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

Techniques d’enquête

Un algorithme d’optimisation appliqué au problème de stratification unidimensionnelle

par José André de Moura Brito, Tomás Moura da Veiga et
Pedro Luis do Nascimento Silva

Date de diffusion : le 27 juin 2019

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 suivants :

- Service de renseignements statistiques 1-800-263-1136
- Service national d'appareils de télécommunications pour les malentendants 1-800-363-7629
- Télécopieur 1-514-283-9350

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.

Techniques d'enquête, juin 2019
Vol. 45, nº 2, pp. 317-337
Statistique Canada, nº 12-001-X au catalogue

Un algorithme d'optimisation appliqué au problème de stratification unidimensionnelle

José André de Moura Brito, Tomás Moura da Veiga et Pedro Luis do Nascimento Silva[1]

Résumé

Ce document présente un nouvel algorithme pour résoudre le problème de stratification unidimensionnelle optimale, lequel se ramène à une détermination des bornes de strate. Lorsque le nombre de strates *H* et la taille totale de l'échantillon *n* sont fixes, on obtient les bornes de strate en minimisant la variance de l'estimateur d'un total pour la variable de stratification. C'est un algorithme qui fait appel à la métaheuristique de l'algorithme génétique biaisé à clés aléatoires (BRKGA) pour trouver la solution optimale. Il a été démontré que cette métaheuristique produit des solutions de bonne qualité à de nombreux problèmes d'optimisation à un prix modeste en temps de calcul. L'algorithme est mis en œuvre dans le package *stratbr* en *R* disponible à partir de CRAN (de Moura Brito, do Nascimento Silva et da Veiga, 2017a). Nous livrons des résultats numériques pour un ensemble de 27 populations, ce qui permet de comparer le nouvel algorithme à certaines méthodes rivales figurant dans la documentation spécialisée. L'algorithme est d'un meilleur rendement que les méthodes plus simples par approximation. Il est également supérieur à quelques autres approches en optimisation. Il est égal en rendement à la meilleure technique d'optimisation que l'on doit à Kozak (2004). Son principal avantage sur la méthode de Kozak réside dans le couplage de la stratification optimale avec la répartition optimale que proposent de Moura Brito, do Nascimento Silva, Silva Semaan et Maculan (2015), d'où l'assurance que, si les bornes de stratification obtenues atteignent l'optimum global, la solution dégagée dans l'ensemble sera aussi l'optimum global pour les bornes de stratification et la répartition de l'échantillon.

Mots-clés : Stratification optimale; algorithme génétique; programmation en nombres entiers; optimisation non linéaire; algorithme métaheuristique BRKGA.

1 Introduction

L'échantillonnage stratifié est une méthode largement employée pour accroître l'efficacité des plans d'échantillonnage. Les études abondent sur la stratification optimale (qui sera examinée plus loin dans le présent document), ce qui témoigne à la fois de l'importance de ce sujet pour les chercheurs et de sa vaste gamme d'applications. Récemment, Hidiroglou et Kozak (2017) ont comparé des méthodes d'optimisation et d'approximation pour la stratification unidimensionnelle de populations asymétriques pour conclure que les méthodes d'optimisation sont supérieures et devraient s'employer dans la pratique.

Nous proposons d'appliquer un nouvel algorithme d'optimisation pour établir les bornes de strate, ce que nous combinons à une méthode globalement optimale de répartition de taille d'échantillon entre les strates définies. Nous traitons le problème de stratification unidimensionnelle au moyen d'une technique d'optimisation globale (métaheuristique) appelée algorithme génétique biaisé à clés aléatoires (BRKGA), laquelle a été proposée par Gonçalves et Resende (2011). Cette technique n'est pas garante d'un optimum global pour les bornes de strate, mais on a démontré qu'elle engendre des solutions de bonne qualité pour de nombreux problèmes d'optimisation à un prix modeste en temps de calcul (voir Gonçalves et Resende, 2004; Gonçalves, Mendes et Resende, 2005; Festa, 2013; Oliveira, Chaves et Lorena, 2017).

