Nº 12-001-X au catalogue
ISSN 1712-5685

Techniques d'enquête

Application des formulations de la programmation en nombres entiers à la répartition optimale dans l'échantillonnage stratifié

par José André de Moura Brito, Pedro Luis do Nascimento Silva, Gustavo Silva Semaan et Nelson Maculan

Date de diffusion : le 17 décembre 2015

Statistique Canada Statistics Canada

Canada

Comment obtenir d'autres renseignements

Pour toute demande de renseignements au sujet de ce produit ou sur l'ensemble des données et des services de Statistique Canada, visiter notre site Web à www.statcan.gc.ca.

Vous pouvez également communiquer avec nous par :

Courriel à STATCAN.infostats-infostats.STATCAN@canada.ca

Téléphone entre 8 h 30 et 16 h 30 du lundi au vendredi aux numéros sans frais suivants :

- Service de renseignements statistiques 1-800-263-1136
- Service national d'appareils de télécommunications pour les malentendants 1-800-363-7629
- Télécopieur 1-877-287-4369

Programme des services de dépôt

- Service de renseignements 1-800-635-7943
- Télécopieur 1-800-565-7757

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois. À cet égard, notre organisme s'est doté de normes de service à la clientèle que les employés observent. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1-800-263-1136. Les normes de service sont aussi publiées sur le site www.statcan.gc.ca sous « Contactez-nous » > « Normes de service à la clientèle ».

Note de reconnaissance

Le succès du système statistique du Canada repose sur un partenariat bien établi entre Statistique Canada et la population du Canada, les entreprises, les administrations et les autres organismes. Sans cette collaboration et cette bonne volonté, il serait impossible de produire des statistiques exactes et actuelles.

Signes conventionnels dans les tableaux

Les signes conventionnels suivants sont employés dans les publications de Statistique Canada :

.	indisponible pour toute période de référence
..	indisponible pour une période de référence précise
...	n'ayant pas lieu de figurer
0	zéro absolu ou valeur arrondie à zéro
0^s	valeur arrondie à 0 (zéro) là où il y a une distinction importante entre le zéro absolu et la valeur arrondie
^{p}	provisoire
^{r}	révisé
x	confidentiel en vertu des dispositions de la *Loi sur la statistique*
^{E}	à utiliser avec prudence
F	trop peu fiable pour être publié
*	valeur significativement différente de l'estimation pour la catégorie de référence (p<0,05)

Une version HTML est aussi disponible.

This publication is also available in English.

Techniques d'enquête, décembre 2015
Vol. 41, nº 2, pp. 451-467
Statistique Canada, nº 12-001-X au catalogue

Application des formulations de la programmation en nombres entiers à la répartition optimale dans l'échantillonnage stratifié

José André de Moura Brito, Pedro Luis do Nascimento Silva, Gustavo Silva Semaan et Nelson Maculan[1]

Résumé

Le problème de la répartition optimale des échantillons dans les enquêtes basées sur un plan d'échantillonnage stratifié a été abordé pour la première fois par Neyman en 1934. Depuis, de nombreux chercheurs ont étudié le problème de la répartition des échantillons dans les enquêtes à plusieurs variables, et plusieurs méthodes ont été proposées. Ces méthodes se divisent essentiellement en deux catégories. La première catégorie englobe les méthodes de répartition qui réduisent les coûts des enquêtes tout en maintenant les coefficients de variation des estimateurs de totaux sous des seuils spécifiés pour toutes les variables d'enquête d'intérêt. La seconde catégorie de méthodes vise à minimiser une moyenne pondérée des variances relatives des estimateurs des totaux étant donné une taille globale maximale d'échantillon ou un coût maximum. Cet article propose une nouvelle approche d'optimisation pour régler le problème de la répartition des échantillons dans les enquêtes à plusieurs variables. Cette approche se fonde sur une formulation de la programmation en nombres entiers binaires. Plusieurs expériences numériques ont démontré que l'approche proposée offre des solutions efficaces à ce problème, qui permettent d'améliorer un « algorithme classique » et peuvent être plus efficaces que l'algorithme de Bethel (1985, 1989).

Mots-clés : Stratification; répartition; programmation en nombres entiers; enquête à plusieurs variables.

1 Introduction

Une grande partie des statistiques produites par les organismes statistiques officiels de nombreux pays proviennent d'enquêtes par sondage. Ces enquêtes couvrent une population bien définie, en fonction notamment du lieu géographique et d'autres critères d'admissibilité, utilisent des bases de sondage appropriées pour guider la sélection de l'échantillon et appliquent certaines procédures bien précises de sélection des échantillons. L'utilisation de procédures « standard » d'échantillonnage probabiliste permet de produire des estimations pour les paramètres de la population cible avec une précision contrôlée, tout en disposant de données généralement tirées de petits échantillons des populations, à une fraction du coût des recensements correspondants.

Au moment de concevoir la stratégie d'échantillonnage, le planificateur d'enquête cherche souvent à optimiser la précision pour les estimations d'enquête les plus importantes, compte tenu du budget disponible. La stratification est un outil important qui permet d'explorer l'information auxiliaire antérieure disponible pour toutes les unités de population en formant des groupes d'unités homogènes, puis en faisant un échantillonnage indépendant dans ces groupes. On fait donc une utilisation très fréquente de la stratification dans un vaste éventail d'enquêtes par sondage.

1. José André de Moura Brito et Pedro Luis do Nascimento Silva, Escola Nacional de Ciências Estatísticas (ENCE/IBGE), R. André Cavalcanti, 106, sala 403, Centro, Rio de Janeiro/RJ. Courriel : jambrito@gmail.com et pedronsilva@gmail.com; Gustavo Silva Semaan, Instituto do Noroeste Fluminense de Educação Superior, Universidade Federal Fluminense - INFES/UFF, Av. João Jasbick, s/n, Bairro Aeroporto - Santo Antônio de Pádua - RJ - CEP 28470-000. Courriel : gustavosemaan@gmail.com; Nelson Maculan, Universidade Federal do Rio de Janeiro (COPPE/UFRJ), Endereço : Av. Horácio Macedo, 2030 - CT, Bloco H, sl. 319 - Cidade Universitária, Ilha do Fundão - Rio de Janeiro, RJ - CEP 21941-914. Courriel : nelson.maculan@gmail.com.

Le présent article met l'accent sur les plans d'échantillonnage par élément (Särndal, Swensson et Wretman 1992), où la base de sondage est d'un enregistrement par unité de population et où une information auxiliaire s'ajoute aux données d'identification et de localisation disponibles pour chaque unité de population. L'échantillonnage stratifié consiste à diviser les N unités dans une population U en H groupes homogènes appelés strates. Ces groupes sont formés en tenant compte d'une ou de plusieurs variables de stratification et de manière à ce que la variance dans les groupes soit faible (problème de la formation des strates).

