N° 11-522-XIF au catalogue

La série des symposiums internationaux de Statistique Canada - Recueil

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

2003

Statistique Canada Statistics Canada

Canada

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

UTILISATION DE DONNÉES AUXILIAIRES POUR DÉTERMINER LE PLAN D'ÉCHANTILLONNAGE DANS UNE ENQUÊTE COMPORTANT PLUSIEURS VARIABLES CLÉS

Anders Holmberg[1]

RÉSUMÉ

Aux stades de la conception et de l'estimation d'une enquête, les grands organismes recourent souvent à des données auxiliaires. Les progrès techniques de la collecte de données et la plus grande accessibilité des registres permettent une utilisation accrue et plus efficiente de tels renseignements. Dans le présent document, nous verrons comment faire un usage efficient de données auxiliaires en échantillonnage portant sur des populations finies. Comme la puissance statistique des variables auxiliaires peut varier selon les paramètres visés et puisqu'un plan d'échantillonnage qui est optimal pour un paramètre ne l'est pas nécessairement pour un autre, il faut concevoir des plans d'échantillonnage représentant un bon compromis.

Nous étenderons les résultats antérieurs au sujet du choix d'un plan d'échantillonnage optimal au cas de plusieurs variables d'intérêt. En minimisant différentes mesures sommaires au stade de la planification, nous proposons une démarche permettant de trouver un plan d'échantillonnage de « bon compromis » qui se révélera d'une grande efficience dans l'ensemble. Nous aborderons la question des critères de précision qui varient selon les estimations en appliquant un algorithme de programmation non linéaire. Notre méthode produit un ensemble de probabilités d'inclusion du premier ordre inégales applicable à des scénarios d'échantillonnage d'une taille tant fixe qu'aléatoire. Un recours souple à des variables auxiliaires dans un plan d'échantillonnage est alors possible. En guise d'exemple, nous évoquons une application à une population d'entreprises suédoises où nous comparons les programmes à variable unique aux autres méthodes qu'emploient les statisticiens d'enquêtes pour résoudre le problème de l'inclusion de plusieurs variables en échantillonnage.

MOTS CLÉS : Enquêtes à plusieurs variables, plan d'échantillonnage optimal, planification des enquêtes, variables auxiliaires.

1. INTRODUCTION

1.1 Observations préliminaires

L'objectif habituel d'une enquête est l'estimation d'un grand nombre de paramètres. Pour le statisticien d'enquête, un grand but au stade de la planification d'une enquête est de trouver une stratégie, c'est-à-dire une combinaison de plan d'échantillonnage et d'estimateurs pour les paramètres les plus importants, qui produit les meilleures estimations de paramètres possible (les plus précises) à un coût donné; en d'autres termes, il s'agit d'obtenir des estimations avec la précision voulue au meilleur coût possible. La théorie d'échantillonnage met des instruments à notre disposition dans un tel but. Nous illustrerons et comparerons les méthodes auxquelles peuvent avoir recours les statisticiens d'enquête au stade de la planification pour régler les problèmes de variables multiples.

Dans sa quête d'estimateurs de haute précision, la théorie de l'échantillonnage met fortement l'accent sur l'utilisation de données auxiliaires. Comme des renseignements auxiliaires sont fréquemment disponibles dans les grands organismes d'enquête, le statisticien peut les mettre à profit et se doter par là d'une stratégie d'une grande efficience. Souvent, la base de sondage comporte au moins quelques variables auxiliaires qui peuvent être utiles. Les spécialistes du domaine traité et les statisticiens d'enquête peuvent connaître les relations entre les variables étudiées et les variables auxiliaires provenant d'enquêtes de même nature menées auprès de populations semblables. Si de telles données auxiliaires ne sont pas disponibles, il sera bon dans bien des cas de réaliser une enquête pilote. Dans le présent document, nous entendrons par « données auxiliaires » les seuls renseignements répondant à la double définition suivante : (i) données relatives à un ensemble de variables (que nous appellerons variables auxiliaires) qui

[1] Anders Holmberg, Statistics Sweden, Klostergatan 23, SE-70189 Örebro, Suède, anders.holmberg@scb.se.

sont disponibles au stade de la planification, ainsi qu'au stade de l'estimation pour tout élément de la population visée; (ii) bonne connaissance a priori de la structure des relations entre les variables importantes à l'étude et les variables auxiliaires (ou des sous-ensembles de ces variables).

