillon équilibré est tiré, la collecte des données peut se poursuivre de la même façon que dans les enquêtes habituelles. L'échantillonnage à bases de sondage multiples couvre un cas intermédiaire, en ce sens qu'aucune bonne base de sondage unique n'existe, mais que plusieurs bases de sondage partielles sont utilisées pour « compenser » leur faiblesses réciproques. Des échantillons distincts sont tirés de chaque base de sondage et la collecte des données se fait comme d'habitude, sauf qu'il est nécessaire de déterminer à

quelle(s) base(s) de sondage chaque unité échantillonnée appartient.

Conjuguées aux approches existantes déjà mises en œuvre, ces trois nouvelles méthodes d'échantillonnage pourraient accroître fortement la souplesse avec laquelle des échantillons peuvent être adaptés à des applications particulières, afin de réduire les coûts d'enquête et d'augmenter la précision des estimateurs.

Bibliographie

- Deville, J.-C., et Tillé, Y. (2005). Variance approximation under balanced sampling. *Journal of Statistical Planning and Inference*, 2, 569-591.
- Fuller, W.A. (2009). Some design properties of a rejective sampling procedure. *Biometrika*, 96(4), 933-944.
- Kott, P.S. (2001). The delete-a-group jackknife. *Journal of Official Statistics*, 17, 521-526.
- Lohr, S. (2007). Recent developments in multiple frame surveys. Dans *ASA Proceedings of the Joint Statistical Meetings*, American Statistical Association, 3257-3264.
- Skinner, C.J., et Rao, J.N.K. (1996). Estimation in dual frame surveys with complex designs. *Journal of the American Statistical Association*, 91, 349-356.
- Thompson, S.K. (2006). Adaptive web sampling. *Biometrics*, 62, 1224-1234.