Compte tenu d'une taille d'échantillon n, une fois les strates définies, le prochain problème consiste à déterminer le nombre d'unités d'échantillonnage à sélectionner dans chaque strate de manière à minimiser la variance d'un estimateur spécifié (problème de la répartition optimale des échantillons). Lorsqu'il s'agit simplement d'estimer le total de la population (ou la moyenne de population) pour une seule variable d'enquête, on peut utiliser la répartition bien connue de Neyman (voir, par exemple, Cochran 1977) pour déterminer la répartition des échantillons. Bien que les enquêtes à une seule variable cible soient rares, la formule de répartition simple de Neyman peut quand même être utile, car la répartition optimale pour une variable cible peut être raisonnable pour d'autres variables d'enquête ayant une corrélation positive avec celle qui est à la base de la répartition optimale.

Lorsqu'une enquête doit produire des estimations dont les niveaux de précision sont spécifiés pour certaines variables d'enquête et qu'il n'y a pas de corrélation étroite entre ces variables, il faut utiliser une méthode de répartition des échantillons qui permet de produire des estimations aux niveaux de précision requis pour toutes les variables d'enquête. Nous avons alors un problème de répartition multivariée optimale des échantillons.

Selon la littérature sur le sujet, dans les cas de ce genre, la répartition de la taille d'échantillon globale n entre les strates peut avoir l'un ou l'autre des objectifs suivants :

(i) minimiser le coût variable total de l'enquête C, à condition que les coefficients de variation (CV) pour les estimations des totaux des m variables d'enquête soient inférieurs aux seuils spécifiés;
(ii) minimiser une somme pondérée des variances (ou des variances relatives) des estimations des totaux pour les m variables d'enquête.

Notons que le CV est simplement la racine carrée de la variance relative.

Cet article présente une nouvelle approche fondée sur l'élaboration et l'application de deux formulations de la programmation en nombres entiers binaires qui répondent à chacun de ces deux objectifs, tout en veillant à ce que la répartition résultante produise l'optimum global. L'article se divise comme suit. La section 2 examine certains concepts et certaines définitions clés de l'échantillonnage stratifié. La section 3 décrit la nouvelle approche proposée ici. La section 4 présente les résultats pour un sous-ensemble d'expériences numériques menées afin de tester l'approche proposée à l'aide de bases de données de populations choisies. La section 5 contient quelques observations finales. L'annexe A donne des informations sur trois populations utilisées dans les expériences numériques présentées à la section 4.

2 L'échantillonnage stratifié et le problème de la répartition optimale

Dans l'échantillonnage stratifié (Cochran 1977; Lohr 2010), une population U constituée de N unités est divisée en H strates $U_1, U_2, \ldots, U_H$ contenant $N_1, N_2, \ldots, N_H$ unités respectivement. Ces strates ne se chevauchent pas (2.1) et forment ensemble la population entière (2.2) de sorte que :

$$U_h \bigcap U_k = \varnothing, \quad h \neq k \tag{2.1}$$

$$\bigcup_{h=1}^{H} U_h = U \tag{2.2}$$

$$N_1 + N_2 + \ldots + N_H = \sum_{h=1}^{H} N_h = N. \tag{2.3}$$

Une fois les strates définies et étant donné une taille d'échantillon globale n, un échantillon indépendant de taille n_h est sélectionné à partir des N_h unités de la strate U_h $(h = 1, \ldots, H)$ de sorte que $n_{\min} \leq n_h \leq N_h \; \forall h$, où $n_{\min}$ est la plus petite taille d'échantillon possible dans une strate donnée, et $n_1 + n_2 + \ldots + n_H = \sum_{h=1}^{H} n_h = n$.

Une taille d'échantillon minimale par strate de $n_{\min} = 2$ est examinée ici, mais cette valeur peut être modifiée au besoin pour satisfaire à des exigences particulières de l'enquête. Une taille d'échantillon minimale de 1 par strate n'est pas recommandée, car cela pourrait mener à des solutions qui nécessitent l'utilisation de méthodes approximatives d'estimation de la variance lorsque les tailles d'échantillon attribuées atteignent ce minimum. Dans la pratique, il pourrait même être sage d'utiliser une taille minimale $n_{\min}$ supérieure à 2, en raison de la non-réponse ou pour d'autres raisons pratiques.

Dans l'hypothèse d'une réponse complète, les données sont recueillies pour toutes les unités de l'échantillon sélectionné et utilisées pour produire des estimations (des totaux, par exemple) pour un ensemble de m variables d'enquête. Soit $y_1, y_2, \ldots, y_m$ les variables d'enquête. La variance de la variable y_j de la strate h est définie comme suit :

$$S_{hj}^2 = \frac{1}{N_h - 1} \sum_{i \in U_h} \left(y_{ij} - \bar{Y}_{hj} \right)^2 \tag{2.4}$$

où y_{ij} est la valeur de y_j pour la i^{e} unité de population et $\bar{Y}_{hj}$ est la moyenne de population pour y_j dans la strate h, donnée par

$$\bar{Y}_{hj} = \frac{1}{N_h} \sum_{i \in U_h} y_{ij} = Y_{hj} / N_h \tag{2.5}$$

pour $h = 1, \ldots, H$ et $j = 1, \ldots, m$. Le total de la population Y_j pour la j^{e} variable d'enquête est $Y_j = \sum_{h=1}^{H} \sum_{i \in U_h} y_{ij} = \sum_{h=1}^{H} Y_{hj}$.

Dans l'échantillonnage aléatoire simple stratifié (EASS), la variance de l'estimateur t_j de Horvitz-Thompson (HT) du total pour la j^{e} variable d'enquête (Cochran 1977) est donnée par :

$$V\left(t_j\right) = \sum_{h=1}^{H} N_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) S_{hj}^2 \tag{2.6}$$

où $t_j = \sum_{h=1}^{H} N_h / n_h \sum_{i \in s_h} y_{ij} = \sum_{h=1}^{H} N_h \bar{y}_{hj}$, $s_h \subset U_h$ est l'ensemble d'étiquettes des n_h unités échantillonnées de la strate h, et $\bar{y}_{hj}$ est la moyenne de l'échantillon de la strate h.

Comme les valeurs de N_h et S_{hj}^2 sont fixées après que les strates ont été définies, la variance de l'estimateur HT t_j du total pour la j^e variable d'enquête en (2.6) dépend uniquement des tailles d'échantillon n_h attribuées aux strates. Cette répartition est importante, parce qu'elle permet au concepteur de l'enquête de contrôler la précision des estimations de l'enquête.