Comme les variables auxiliaires peuvent avoir des usages fort divers au double stade de la conception et de l'estimation, les efforts en vue d'utiliser cette information pour obtenir des estimateurs efficients conduisent à un grand nombre de techniques d'échantillonnage et d'estimation. (Les exercices de répartition optimale en échantillonnage stratifié, en échantillonnage avec probabilité proportionnelle à la taille et en estimation par régression ou étalonnage sont tous fondés sur les propriétés de variables auxiliaires.) Si les données auxiliaires sont individuellement choisies et peuvent être différentes selon l'estimateur, elles sont communes dans le cas des plans d'échantillonnage et influeront donc sur toutes les estimations de paramètres. Il s'ensuit que, lorsqu'on planifie une enquête comportant plusieurs variables, le choix d'un plan – puisqu'il influe sur toutes les estimations – importe relativement plus que le choix d'estimateurs. Si un bon estimateur peut parfois compenser un mauvais plan d'échantillonnage et permettre une précision acceptable, on peut tout aussi bien penser qu'on perd de la précision par rapport à ce qui serait réalisable avec le même estimateur et un meilleur plan. Avant d'appliquer un plan qui prévoit le recours à des données auxiliaires, le statisticien doit donc en examiner de près les effets sur tous les estimateurs clés. Nous nous concentrerons sur ces effets et sur le choix d'un bon compromis là où les variables auxiliaires se prêtent à de nombreux usages. Le contenu théorique de ce document résume les résultats antérieurs sur l'optimalité des plans d'échantillonnage et sera ici appliqué à une population d'entreprises suédoises.

1.2 Énoncé du problème et quelques éléments de notation

Au départ, nous aurons besoin de quelques explications du contexte théorique et de certains éléments de notation. Supposons que nous planifions une enquête portant sur une population finie $U=\{1,\ldots,k,\ldots,N\}$ et que les paramètres clés à estimer sont les totaux de population des variables inconnues à l'étude $y_1,\ldots,y_q,\ldots,y_Q$, c'est-à-dire $\mathbf{t}=\left(t_{y_1},\ldots,t_{y_q},\ldots,t_{y_Q}\right)'$, où $t_{y_q}=\sum_{k\in U}y_{qk}=\sum_U y_{qk}$. Nous appliquerons un plan d'échantillonnage sans remise $p(\bullet)$ avec probabilités d'inclusion du premier ordre $\pi_k\ (k=1,\ldots,N)$ et probabilités d'inclusion du second ordre $\pi_{kl}\ (k,l=1,\ldots,N)$ pour tirer un échantillon aléatoire $s\subseteq U$ de taille n_s et estimerons chacun des totaux de population par les estimateurs $\hat{\mathbf{t}}=\left(\hat{t}_{y_1},\ldots,\hat{t}_{y_q},\ldots,\hat{t}_{y_Q}\right)'$.

Parmi les critères menant au choix du plan d'échantillonnage, il devrait y avoir un critère de précision. En temps normal, un tel critère se définit par des fonctions de la variance des estimateurs. Ainsi, on peut planifier de minimiser une fonction f de la variance de tous les Q estimateurs en respectant des restrictions spécifiées v_q sur les fonctions g de chaque variance d'estimateur (ou d'approximation de variance). En d'autres termes, nous minimisons

$$f(g(V(\hat{t}_{y_1})),g(V(\hat{t}_{y_2})),\ldots,g(V(\hat{t}_{y_Q}))) \tag{1}$$

avec des restrictions sur

$$g\left(V\left(\hat{t}_{y_q}\right)\right)\le v_q\ (q=1,\ldots,Q). \tag{2}$$

Pour être d'un usage pratique en planification d'enquête, un critère général comme celui que nous venons de présenter doit être simplifié, concrétisé et adapté à ce qui est connu et disponible, c'est-à-dire à des données auxiliaires pour l'essentiel. Nous supposons ici l'existence de P variables auxiliaires accessibles au stade de la planification. Désignées par $u_1,\ldots,u_p,\ldots,u_P$, leurs valeurs $u_{pk}\ (p=1,\ldots,P)$ sont connues pour tout élément *k* de la population.

En planification, on peut aussi s'aider de modèles statistiques. Si on dispose de renseignements utiles a priori sur les relations entre les variables étudiées et les variables auxiliaires, on pourra élaborer des modèles qui décrivent ces relations. Pour citer un bon exemple courant, on peut formuler des modèles linéaires ξ_q, $\left(y_{qk} = \mathbf{x}'_{qk}\boldsymbol{\beta}_q + \varepsilon_{qk}\right)$ pour les variables étudiées avec $E_{\xi_q}\left(\varepsilon_{qk}\right) = 0$, $V_{\xi_q}\left(\varepsilon_{qk}\right) = \sigma^2_{qk}$ et $E_{\xi_q}\left(\varepsilon_{qk}\varepsilon_{ql}\right) = 0 \ \ (k \neq l)$. En d'autres termes,

$$\begin{aligned} E_{\xi q}\left(y_{qk}\right) &= \mathbf{x}'_{qk}\boldsymbol{\beta}_q \\ V_{\xi q}\left(y_{qk}\right) &= \sigma^2_{qk}, \end{aligned} \tag{3}$$

où $\mathbf{x}'_{qk} = \left(x_{1_q k}, \ldots, x_{j_q k}, \ldots, x_{J_q k}\right)$ est un ensemble approprié de J_q variables auxiliaires (positives) formées à partir de $u_1, \ldots, u_p, \ldots, u_P$ et où $\boldsymbol{\beta}_q = \left(\beta_{1_q}, \ldots, \beta_{j_q}, \ldots, \beta_{J_q}\right)'$ et σ^2_{qk} sont des paramètres de modélisation.

Les modèles sont utiles, puisqu'il nous faut – lorsque nous planifions une enquête en fonction d'un critère spécifié de précision – un substitut pour la variance inconnue des estimateurs. Il nous faut en outre une certaine mesure pouvant nous aider à discriminer et faire un choix entre les stratégies possibles. Cette mesure doit être fonction des données auxiliaires, qui sont les seuls renseignements disponibles. La variance anticipée, définie par Isaki (1970) (voir aussi Isaki et Fuller, 1982) est une de ces mesures. Si $\hat{t}_{y_q}$ est un estimateur de t_{y_q} et que (3) s'interprète comme les moments d'une surpopulation dont est tirée une population finie, la variance anticipée est la variance de $\hat{t}_{y_q} - t_{y_q}$ aussi bien sur le modèle ξ_q que sur le plan d'échantillonnage. En d'autres termes,

$$E_{\xi_q} E_p\left[\left(\hat{t}_{y_q} - t_{y_q}\right)^2\right] - \left[E_{\xi_q} E_p\left(\hat{t}_{y_q} - t_{y_q}\right)\right]^2.$$

On peut se reporter à la variance anticipée pour comparer les propriétés des stratégies possibles. Dans le présent document, les estimateurs ponctuels $\hat{\mathbf{t}}$ font partie de la famille des estimateurs par régression généralisée (GREG) (régression généralisée). L'estimateur GREG se définit comme

$$\hat{t}_{y_q r} = \hat{t}_{y_q \pi} + \left(\mathbf{t}_{x_q} - \hat{\mathbf{t}}_{x_q \pi}\right)' \hat{\mathbf{B}}_q. \tag{4}$$

Dans ce cas, $\hat{t}_{y_q \pi} = \sum_{k \in s} y_{qk} / \pi_k = \sum_s y_{qk} / \pi_k$ est l'estimateur bien connu de Horvitz-Thompson ou l'estimateur π, $\mathbf{t}_{x_q} = \left(t_{x_{1q}}, \ldots, t_{x_{jq}}, \ldots, t_{x_{Jq}}\right)'$ est un vecteur de dimension J_q pour les totaux x_q, $\hat{\mathbf{t}}_{x_q \pi}$ est un vecteur d'estimateurs π correspondants et

$$\hat{\mathbf{B}}_q = \left(\sum_s \frac{\mathbf{x}_{qk}\mathbf{x}'_{qk}}{c_{qk}\pi_k}\right)^{-1} \sum_s \frac{\mathbf{x}_{qk} y_{qk}}{c_{qk}\pi_k} \tag{5}$$

est un vecteur estimé de coefficients de régression, où c_{qk} est une constante appropriée. (On trouvera des détails sur l'estimation REGG dans Särndal, Swensson et Wretman (1992), sections 6.4 à 6.7.)

2. PLANIFICATION OPTIMALE D'UN PLAN D'ÉCHANTILLONNAGE

2.1 Contexte théorique d'un traitement à variable unique

Si un modèle ξ_q est bien spécifié, l'approximation de la variance anticipée de $\hat{t}_{y_q r}$ peut s'écrire sous la forme $ANV_q(\hat{t}_{y_q r}) = \sum_U \left(\pi_k^{-1} - 1\right)\sigma_{qk}^2$. Le plan d'échantillonnage qui minimise $ANV_q(\hat{t}_{y_q r})$ (pour un q donné, une taille déterminée d'échantillon et sous la restriction $0 < \pi_k \leq 1$) est celui où $\pi_k \propto \sigma_{qk}$ $\left(k = 1, \ldots, N\right)$. Pour les plans à taille fixe, d'autres auteurs ont présenté des résultats qui indiquent aussi qu'on parvient à l'optimalité lorsque $\pi_k \propto \sigma_{qk}$; voir Hajek (1959) (estimateurs linéaires sans biais de plan d'échantillonnage), Brewer (1963) (estimation par quotient) et Cassel, Särndal et Wretman (1976) (estimateurs de différences généralisées). Pour un estimateur π coïncidant avec l'estimateur par quotient, Godambe (1955) présente aussi des résultats du même ordre.