En général, le planificateur d'enquête qui procède à la répartition cherche à équilibrer l'atteinte de la précision souhaitée pour chacune des variables d'intérêt de l'enquête et le coût de l'enquête. L'importance de ce problème et la complexité des calculs connexes ont motivé de nombreuses contributions, qui sont axées sur un des deux objectifs associés au problème de la répartition, comme il est expliqué dans la section 1. Voir, par exemple, Kokan (1963), Folks et Antle (1965), Kokan et Khan (1967), Huddleston, Claypool et Hocking (1970), Kish (1976), Bethel (1985, 1989), Chromy (1987), Valliant et Gentle (1997), Khan et Ahsan (2003), García et Cortez (2006), Kozak (2006), Day (2010), Khan, Ali et Ahmad (2011), Ismail, Nasser et Ahmad (2011), Khan, Ali, Raghav et Bari (2012).

Tous les auteurs précités appliquent des méthodes fondées sur la théorie de la programmation linéaire, la programmation convexe, la programmation dynamique, la programmation multi-objectifs et de méthodes heuristiques pour essayer de résoudre le problème de la répartition multivariée optimale. Nous proposons ici deux formulations de la programmation en nombres entiers pour résoudre le problème.

Formulation A

$$\text{Minimiser} \sum_{h=1}^{H} c_h n_h \tag{2.7}$$

$$\text{s.c. } n_{\min} \le n_h \le N_h, \; h = 1, \ldots, H \tag{2.8}$$

$$\sqrt{V\left(t_j\right)} / Y_j \le \text{CV}_j \quad j = 1, \ldots, m \tag{2.9}$$

$$n_h \in Z_+ \quad h = 1, \ldots, H \tag{2.10}$$

où c_h représente le coût de l'enquête au niveau de l'unité pour l'échantillonnage de la strate h.

Dans cette formulation, la fonction objectif à minimiser (2.7) correspond au budget global des coûts variables pour l'enquête (désigné par C). Si les coûts de l'enquête au niveau de l'unité pour l'échantillonnage des différentes strates sont inconnus ou présumés être les mêmes, alors les coûts c_h peuvent tous être fixés à un, et l'autre fonction objectif à minimiser est $n = \sum_{h=1}^{H} n_h$, soit la taille d'échantillon globale.

La contrainte (2.8) permet de s'assurer qu'au moins $n_{\min}$ unités sont attribuées à chaque strate, et que la taille d'échantillon ne dépassera pas la taille de population pour la strate.

La contrainte (2.9) permet de s'assurer que le CV de l'estimateur HT du total pour chaque variable d'enquête est inférieur à un seuil prédéterminé CV_j $(j = 1, \ldots, m)$ appelé CV cible. Enfin, la contrainte (2.10) permet de s'assurer que toutes les tailles d'échantillon attribuées sont des nombres entiers.

Il est à noter que les contraintes (2.9) peuvent être réécrites comme suit :

$$\frac{V(t_j)}{Y_j^2 \ \mathrm{CV}_j^2} \leq 1 \, , j = 1, \ldots, m. \tag{2.11}$$

Le remplacement du numérateur en (2.11) par l'équation (2.6) mène alors à :

$$\sum\nolimits_{h=1}^{H} \left(\frac{N_h^2 S_{hj}^2}{n_h Y_j^2 \mathrm{CV}_j^2} - \frac{N_h S_{hj}^2}{Y_j^2 \mathrm{CV}_j^2} \right) \leq 1, \quad j = 1, \ldots, m. \tag{2.12}$$

Si l'on utilise la définition

$$p_{hj} = \frac{N_h \ S_{hj}^2}{Y_j^2 \ \mathrm{CV}_j^2} \tag{2.13}$$

les contraintes (2.12) peuvent s'écrire sous la forme :

$$\sum\nolimits_{h=1}^{H} \left(\frac{N_h \ p_{hj}}{n_h} - \ p_{hj} \right) \leq 1, \quad j = 1, \ldots, m. \tag{2.14}$$

Formulation B

$$\text{Minimiser} \sum\nolimits_{j=1}^{m} w_j \frac{1}{Y_j^2} \left[\sum\nolimits_{h=1}^{H} N_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h} \right) S_{hj}^2 \right] \tag{2.15}$$

$$\text{s.c. } n_{\min} \leq n_h \leq N_h, \quad h = 1, \ldots, H \tag{2.16}$$

$$\sum_{h=1}^{H} c_h n_h \leq C \tag{2.17}$$

$$n_h \in Z_+ \quad h = 1, \ldots, H \tag{2.18}$$

$$0 < w_j < 1 \ \ \forall j \ \text{ et } \sum\nolimits_{j=1}^{m} w_j = 1 \tag{2.19}$$

où w_j sont des poids propres aux variables, fixés a priori pour représenter l'importance relative des variables de l'enquête. Les poids propres aux variables w_j sont fixés par des spécialistes ou par les

concepteurs de l'enquête. S'ils ne sont pas fixés, des poids relatifs égaux peuvent être attribués à toutes les variables de l'enquête prises en considération.

Dans cette formulation, la fonction objectif (2.15) à minimiser correspond à une somme pondérée des variances relatives des estimations du total pour les m variables d'enquête. Nous utilisons des variances relatives, parce que différentes variables d'enquête peuvent être mesurées dans différentes unités et que la somme des variances ne serait donc pas significative. Si l'on examine (2.15), il est clair que son minimum est atteint quand

$$\sum_{j=1}^{m} w_j \frac{1}{Y_j^2}\left[\sum_{h=1}^{H} N_h^2 \left(\frac{1}{n_h}\right) S_{hj}^2\right]$$

est minimum, puisque le dernier terme

$$\sum_{j=1}^{m} w_j \frac{1}{Y_j^2}\left[\sum_{h=1}^{H} N_h^2 \left(-\frac{1}{N_h}\right) S_{hj}^2\right]$$

ne dépend pas des tailles d'échantillon de strate. La fonction objectif (2.15) peut donc être réécrite comme suit :

$$\text{Minimiser } \sum_{j=1}^{m} w_j \frac{1}{Y_j^2}\left(\sum_{h=1}^{H} \frac{N_h^2}{n_h} S_{hj}^2\right). \tag{2.20}$$

La contrainte (2.16) est la même que la contrainte (2.8) appliquée dans la formulation A. La contrainte (2.17) permet de s'assurer que le coût variable total de l'enquête ne dépassera pas le budget alloué C. Comme la contrainte (2.10) de la formulation A, la contrainte (2.18) permet de s'assurer que toutes les tailles d'échantillon attribuées sont des nombres entiers. La contrainte (2.19) permet de s'assurer que les poids d'échantillonnage préférentiel d'importance sont appropriés pour l'agrégation des variances relatives des totaux estimatifs pour chacune des variables de l'enquête.

Lorsque les coûts d'enquête au niveau de l'unité c_h par strate sont inconnus ou peuvent être présumés égaux, la contrainte (2.17) peut être remplacée par $\sum_{h=1}^{H} n_h \leq n,$ où n est la taille d'échantillon globale (maximale).