Dans la plupart des présentations, la notion d'optimalité a à voir avec une variable unique à l'étude. On en relève des exemples dans la théorie ($\pi_k \propto \sigma_{qk}$) déjà évoquée (et qui normalement mène au choix de probabilités d'inclusion inégales), dans celle de la stratification optimale et de la répartition optimale en échantillonnage stratifié et enfin dans la théorie de l'échantillonnage à plusieurs degrés où le nombre d'unités peut faire l'objet d'une détermination optimale pour chaque degré. (On trouvera dans Rao (1979) et Bellhouse (1981) des examens des divers résultats d'optimalité de plans d'échantillonnage pour les enquêtes par sondage.) Il n'est pas si simple cependant d'appliquer ces exemples aux tâches pratiques de planification d'une enquête comportant plusieurs variables étudiées dont les relations avec les variables auxiliaires sont différentes. Nous allons proposer une approche.

2.2 Extension aux plans d'échantillonnage d'une enquête comportant plusieurs variables

Si peu d'auteurs traitent du problème des enquêtes par sondage comportant plusieurs variables, c'est peut-être parcequ'il n'y a pas de critère évident d'optimalité dans ce cas. Voici un exemple de la façon possible d'appliquer la théorie ($\pi_k \propto \sigma_{qk}$) à la planification de telles enquêtes. D'abord, nous reformulons notre critère (1) et remplaçons les estimations de variances inconnues par les approximations de variances anticipées $ANV_q(\hat{t}_{y_q r})$ $(q = 1, \ldots, Q)$ et, comme nous avons seulement les moyennes du stade de la planification, nous devons remplacer les σ_{qk}^2 inconnus par des estimations au jugé (guesstimates) de la structure de variance. Nous désignerons ces estimations au jugé aux fins de la planification par le symbole du tilde, $\tilde{\sigma}_{qk}^2$. (Au chapitre 7 de Brewer (2002), on parle de choix courants et acceptables de $\tilde{\sigma}_{qk}^2$.) Si nous disposons de bonnes estimations au jugé et appliquons la théorie ($\pi_k \propto \sigma_{qk}$), un plan d'échantillonnage à privilégier dans un traitement à variable unique se caractériserait par un ensemble préféré de probabilités d'inclusion du premier ordre où $\pi_k \propto \tilde{\sigma}_{qk}$. Toutefois, dans un traitement à variables multiples, il pourrait y avoir un tel ensemble préféré pour chacune des variables clés et peut-être autant de plans optimaux à privilégier qu'il y a de variables clés. En pareil cas, on doit soigneusement examiner l'effet global des différents plans possibles avant de décider du plan définitif. Dans bien des cas, il nous faut trouver un plan de compromis (à cause de considérations reliées aux variables multiples).

Dans l'examen des divers effets de plan d'échantillonnage, il est acceptable de calculer pour chaque q $\left(q = 1, \ldots, Q\right)$ l'ensemble préféré de probabilités d'inclusion du premier ordre qui est nécessaire à la minimisation d' $ANV_q(\hat{t}_{y_q r})$. Les valeurs des ensembles en question seront désignées par $\tilde{\pi}_{q(opt)k}$. Ainsi, il y a au stade de la planification des calculs reposant sur des estimations au jugé et des relations posées par le modèle. Le minimum d' $ANV_q(\hat{t}_{y_q r})$ à employer à ce stade est alors donné par $ANV_{q\min}(\hat{t}_{y_q r}) = \sum_U \left(\tilde{\pi}_{q(opt)k}^{-1} - 1\right)\tilde{\sigma}_{qk}^2$ pour $\left(q = 1, \ldots, Q\right)$.

Comme critère d'optimisation dans un traitement à variables multiples, la fonction objective à minimiser f pourrait simplement être (comme ci-après) une moyenne arithmétique (peut-être pondérée) de rapports relatifs où on tient compte de la précision de tous les $\hat{t}_{y_q r}$. On a alors un genre de fonction de perte que nous pourrions appeler perte globale anticipée d'efficience relative (« Anticipated Overall Relative Efficiency Loss » ou *ANOREL*), qui se définit comme

$$ANOREL = \sum_{q=1}^{Q} H_q \frac{ANV_q(\hat{t}_{y_q r})_{p_i}}{ANV_{q\min}(\hat{t}_{y_q r})}, \tag{6}$$

où les H_q $(q = 1,\ldots,Q)$ sont des poids (que l'on somme à 1) qui traduisent l'importance relative des paramètres à estimer. $ANV_q(\hat{t}_{y_q r})_{p_i}$ est l'approximation de la variance anticipée de $\hat{t}_{y_q r}$ dans un plan $p_i(\bullet)$ avec $\pi_k = \pi_{p_i k}$. Pour une taille d'échantillon donnée, on obtient le minimum d'*ANOREL* si

$$\pi_k \propto \sqrt{\sum_{q=1}^{Q} H_q \frac{\tilde{\sigma}_{qk}^2}{\sum_U \left(\tilde{\pi}_{q(opt)k}^{-1} - 1\right)\tilde{\sigma}_{qk}^2}}. \tag{7}$$

Toutefois, en cas de restrictions v_q (limites supérieures) sur les rapports, c'est-à-dire

$$\frac{ANV_q\left(\hat{t}_{y_q r}\right)_{p_i}}{ANV_{q\min}\left(\hat{t}_{y_q r}\right)} \le v_q \;\; q = 1,\ldots,Q, \tag{8}$$

la minimisation de (6) devient un problème à résoudre d'optimisation non linéaire, que l'on peut exprimer comme dans les formules (9) et (10) : on minimise la fonction objective

$$f(\boldsymbol{\pi}) = \sum_{q=1}^{Q} H_q \sum_U (\pi_k^{-1} - 1) \frac{\tilde{\sigma}_{qk}^2}{\sum_U (\tilde{\pi}_{q(opt)k}^{-1} - 1)\tilde{\sigma}_{qk}^2} \tag{9}$$

en tenant compte des $2N + Q + 1$ restrictions

$$\begin{aligned} &0 < \pi_k \le 1 \quad k = 1,\ldots,N \\ g_0(\boldsymbol{\pi}) &= \textstyle\sum_U \pi_k - n = 0 \\ g_q(\boldsymbol{\pi}) &= \textstyle\sum_U \left(\pi_k^{-1} - 1\right) \dfrac{\tilde{\sigma}_{qk}^2}{\sum_U \left(\tilde{\pi}_{q(opt)k}^{-1} - 1\right)\tilde{\sigma}_{qk}^2} \le v_q \;\; q = 1,\ldots,Q. \end{aligned} \tag{10}$$

On trouvera des exemples d'autres fonctions objectives et une description détaillée du modèle d'optimisation dans Holmberg (2002) et Holmberg, Flisberg et Rönnqvist (2003).

Remarque : La démarche de planification adoptée ici qui permet de trouver une bonne stratégie pour une enquête par sondage donne lieu à une décision portant sur une catégorie de *plans d'échantillonnage préférés* se caractérisant par un ensemble préféré de probabilités d'inclusion du premier ordre. Pour appliquer un plan conforme à cet ensemble, il nous faut un plan de sélection d'échantillon. Nous n'aborderons pas cette question, mais diverses solutions sont possibles, qu'il s'agisse de plans à taille aléatoire comme dans un échantillonnage de Poisson ou de plans à taille fixe comme dans l'échantillonnage de Rao-Sampford ou de Pareto (qu'ont indépendamment proposé

Saavedra (1995) et Rosén (1997)). Si on ne juge pas importante une estimation de variance qui soit (du moins approximativement) sans biais, l'échantillonnage systématique est une autre solution possible.

2.3 Observations sur les méthodes applicables au choix d'un plan d'échantillonnage dans un traitement à variables multiples

En quoi cette extension évoquée aux plans d'échantillonnage à plusieurs variables est-elle avantageuse? Il semble y avoir trois manières courantes d'aborder ce problème dans la pratique : (i) on écarte le problème et le ramène à un problème de traitement à variable unique par un jugement reposant sur l'expérience; (ii) pour chacun des paramètres importants, on étudie l'effet de diverses mesures individuelles de variables, puis on choisit un plan d'échantillonnage qui paraît constituer le meilleur compromis; (iii) on trouve un genre de critère général applicable au traitement à variables multiples, peut-être mécaniquement par une généralisation de concepts applicables au traitement à variable unique (comme à la section 2.2), et on choisit l'opération qui optimise ce critère.