Les formulations A et B présentent une non-linéarité : la contrainte (2.9) ou (2.14) dans la formulation A, et la fonction objectif dans la formulation B. La première solution possible pour résoudre le problème de non-linéarité dans ces deux formulations serait une des méthodes de programmation non linéaire ou de programmation convexe (Bazaraa, Sheralli et Shetty 2006; Luenberger et Ye 2008) qui peuvent traiter les contraintes, comme les méthodes fondées sur les pénalités et celles fondées sur les multiplicateurs. Quoi qu'il en soit, l'application de ces méthodes tend à produire des solutions (ensembles de tailles d'échantillon à répartir entre les strates) qui ne sont généralement pas des nombres entiers. De plus, lorsque les résultats de ces solutions sont arrondis afin d'obtenir des tailles d'échantillon pratiques, rien ne garantit l'obtention d'un optimum global (Wolsey 1998) en termes de minimisation des fonctions d'objectif correspondantes.

Ou bien, étant donné que les solutions (tailles d'échantillon) doivent être des nombres entiers, on pourrait envisager d'appliquer des méthodes de programmation en nombres entiers, comme *Branch and Bound* (Land et Doig 1960; Wolsey 1998; Wolsey et Nemhauser 1999). Cependant, la non-linéarité présente dans les deux formulations empêche l'application immédiate de ces méthodes.

Cela dit, dans la prochaine section, nous proposons deux nouvelles formulations pour la programmation en nombres entiers qui contournent ces problèmes et équivalent à la formulation A, définie à la fois par (2.7), (2.8), (2.9) et (2.10), et la formulation B, définie à la fois par (2.20), (2.16), (2.17), (2.18) et (2.19). Plus précisément, il est possible d'obtenir, à partir de la résolution de ces nouvelles formulations, des tailles d'échantillon qui sont des nombres entiers (n_h) pour la répartition des échantillons, qui satisfont aux contraintes établies pour chaque problème et qui aboutissent à un optimum global (Wolsey 1998), que ce soit pour la fonction objectif définie en (2.7) ou pour celle définie en (2.20).

3 Formulations proposées

Du point de vue de l'optimisation, pour résoudre les problèmes définis par (2.7)-(2.10) ou par (2.16)-(2.20), il faut déterminer $n_1, n_2, \ldots n_H$ à partir des ensembles définis par $A_h = \{n_{\min}, \ldots, N_h\}$, $h = 1, \ldots, H$, que les contraintes de chacun de ces problèmes sont satisfaites et que la fonction objectif correspondante est minimisée. Comme il est indiqué plus haut, une taille d'échantillon minimale standard par strate de $n_{\min} = 2$ est examinée ici pour définir les ensembles A_h, mais cette valeur peut être modifiée au besoin pour répondre aux exigences particulières d'une enquête donnée.

Selon cette approche, une nouvelle formulation peut être considérée où les variables décisionnelles sont des variables indicatrices à partir desquelles des éléments des ensembles A_h $(h = 1, \ldots, H)$ seront choisis. À cette fin, nous présentons la variable binaire x_{hk}, qui prend la valeur 1 si la taille d'échantillon $k \in A_h$ est attribuée à la strate h, et la valeur 0 si cette taille d'échantillon n'est pas attribuée à la strate $h, h = 1, \ldots, H$.

Compte tenu des formulations présentées plus haut et des nouvelles variables binaires, nous pouvons proposer deux formulations de la programmation en nombres entiers où les variables décisionnelles (c'est-à-dire les inconnus à déterminer) sont du type $0-1$, configurant ainsi un problème de programmation en nombres entiers binaires (Wolsey et Nemhauser 1999). La formulation équivalente à la formulation A est donnée par :

Formulation C

$$\text{Minimiser} \sum_{h=1}^{H} c_h \left(\sum_{k=n_{\min}}^{N_h} k \; x_{hk} \right) \tag{3.1}$$

$$\text{s.c.} \sum_{k=n_{\min}}^{N_h} x_{hk} = 1 \;\; \forall h = 1, \ldots, H \tag{3.2}$$

$$\sum_{h=1}^{H} N_h \; p_{hj} \left(\sum_{k=n_{\min}}^{N_h} \frac{x_{hk}}{k} \right) - \sum_{h=1}^{H} p_{hj} \leq 1, \;\; j = 1, \ldots, m \tag{3.3}$$

$$x_{hk} \in \{0,1\}, \quad k = n_{\min}, \ldots, N_h, h = 1, \ldots, H. \tag{3.4}$$

Dans la formulation C, la contrainte (3.2) permet de s'assurer que, pour chacune des strates, il y aura exactement une variable x_{hk} prenant la valeur 1. Cela revient à choisir une seule valeur k (taille d'échantillon) dans chaque ensemble $A_h\ (h = 1, \ldots, H)$. La contrainte (3.3) est équivalente à la contrainte (2.9) ou à son équivalent (2.14) dans la formulation A. Cette formulation envisage des coûts d'enquête unitaires variables pour les différentes strates. Si cela n'est pas nécessaire, la fonction objectif en (3.1) peut être redéfinie comme suit :

$$\text{Minimiser} \sum_{h=1}^{H} \sum_{k=n_{\min}}^{N_h} k\ x_{hk}. \tag{3.5}$$

L'exemple suivant aide à comprendre la formulation proposée.

Exemple 1 : Supposons trois strates de population $(H = 3)$ où $N_1 = 3, N_2 = 5$ et $N_3 = 4$, des coûts d'enquête unitaires égaux pour toutes les strates (disons $c_h = 1\ \forall h$) et une seule variable d'enquête $(m = 1)$. La formulation C se présenterait alors comme suit :

$$\text{Minimiser } 1\,x_{11} + 2\,x_{12} + 3\,x_{13} + 1\,x_{21} + 2\,x_{22} + 3\,x_{23} + 4\,x_{24} + 5\,x_{25} + 1\,x_{31} + 2\,x_{32} + 3\,x_{33} + 4\,x_{34} \tag{3.6}$$

$$\text{s.c. } x_{11} + x_{12} + x_{13} = 1 \tag{3.7}$$

$$x_{21} + x_{22} + x_{23} + x_{24} + x_{25} = 1 \tag{3.8}$$

$$x_{31} + x_{32} + x_{33} + x_{34} = 1 \tag{3.9}$$

$$N_1\ p_{11}\left(1\,x_{11} + \frac{1}{2}\,x_{12} + \frac{1}{3}\,x_{13}\right) - p_{11} + N_2\ p_{21}\left(1\,x_{21} + \frac{1}{2}\,x_{22} + \frac{1}{3}\,x_{23} + \frac{1}{4}\,x_{24} + \frac{1}{5}\,x_{25}\right) - p_{21} + N_3\ p_{31}\left(1\,x_{31} + \frac{1}{2}\,x_{32} + \frac{1}{3}\,x_{33} + \frac{1}{4}\,x_{34}\right) - p_{31} \le 1 \tag{3.10}$$

$$x_{11}, x_{12}, x_{13}, x_{21}, x_{22}, x_{23}, x_{24}, x_{25}, x_{31}, x_{32}, x_{33}, x_{34} \in \{0,1\}. \tag{3.11}$$