Le premier choix pourrait ne pas être aussi mauvais qu'il le semble. Dans maintes situations, quelques variables importantes à l'étude peuvent présenter des propriétés semblables et la perte totale de précision pourrait être acceptable si on choisit de se concentrer sur une de ces variables seulement. Toutefois, la deuxième solution est meilleure, car le choix d'un plan se fait expressément en fonction de considérations relatives au traitement à variables multiples. Une méthode décrite par Kott et Bailey (2000) qui s'appelle la *méthode de sélection maximale de Brewer* en est un exemple. Ce n'est pas une méthode d'optimisation, mais elle est d'une application simple et garantit la précision recherchée de tous les estimateurs visés au stade de la planification. Diverses méthodes par moyenne de mesures clés appartiennent à cette catégorie de techniques. Une possibilité est de tirer un échantillon stratifié où la répartition est basée sur la moyenne de diverses répartitions optimales qui concernent individuellement les variables. Une autre possibilité, pour un échantillonnage aux probabilités proportionnelles à la taille avec σ_q comme mesure de taille, est de prendre la moyenne ou la médiane des Q différents σ_q se rapportant aux variables individuelles.

En ce qui concerne la troisième voie (par optimisation) que nous avons mentionnée, un certain nombre d'auteurs ont présenté des solutions pour un échantillonnage stratifié à variables multiples, dont Dalenius (1957), Chatterjee (1968), Hughes et Rao (1979) et Chromy (1987). Voici d'autres références utiles à cet égard : Sigman et Monsour (1995), qui ont conçu une méthode de programmation non linéaire semblable à celle de la section 2.2 pour l'échantillonnage πps de Poisson et l'estimateur π; Saavedra (1999), qui a appliqué ces idées à l'aide de l'algorithme proposé par Chromy en vue de déterminer des probabilités à utiliser pour l'échantillonnage πps de Pareto dans une enquête de prix et de volumes sur les produits pétroliers.

Avec cette troisième voie, l'avantage est, du moins en théorie, l'obtention de solutions optimales. Ajoutons que, en choisissant le critère d'optimisation, nous exerçons aussi un certain contrôle sur les propriétés recherchées en planification d'échantillonnage par opposition aux solutions intuitives à caractère spécial. Toutefois, comme pour toutes les techniques possibles, le succès dépendra en définitive du critère et des hypothèses de base retenus par le concepteur de l'enquête.

3. ILLUSTRATION DE L'EXTENSION AU TRAITEMENT À VARIABLES MULTIPLES

3.1 Description de l'application étudiée

Nous illustrons l'extension à un traitement à variables multiples de la théorie $\pi_k \propto \sigma_{qk}$ exposée à la section 2.2 par une application simple sur une population d'entreprises suédoises. Les données sont extraites d'un registre administratif de ces entreprises et, dans l'exemple que nous allons présenter, nous étudions les entreprises appartenant à la branche d'activité « fabricants de produits métalliques (sans les machines ni les dispositifs) ». Pour

les éléments de cette population $k = 1, \ldots, 2292$, nous disposons des valeurs de quatre variables auxiliaires : *nombre de travailleurs* u_{1k}, *roulement du personnel* u_{2k}, *dépenses de personnel* u_{3k} et *investissements* u_{4k}. Supposons que l'enquête vise à l'estimation des totaux annuels de population de ces variables pour une période de référence postérieure à celle des données du registre. Ainsi, sauf pour la période de référence, les définitions des quatre variables auxiliaires correspondent à celles des variables étudiées $q = 1, \ldots, 4$, t_{y_1} étant le nombre total de travailleurs, t_{y_2} le roulement total, et ainsi de suite. L'expérience nous dit que, pour le nombre de travailleurs, le roulement et les dépenses de personnel, un simple modèle par quotient décrit assez bien les relations entre la variable auxiliaire et la variable étudiée. Ainsi, pour $q = 1, 2, 3$, nos modèles du stade de la planification (dans un traitement à variable unique) sont

$$\begin{aligned} E_{\xi q}\left(y_{qk}\right) &= \beta_q u_{qk} \\ V_{\xi q}\left(y_{qk}\right) &= \sigma_{qk}^2 = u_{qk}. \end{aligned} \tag{11}$$

Nous nous reportons, par conséquent, aux u_{qk} en guise d'estimations au jugé $\tilde{\sigma}_{qk}^2$ pour $q = 1, 2, 3$. Pour la variable des investissements $q = 4$, des diagrammes de dispersion et des estimations venant de données relatives aux années antérieures indiquent que l'hétéroscédasticité de la variance comme (11) n'a pas sa place. Comme solution alternative dans ce cas, nous employons $\tilde{\sigma}_{4k}^2 = 1$. Les quatre variables étudiées sont jugées d'une même importance, et nous supposons que notre budget permet de constituer un échantillon d'une taille espérée $E_p(n_s) = 290$. Il est également nécessaire que notre plan d'échantillonnage soit tel qu'aucun $ANV_q(\hat{t}_{y_q r})_{p_i}$ n'excède $ANV_{q\min}(\hat{t}_{y_q r})$ de plus de 10 %, c'est-à-dire que $\nu_q = 1{,}1$ $(q = 1, 2, 3, 4)$. Avec cette dernière exigence, il est improbable qu'un plan autre qu'une solution du problème d'optimisation (9-10) soit satisfaisant, mais nous ne pouvons en être certains à moins de procéder à certaines comparaisons au stade de la planification.

3.2 Calcul des pertes d'efficience relative au stade de la planification

Pour faire un bon choix de plan d'échantillonnage, nous préparons en planification un certain nombre de diagnostics qui illustrent les propriétés des divers plans possibles. Nous pouvons, par exemple, calculer les pertes anticipées d'efficience pour les divers plans envisagés. Si nous le faisons dans notre application à une enquête comportant plusieurs variables, nous obtenons les résultats qui suivent.

Étant donné les conditions énoncées à la section 3.1, six plans s'offrent d'emblée. Soit $p_1, \ldots, p_4$ les plans qui découlent du choix de $\pi_k \propto \tilde{\sigma}_{qk}$, c'est-à-dire les plans tirés du calcul de $\tilde{\pi}_{q(opt)k}$ $(q = 1, 2, 3, 4)$. Soit p_5 le plan par choix de π_k selon l'équation (7) et p_6 un plan par solution du problème d'optimisation aux équations (9) et (10). Nous récapitulons les diagnostics de ces plans au tableau 1, dont les cellules présentent les valeurs de perte d'efficience relative, c'est-à-dire $100[(ANV_q(\hat{t}_{y_q r})_{p_i} \big/ ANV_{q\min}(\hat{t}_{y_q r})) - 1]$.

Tableau 1 : Pertes anticipées d'efficience relative en pourcentage pour les six plans d'échantillonnage envisagés					
	Variables étudiées				
Plan envisagé	y_1	y_2	y_3	y_4	*Moyenne*
p_1	0	7,3	2,0	21,0	7,6
p_2	6,1	0	4,3	33,0	10,9
p_3	1,8	5,0	0	24,4	7,8
p_4	31,6	51,2	36,1	0	29,7
p_5	1,7	4,9	1,9	14,0	5,6
p_6	3,4	7,0	4,0	10,0	6,1

Rien d'étonnant à ce que nous constations au tableau 1 que le plan p_1 (optimal pour le nombre de travailleurs) et le plan p_3 (optimal pour les dépenses de personnel) se ressemblent fort. On peut s'attendre à ce qu'ils donnent de bons résultats dans l'estimation de t_{y1} et t_{y3} et de passablement bons dans l'estimation de t_{y2}. Des propriétés analogues (mais avec une plus grande perte moyenne d'efficience) sont indiquées par les pertes anticipées d'efficience pour le plan p_2. Aucun des trois plans n'est satisfaisant pour la variable des investissements y_4. Les plans de compromis p_5 et p_6, qui sont fondés sur les extensions au traitement à variables multiples et sur l'optimisation d'ANOREL à l'équation (6), sont préférables à cet égard. La perte moyenne d'efficience est plus grande pour le plan p_6 (6,1 %) que pour le plan p_5. C'est le prix à payer pour le respect de la restriction selon laquelle $ANV_q\left(\hat{t}_{yq^r}\right)_{p_1} / ANV_{q\min}\left(\hat{t}_{yq^r}\right) \leq 1{,}1$ pour tout $q = 1,2,3,4$.