On peut aussi adapter la formulation B à cette nouvelle approche utilisant les variables binaires comme suit :

Formulation D

$$\text{Minimiser} \sum_{j=1}^{m} w_j \frac{1}{Y_j^2} \sum_{h=1}^{H} \left(\sum_{k=n_{\min}}^{N_h} \frac{x_{hk}}{k} \right) N_h^2 S_{hj}^2 \tag{3.12}$$

$$\text{s.c.} \sum_{k=n_{\min}}^{N_h} x_{hk} = 1 \quad \forall h = 1, \ldots, H \tag{3.13}$$

$$\sum_{h=1}^{H} c_h \left(\sum_{k=n_{\min}}^{N_h} k\ x_{hk} \right) \le C \tag{3.14}$$

$$x_{hk} \in \{0,1\},\ k = n_{\min},\ldots,N_h, h = 1,\ldots,H. \tag{3.15}$$

Dans la formulation D, la fonction objectif (3.12) est équivalente à la fonction objectif (2.20). La contrainte (3.13) est équivalente à la contrainte (2.16). La contrainte (3.14) est équivalente à la contrainte (2.17) et permet de s'assurer que le coût variable total de l'enquête ne dépassera pas le budget alloué C. Si nous ne connaissons pas les coûts d'enquête unitaires par strate ou si nous supposons qu'ils sont les mêmes dans toutes les strates, nous remplaçons la contrainte (3.14) par

$$\sum_{h=1}^{H}\sum_{k=n_{\min}}^{N_h} k\ x_{hk} \le n. \tag{3.16}$$

L'exemple qui suit illustre la formulation proposée.

Exemple 2 : Supposons deux strates de population $(H=2)$ où $N_1=3$ et $N_2=4$, deux variables d'enquête $(m=2)$, des coûts d'enquête unitaires égaux pour les deux strates, un poids d'échantillonnage préférentiel w_j égal à ½ pour les deux variables d'enquête et une taille d'échantillon totale de $n=5$. La formulation D se présenterait alors comme suit :

$$\begin{aligned}\text{Minimiser } &\left[x_{11}\frac{N_1^2}{1} + x_{12}\frac{N_1^2}{2} + x_{13}\frac{N_1^2}{3}\right]\frac{1}{2}\left(\frac{S_{11}^2}{Y_1^2}+\frac{S_{12}^2}{Y_2^2}\right)+\\ &\left[x_{21}\frac{N_2^2}{1} + x_{22}\frac{N_2^2}{2} + x_{23}\frac{N_2^2}{3} + x_{24}\frac{N_2^2}{4}\right]\frac{1}{2}\left(\frac{S_{21}^2}{Y_1^2}+\frac{S_{22}^2}{Y_2^2}\right)\end{aligned} \tag{3.17}$$

$$\text{s.c. } x_{11} + x_{12} + x_{13} = 1 \tag{3.18}$$

$$x_{21} + x_{22} + x_{23} + x_{24} = 1 \tag{3.19}$$

$$1\,x_{11} + 2\,x_{12} + 3\,x_{13} + 1\,x_{21} + 2\,x_{22} + 3\,x_{23} + 4\,x_{24} \le 5 \tag{3.20}$$

$$x_{11}, x_{12}, x_{13}, x_{21}, x_{22}, x_{23}, x_{24} \in \{0,1\}. \tag{3.21}$$

Dans cet article, nous avons résolu ces deux formulations en appliquant une méthode d'énumération implicite appelée *Branch and Bound*. Les méthodes *Branch and Bound* (Wolsey 1998; Wolsey et Nemhauser 1999) permettent de résoudre les problèmes de programmation en nombres entiers binaires de manière optimale et efficiente, en considérant la résolution d'un sous-ensemble de problèmes qui peuvent être résolus localement pour le problème. Ces méthodes ont été développées à partir du travail de pionnier effectué par Land et Doig (1960).

Nous avons résolu les formulations C et D de l'approche, appelée BSSM (Brito, Silva, Semaan et Maculan) au moyen du module R *Rglpk*. Le code R que nous avons développé est disponible sur demande. Le module *Rglpk* contient un ensemble de procédures qui peut servir à résoudre les problèmes de programmation linéaire et de programmation en nombres entiers.

À des fins de comparaison dans le cas de notre formulation C, nous avons également examiné dans nos illustrations numériques un algorithme proposé par (Bethel 1985 et 1989), qui est disponible dans le

module R *SamplingStrata*. Cet algorithme se fonde sur le théorème de Kuhn-Tucker et utilise les multiplicateurs de Lagrange (Bazaraa et coll. 2006). Dans le cas de notre formulation D, nous avons comparé notre approche à une « méthode classique » proposée dans Cochran (1977, section 5.A.4), comme le suggérait le rédacteur adjoint.

4 Résultats numériques

Cette section présente des résultats en vue de l'application des approches de répartition multivariée optimale sélectionnées à un ensemble de base de données de population. Les approches examinées comprennent :

- les algorithmes BSSM développés pour résoudre les formulations C et D présentées dans la section 3;
- une version améliorée de l'algorithme de Bethel (Bethel 1989) développée par Ballin et Barcarolli (2008);
- la méthode classique proposée dans Cochran (1977, section 5.A.4).

Onze ensembles de données sur la population ont été utilisés à des fins d'illustration numérique, mais pour des raisons d'espace, nous présentons ici les résultats pour seulement trois des populations. Les trois populations sélectionnées sont décrites aux tableaux A1 à A6 de l'annexe A. Le tableau A1 fournit une brève description de chaque population d'enquête ainsi que la liste des variables d'enquête correspondantes. Le tableau A2 explique comment nous avons stratifié chaque population avant de déterminer la répartition optimale. En particulier, pour l'ensemble de données d'enquête appelé *MunicSw* les strates avaient déjà été définies. Les deux autres populations ont été stratifiées à l'aide d'un algorithme de stratification disponible dans le module R *stratification* ou d'une méthode de mise en grappes $k-\text{means}$ classique disponible dans le module *base* de R.

Le tableau A3 présente le nombre de strates de population (H), le nombre de variables d'enquête (m), et la taille de population (N) pour chacune des populations examinées. Les tableaux A4 à A6 fournissent les chiffres de population, les moyennes et les écarts types par strate pour les variables d'enquête examinées dans chacune des trois populations observées.