4. OBSERVATIONS EN CONCLUSION

Dans l'application qui précède, seul le plan p_6 par optimisation dans un contexte de variables multiples se révèle satisfaisant et il est globalement supérieur en efficience aux plans par optimisation individuelle de variables. Pour le praticien, le caractère fastidieux des calculs est un inconvénient possible d'une démarche par optimisation. Il est heureux que, avec un programme informatique, l'extension proposée au traitement à variables multiples soit seulement un peu plus complexe que les méthodes applicables à un traitement à variable unique. Ajoutons que la méthode mise de l'avant repose dans une large mesure sur des calculs qui devraient de toute manière se faire dans tout exercice sérieux de planification. Avec l'extension en contexte de variables multiples, un point important à faire valoir est celui de la souplesse de la solution à trouver pour exploiter des données auxiliaires de façon exhaustive dans la planification d'échantillonnage. Le statisticien d'enquête a le loisir de retenir toute combinaison appropriée de ces données pour toutes les variables clés, mais il faut bien dire que, dans le cas des enquêtes à plusieurs variables, le travail consistant à obtenir et à analyser les diagnostics qui justifieront le choix définitif d'un plan demeure le plus important. Les statistiques qui illustrent les effets de plan d'échantillonnage dans un traitement à variables multiples sont particulièrement utiles non seulement pour la décision de conception du statisticien d'enquête, mais aussi dans les discussions à tenir avec les non-statisticiens associés à cette prise de décisions. Avec un programme d'optimisation non linéaire, il est à la fois rapide et relativement facile de procéder à de nouveaux calculs si on veut, par exemple, cerner les effets de modifications des restrictions et de la taille d'échantillon.

RÉFÉRENCES

Bellhouse, D. R., (1984). A review of optimal designs in survey sampling. Canadian Journal of Statistics, 12 pp 53-65.

Brewer, K.R.W. (1963). Ratio Estimation and Finite population: Some results deductible from the assumption of an underlying stochastic process. Australian Journal of Statistics 5, 93–105.

Brewer, K.R.W. (2002). Combined Survey Sampling Inference, Weighting Basu's Elephants. Arnold, London.

Cassel, C.M., Särndal, C-E. and Wretman, J. (1976). Some results on generalized difference estimators and generalized regression estimators for finite populations. Biometrika 63, 615–620

Chatterjee, S. (1968). Multivariate stratified surveys. Journal of the American Statistical Association 63, 530–534.

Chromy, J. (1987). Design Optimization with Multiple Objectives. Proceedings of the Section on Survey Research Methods, American Statistical Association 1987 194–199.

Dalenius, T. (1957). Sampling in Sweden. Almquist & Wiksell, Stockholm

Godambe, V.P. (1955). A unified theory of sampling from finite populations. Journal of the Royal Statistical Society, Series B, 17 269–278.

Hajek, J., (1959). Optimum strategy and other problems in probability sampling. Casopis pro Pestovani Matematiky 84 387–421.

Holmberg, A., (2002). A multiparameter perspective on the choice of sampling design in surveys. Statistics in Transition, 5, no 6, pp. 969–994.

Holmberg, A., Flisberg, P., and Rönnqvist, M., (2003). On the choice of optimal design in business surveys with several important study variables. In: Holmberg, A., (2003). Essays on Model Assisted Survey Planning. Comprehensive Summaries of Uppsala Dissertations from the Faculty of Social Sciences 126, Doctoral thesis Uppsala University, Sweden.

Hughes, E., and Rao, J.N.K. (1979) Some problems of optimal allocation in sample surveys involving inequality constraints. Communications in Statistics A 8, 1551–1574.

Isaki, C.T. (1970). Survey designs utilizing prior information. Doctoral thesis, Iowa State University, Ames, Iowa.

Isaki, C.T. and Fuller, W.A. (1982). Survey design under the regression superpopulation model, Journal of the American Statistical Association, 77, 89–96

Kott, P.S. and Bailey, J.T. (2000). The Theory and Practice of Maximal Brewer Selection with Poisson PRN Sampling, Proceedings of the second International Conference on Establishment Surveys, June 17–21, 2000, Buffalo 269–279.

Rao, J.N.K., (1979). Optimization in the design of sample surveys. In: J.S. Rustagi (ed.), Optimization methods in Statistics: Proceedings of an International Conference. New York, Academic Press, pp 419-434

Rosén, B., (1997). On sampling with Probability Proportional to Size, Journal of Statistical Planning and Inference, 62, 159-191.

Saavedra, P.J. (1995). Fixed Sample Size PPS Approximations with a Permanent Random Number. Proceedings of the section on Survey research Methods Joint Statistical Meetings, American Statistical Association, 697-700.

Saavedra, P.J. (1999). Application of the Chromy Algorithm with Pareto Sampling, Proceedings of the Section on Survey Research Methods Joint Statistical Meetings, American Statistical Association 1999 355–358.

Särndal, C-E., Swensson, B. and Wretman, J. (1992). Model Assisted Survey Sampling. Springer, New York.

Sigman, R.S. and Monsour, N.J. (1995). Selecting Samples from List Frames of Businesses, in Cox, B.G., Binder, D.A., Chinnappa, N., Christianson, A., Colledge, M.J., and Kott, P.S. (eds) Business Survey Methods, New York: Wiley, 153–169.