Les résultats de toutes les expériences numériques décrites dans cet article ont été obtenus à l'aide des modules R et des fonctions mentionnées, sur un ordinateur de bureau fonctionnant sous Windows 7 avec 24 Go de mémoire vive et 8 processeurs i7 de 3,40 GHz. Le temps de traitement allait de quelques millisecondes (pour la population *MunicSw*, relativement petite) à moins de quatre secondes (pour la population *SchoolsNortheast* plus nombreuse, sous la formulation C). Cela démontre que les formulations proposées offrent une solution de rechange pratique et efficace aux problèmes de répartition multivariée optimale de petite et moyenne tailles, pour les populations de tailles (N) qui comptent des milliers et même des dizaines de milliers d'unités.

Les tableaux 4.1 à 4.3 fournissent les coefficients de variation $\left(\text{CV}_j\right)$ cibles pour chaque variable d'enquête, les tailles d'échantillon obtenues en utilisant l'algorithme pour résoudre la formulation C (n_{BSSM}) proposée et l'algorithme de Bethel (n_{Bethel}), et les coefficients de variation obtenus pour les

estimateurs des totaux des variables d'enquête examinées dans chaque population sous les deux algorithmes comparés.

Tableau 4.1
Résultats pour la population *CoffeeFarms*

	Algorithme pour la formulation C				Algorithme de Bethel			
CV_j (%)	n_{BSSM}	$CV(t_1)$ (%)	$CV(t_2)$ (%)	$CV(t_3)$ (%)	n_{Bethel}	$CV(t_1)$ (%)	$CV(t_2)$ (%)	$CV(t_3)$ (%)
5	**2 545**	1,24	5,00	2,92	2 546	1,23	5,00	2,91
10	**754**	3,30	10,00	7,01	755	3,30	9,99	7,07
15	**347**	5,21	15,00	11,01	349	5,11	14,95	10,85

Tableau 4.2
Résultats pour la population *SchoolsNortheast*

	Algorithme pour la formulation C			Algorithme de Bethel		
CV_j (%)	n_{BSSM}	$CV(t_1)$ (%)	$CV(t_2)$ (%)	n_{Bethel}	$CV(t_1)$ (%)	$CV(t_2)$ (%)
2	**1 624**	2,00	1,79	1 628	2,00	1,78
5	**294**	5,00	4,31	299	4,96	4,23
10	**80**	9,93	8,24	83	9,72	8,13

Tableau 4.3
Résultats pour la population *MunicSw*

	Algorithme pour la formulation C					Algorithme de Bethel				
CV_j (%)	n_{BSSM}	$CV(t_1)$ (%)	$CV(t_2)$ (%)	$CV(t_3)$ (%)	$CV(t_4)$ (%)	n_{Bethel}	$CV(t_1)$ (%)	$CV(t_2)$ (%)	$CV(t_3)$ (%)	$CV(t_4)$ (%)
5	**1 527**	2,01	3,88	5,00	4,41	1 529	2,00	3,88	4,99	4,40
10	**761**	3,61	7,27	9,99	8,77	763	3,60	7,25	9,97	8,75
15	**439**	5,01	10,22	14,98	13,07	441	4,95	10,16	14,94	13,03

Comme prévu, dans tous les cas, les tailles d'échantillon obtenues par résolution de la formulation C étaient inférieures (en gras) ou égales à celles obtenues en utilisant l'algorithme de Bethel. Cependant, les améliorations n'étaient généralement pas substantielles. L'algorithme proposé a quand même permis d'améliorer la meilleure méthode actuelle dans les neuf scénarios examinés (trois populations multipliées par trois niveaux pour les CV cibles). Les améliorations semblaient un peu plus importantes pour la population *SchoolsNortheast* où le nombre de strates est plus élevé. Des résultats semblables (non présentés ici pour des raisons de concision, mais disponibles auprès des auteurs sur demande) ont été obtenus pour les huit autres populations examinées dans une version initiale de cet article.

Les tableaux 4.4 à 4.6 présentent les résultats de l'application de la formulation D et de la méthode classique proposée dans Cochran (1977, section 5.A.4) aux trois mêmes populations observées. L'objectif est maintenant de minimiser la variance relative pondérée des estimations HT du total, tout en maintenant la taille d'échantillon globale ou le coût. La première ligne de chacun de ces tableaux contient les tailles d'échantillon totales prises en considération pour la répartition. Ces tailles d'échantillon correspondent aux fractions d'échantillonnage de 10 %, 20 % et 30 % des tailles de population correspondantes (N) apparaissant sur la deuxième ligne de chacun des tableaux. Les lignes suivantes présentent la répartition

de l'échantillon total entre les strates, les coefficients de variation obtenus pour les estimations HT des totaux des variables d'enquête compte tenu de la répartition, et la somme des coefficients de variation $(\Sigma \mathrm{CV}(t_i))$, qui est une mesure agrégée de l'efficience pour toutes les variables d'enquête.

Les poids d'échantillonnage préférentiel étaient considérés comme égaux pour toutes les variables d'enquête, et les coûts d'enquête unitaires étaient considérés comme égaux dans toutes les strates, dans chaque population, pour ces applications.

Tableau 4.4
Résultats pour la population *CoffeeFarms*

n	**2,047**		**4,094**		**6,142**	
Fraction d'échantillonnage	**10 %**		**20 %**		**30 %**	
Résultat	**BSSM-D**	**Classique**	**BSSM-D**	**Classique**	**BSSM-D**	**Classique**
n_1	1 174	1 124	2 483	2 340	3 792	3 625
n_2	662	737	1 400	1 544	2 139	2 306
n_3	211	186	211	210	211	211
$\mathrm{CV}(t_1)$	**1,02**	1,14	**0,62**	0,62	**0,42**	0,42
$\mathrm{CV}(t_2)$	**5,78**	5,79	**3,62**	3,65	**2,61**	2,63
$\mathrm{CV}(t_3)$	**2,86**	2,98	**1,73**	1,73	1,19	**1,17**
$\Sigma\mathrm{CV}(t_i)$	**9,66**	9,91	**5,97**	6,00	**4,22**	4,22

Tableau 4.5
Résultats pour la population *SchoolsNortheast*

n	**7 508**		**15 017**		**22 525**	
Fraction d'échantillonnage	**10 %**		**20 %**		**30 %**	
Résultat	**BSSM-D**	**Classique**	**BSSM-D**	**Classique**	**BSSM-D**	**Classique**
n_1	82	58	82	60	82	66
n_2	36	33	62	53	53	62
n_3	7	6	7	6	7	6
n_4	206	214	465	433	771	611
n_5	1 083	1 000	2 091	1 962	2 671	2 121
n_6	447	452	891	914	1 428	1 436
n_7	361	371	711	750	1 182	1 175
n_8	2 995	2 989	5 963	6 055	9 088	9 634
n_9	976	1 023	1 965	2 069	3 078	3 229
n_{10}	399	419	800	849	1 331	1 338
n_{11}	797	813	1 742	1 647	2 596	2 612
n_{12}	119	130	238	219	238	235
$\mathrm{CV}(t_1)$	**0,86**	0,98	**0,54**	0,69	**0,39**	0,54
$\mathrm{CV}(t_2)$	0,73	**0,72**	**0,47**	0,47	0,35	**0,34**
$\Sigma\mathrm{CV}(t_i)$	**1,59**	1,70	**1,01**	1,16	**0,74**	0,88

Tableau 4.6
Résultats de la formulation D pour la population *MunicSw*

n	290		579		869	
Fraction d'échantillonnage	10 %		20 %		30 %	
Résultat	**BSSM-D**	**Classique**	**BSSM-D**	**Classique**	**BSSM-D**	**Classique**
n_1	67	59	134	118	202	182
n_2	68	77	136	153	206	233
n_3	40	35	80	70	120	107
n_4	58	47	116	93	171	128
n_5	32	43	65	85	97	129
n_6	16	21	31	43	47	65
n_7	9	8	17	17	26	25
$\mathbf{CV}(t_1)$	5,93	**5,40**	4,01	**3,61**	3,10	**2,75**
$\mathbf{CV}(t_2)$	12,53	**12,24**	8,36	**8,12**	6,36	**6,14**
$\mathbf{CV}(t_3)$	**19,49**	20,19	**12,46**	13,01	**8,95**	9,56
$\mathbf{CV}(t_4)$	**16,91**	17,45	**10,85**	11,27	**7,84**	8,30
$\mathbf{\Sigma CV}(t_i)$	**54,86**	55,28	**35,68**	36,01	**26,25**	26,75

Comme prévu, dans les trois cas, les coefficients de variation obtenus en résolvant la formulation D étaient inférieurs (en gras) à ceux obtenus en utilisant l'algorithme classique. Cependant, l'algorithme classique produisait des CV plus petits pour certaines des variables d'enquête, particulièrement pour la population *MunicSw* Les améliorations n'étaient généralement pas très importantes, mais elles étaient là encore un peu plus marquées pour la population *SchoolsNortheast*. Dans cette comparaison, cependant, les répartitions obtenues sont très différentes selon la méthode employée.

5 Observations finales

Dans cet article, nous avons proposé deux nouvelles formulations permettant d'obtenir le minimum global dans les problèmes de répartition multivariée optimale. On peut appliquer ces formulations exactes de la programmation en nombres entiers de façon efficace en utilisant un logiciel commercial (à savoir le module R *Rglpk*). De plus, les formulations proposées permettent de définir les tailles d'échantillon minimales par strate, ce qui est très utile dans la pratique pour éviter les répartitions avec des tailles d'échantillon inférieures à 2, par exemple, qui rendraient difficile l'estimation de la variance. Ces tailles d'échantillon minimales peuvent être fixées à des valeurs plus élevées (par exemple 5, 10, 30 ou un autre chiffre) afin de s'assurer que les échantillons sont assez grands pour tolérer certains cas de non-réponse ou qu'une estimation est possible pour chaque strate si les strates sont utilisées comme domaines d'estimation.

L'approche proposée améliore les méthodes existantes en s'attaquant directement au problème de répartition et en tenant compte de la non-linéarité de la fonction objectif ou des contraintes, ainsi que de l'exigence selon laquelle les tailles d'échantillon pour les strates doivent être des nombres entiers. Dans la

littérature sur ce sujet, les méthodes antérieures ne garantissent pas l'obtention d'un optimum global ou elles produisent des répartitions à valeur réelle qui doivent être arrondies à des nombres entiers.

Dans la pratique, les répartitions à valeur réelle ne constituent pas un problème majeur, à moins que les tailles de population par strate N_h soient très petites ou que le nombre de strates soit très élevé. Dans le premier cas, l'échantillonnage d'une unité de plus ou de moins peut faire une grande différence dans les fractions d'échantillonnage, ce qui peut avoir d'importantes incidences sur les variances. Dans le deuxième cas, l'arrondissement des tailles d'échantillon attribuées peut faire une différence dans la taille d'échantillon totale n. Lorsque toutes les tailles de population par strate N_h sont relativement grandes et que le nombre de strates est raisonnable, l'arrondissement des tailles d'échantillon qui ne sont pas des nombres entiers ne cause pas de problème.

Dans cet article, nous avons effectué quelques calculs visant essentiellement à démontrer la faisabilité de l'approche proposée. La formulation C de l'approche proposée permet d'obtenir des résultats comparables à ceux obtenus avec la méthode de Bethel, en plus de produire des répartitions à valeurs entières qui correspondent à l'optimum global. Cependant, comme peu de différences ont été constatées entre la méthode de BSSM et celle de Bethel dans les applications examinées, il n'y aurait guère d'avantages à adopter la méthode de BSSM. Les résultats obtenus avec la formulation D représentaient des améliorations modestes par rapport à ceux obtenus avec la méthode classique employée dans la comparaison.

D'autres recherches sont requises pour tester l'approche face à des problèmes plus importants et pour en évaluer les mérites par rapport à d'autres méthodes dans d'autres scénarios pratiques. Un avantage important de l'approche proposée est qu'on peut appliquer les deux formulations en utilisant un logiciel commercial, comme il est expliqué plus haut.

Remerciements

Cette étude a été financée par la subvention de recherche E-26/111.947/2012 de FAPERJ.

Annexe A

Description des populations d'enquête examinées dans l'expérience numérique

Tableau A1
Description des populations

Population	Description	Variables d'enquête (y)
CoffeeFarms	Plantations de café dans l'État de Paraná, au Brésil, d'après le recensement agricole de 1996.	Nombre de caféiers Superficie agricole totale Production de café
SchoolsNortheast	Données tirées du recensement des écoles de 2012, par école, région du Nord-Est du Brésil.	Nombre de salles de classe Nombre d'employés
MunicSw	Données sur les municipalités suisses tirées du module *SamplingStrata*.	Superficie agricole Superficie industrielle Nombre de ménages Population

Tableau A2
Stratification des populations

Population	Stratification
CoffeeFarms	Stratification en fonction de la variable Nombre de caféiers, en utilisant l'algorithme de Kozak disponible dans le module *Stratification*
SchoolsNortheast	Douze strates ont été formées en tenant compte du type d'école (4 catégories) et du nombre d'élèves (3 catégories). La stratification par taille des écoles a été effectuée en utilisant l'algorithme de mise en grappes $k-\text{means}$ à l'intérieur de chaque type d'école.
MunicSw	Cette population est disponible dans le module *SamplingStrata* et les strates correspondent aux régions de la Suisse.

Tableau A3
Nombre de strates, nombre de variables d'enquête et taille totale pour les populations d'enquête examinées

Population	H	m	N
CoffeeFarms	3	3	20 472
SchoolsNortheast	12	2	75 084
MunicSw	7	4	2 896

Tableau A4
Résumés de la population par strate – *CoffeeFarms*

Résumé	Strate		
	$h=1$	$h=2$	$h=3$
N_h	17 821	2 440	211
$\bar{Y}_{1h}$	4 291	26 688	218 712
$\bar{Y}_{2h}$	22	84	488
$\bar{Y}_{3h}$	2 671	13 204	129 033
S_{h1}	2 873	15 541	193 366
S_{h2}	69	262	583
S_{h3}	4 611	24 704	200 447

Tableau A5
Résumés de la population par strate – *SchoolsNortheast*

Strate	N_h	$\bar{Y}_{1h}$	$\bar{Y}_{2h}$	S_{h1}	S_{h2}
$h=1$	82	45,1	54,0	309,2	24,9
$h=2$	63	23,9	146,3	14,4	92,6
$h=3$	7	80,9	700,4	29	342,5
$h=4$	783	16,2	95,7	6,4	49,5
$h=5$	2 676	10,9	57,7	21,6	23,7
$h=6$	3 958	6,1	26,7	4,2	17,9
$h=7$	2 172	13,6	76,8	5,7	27,9
$h=8$	45 243	2,5	9,3	3	8,8
$h=9$	9 674	7,7	38,0	3,2	17,9
$h=10$	1 743	17,3	49,1	9,2	36,7
$h=11$	8 445	7,3	15,3	4,1	13,5
$h=12$	238	37,7	140,8	18,4	88,9

Tableau A6
Résumés de la population par strate – *MunicSw*

Résumé	Strate						
	$h = 1$	$h = 2$	$h = 3$	$h = 4$	$h = 5$	$h = 6$	$h = 7$
N_h	589	913	321	171	471	186	245
$\bar{Y}_{1h}$	262,5	367,2	262,7	438,0	429,5	668,9	47,0
$\bar{Y}_{2h}$	5,5	5,3	9,7	13,3	7,9	11,0	4,1
$\bar{Y}_{3h}$	963,9	782,1	1 345,2	3 319,1	906,0	1 465,2	550,7
$\bar{Y}_{4h}$	2 252,5	1 839,4	3 099,5	7 297,7	2 226,0	3 675,8	1 252,4
S_{h1}	220,5	342,4	173,2	290,2	414,2	568,7	65,3
S_{h2}	15,1	13,0	19,4	29,7	14,9	15,5	8,2
S_{h3}	4 600,9	2 794,7	5 003,5	14 610,0	2 178,6	2 802,1	1 197,5
S_{h4}	9 540,3	5 621,6	9 764,5	28 589,4	4 759,4	5 914,5	2 514,9

Bibliographie

Ballin, M., et Barcaroli, G. (2008). Optimal stratification of sampling frames in a multivariate and multidomain sample design. *Contributi ISTAT*, 10.

Bazaraa, M.S., Sherali, H.D. et Shetty, C.M. (2006). *Nonlinear Programming: Theory and Algorithms*. New York : John Wiley & Sons, Inc, Third Edition.

Bethel, J. (1985). An optimum allocation algorithm for multivariate surveys. *Proceedings of the Survey Research Methods Section,* American Statistical Association, 209-212.

Bethel, J. (1989). Répartition de l'échantillon dans les enquêtes à plusieurs variables. *Techniques d'enquête*, 15, 1, 49-60.

Chromy, J. (1987). Design optimization with multiple objectives. *Proceedings of the Survey Research Methods Section,* American Statistical Association, 194-199.

Cochran, W.G. (1977). *Sampling Techniques*. Third Edition – Wiley.

Day, C.D. (2010). A multi-objective evolutionary algorithm for multivariate optimal allocation. *Proceedings of the Survey Research Methods Section,* American Statistical Association.

Folks, J.L., et Antle, C.E. (1965). Optimum allocation of sampling units to strata when there are R responses of interest. *Journal of the American Statistical Association*, 60 (309), 225-233.

García, J.A.D., et Cortez, L.U. (2006). Optimum allocation in multivariate stratified sampling: Multi-objective programming. *Comunicaciones Del Cimat*, no I-06-07/28-03-2006.

Huddleston, H.F., Claypool, P.L. et Hocking, R.R. (1970). Optimal sample allocation to strata using convex programming. *Journal of the Royal Statistical Society, Series C*, 19 (3).

Ismail, M.V., Nasser, K. et Ahmad, Q.S. (2011). Solution of a multivariate stratified sampling problem through Chebyshev's Goal programming. *Pakistan Journal of Statistics and Operation Research*, vol. vii, 1, 101-108.

Khan, M.G.M., et Ahsan, M.J. (2003). A note on optimum allocation in multivariate stratified sampling. *The South Pacific Journal of Natural Science*, 21, 91-95.

Khan, M.F., Ali, I. et Ahmad, Q.S. (2011). Chebyshev approximate solution to allocation problem in multiple objective surveys with random costs. *American Journal of Computational Mathematics*, 1, 247-251.

Khan, M.F., Ali, I., Raghav, Y.S. et Bari, A. (2012). Allocation in multivariate stratified surveys with non-linear random cost function. *American Journal of Operations Research*, 2, 100-105.

Kish, L. (1976). Optima and proxima in linear sample designs. *Journal of the Royal Statistical Society, Series A*, 139 (1), 80-95.

Kokan, A.R. (1963). Optimum allocation in multivariate surveys. *Journal of the Royal Statistical Society, Series A*, 126 (4), 557-565.

Kokan, A.R., et Khan, S. (1967). Optimum allocation in multivariate surveys: An analytical solution. *Journal of the Royal Statistical Society, Series B*, 29 (1), 115-125.

Kozak, M. (2006). Multivariate sample allocation: Application of random search method. *Statistics in Transition*, 7 (4), 889-900.

Land, A.H., et Doig, A.G. (1960). An Automatic method for solving discrete programming problems. *Econometrica*, 28 (3), 497-520.

Lohr, S.L. (2010). *Sampling: Design and Analysis*, Second edition. Brooks/Cole, Cengage Learning.

Luenberger, D.G., et Ye, Y. (2008*). Linear and Non-Linear Programming*, Third Edition. Springer.

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

Valliant, R., et Gentle, J.E. (1997). An application of mathematical programming to sample allocation. *Computational Statistics & Data Analysis*, 25, 337-360.

Wolsey, L.A. (1998). *Integer Programming*. Wiley-Interscience Series in Discrete Mathematics and Optimization.

Wolsey, L.A., et Nemhauser, G.L. (1999). *Integer and Combinatorial Optimization*. Wiley-Interscience Series in Discrete Mathematics and Optimization.