Notre méthode de répartition de l'échantillon en fonction d'une stratification définie (voir de Moura Brito et coll., 2015), c'est-à-dire d'une stratification avec une variable spécifiée et un nombre donné de strates, est fondée sur une formulation en programmation en nombres entiers et dégage toujours

1. José André de Moura Brito, Escola Nacional de Ciências Estatísticas, Rua André Cavalcanti, 106, Centro, RJ Rio de Janeiro, 20231-050. Courriel : jambrito@gmail.com; Tomás Moura da Veiga, SHIN CA 08 Lote 01 Torre 01, 101 Lago Norte, district fédéral de Brasília, 71503-508. Courriel : tmvtomas@gmail.com; Pedro Luis do Nascimento Silva, Escola Nacional de Ciências Estatísticas, Rua André Cavalcanti, 106 Centro, Rio de Janeiro, RJ, 20231-050. Courriel : pedronsilva@gmail.com.

un optimum global en minimisant soit la taille totale de l'échantillon en tenant compte de contraintes de précision, soit la variance pour une taille totale d'échantillon fixe, tout en assurant une répartition de l'échantillon en nombres entiers et en permettant la spécification des bornes inférieure et supérieure de taille d'échantillon par strate, comme on a souvent à le faire dans les applications pratiques. Cette méthode est appliquée avec le package *stratbr* en R (voir de Moura Brito et coll., 2017a), constituant ainsi une solution de rechange pratique aux méthodes existantes par approximation et se révélant clairement plus efficace. Elle se compare aussi favorablement à d'autres méthodes d'optimisation qui ne garantissent pas une répartition optimale en fonction de la stratification.

Nous avons comparé cette nouvelle méthode à celles que proposent Dalenius et Hodges (1959), Gunning et Horgan (2004), Kozak (2004, 2006), Keskintürk et Er (2007) et de Moura Brito, Silva Semaan, Fadel et Brito (2017b) à l'aide de 27 populations d'enquête réelles ou artificielles. Notre étude empirique est bien plus large que celle d'Hidiroglou et Kozak (2017), qui n'ont utilisé que deux populations dans leur comparaison. Elle est aussi plus large que les autres études du passé.

Nous n'avons pas envisagé – comme il est indiqué – de comparer notre méthode aux arbres de classification ou de régression ou à d'autres algorithmes d'apprentissage machine qui synthétisent une ou plusieurs covariables en formant des groupes pouvant servir de strates. La grande raison en est que, avec de telles méthodes, on ne considère pas la variance de l'estimateur de l'échantillon cible ni à la taille d'échantillon en tenant compte de contraintes de précision en tant que critères d'optimisation. Il leur serait donc impossible de dégager l'optimum dans le problème que nous désirons traiter. Précisons que, dans les arbres de classification ou de régression, l'analyste doit toujours spécifier une « variable de réponse » s'ajoutant aux variables prédictives ou auxiliaires. Dans bien des cas types d'échantillonnage, l'analyste n'a pas accès à des données sur une telle « variable de réponse » et doit plutôt viser à minimiser la variance de l'estimateur pour la taille totale ou la variable de stratification (comme c'est le cas dans la plupart des études consacrées à ce thème).

Nous avons seulement considéré le « problème de stratification unidimensionnelle », signifiant qu'une seule mesure de taille est utilisée pour la stratification, mais il est toujours possible d'employer un modèle prédictif ou une autre technique de réduction de variable pour récapituler les variables ou les covariables auxiliaires en une variable unique x ou variable de taille aux fins de la méthode que nous proposons. Notre approche pourrait néanmoins être étendue à une stratification multidimensionnelle avec répartition optimale selon la nature des composantes de l'approche.

Nous avons structuré notre article de la manière suivante : la section 2 énonce les concepts clés de l'échantillonnage stratifié; la section 3 décrit en détail le problème de stratification; la section 4 présente l'algorithme génétique biaisé à clés aléatoires (BRKGA) et son application nouvelle à notre problème de stratification en combinaison avec la méthode de répartition optimale proposée par de Moura Brito et coll. (2015); la section 5 livre les résultats de l'application de la méthode proposée par rapport à cinq autres méthodes figurant dans la documentation spécialisée, comme nous l'avons indiqué; la section 6 tire enfin des conclusions de cette analyse comparative.

2 Échantillonnage stratifié

Dans un échantillonnage stratifié, la première étape consiste à répartir les éléments de la population cible entre des sous-groupes bien définis, de préférence homogènes, mutuellement exclusifs et exhaustifs appelés

strates. Chaque élément (unité) de la population est visé par l'enquête et fournit l'information recherchée. Les unités d'enquête peuvent être les ménages, les particuliers, les exploitations agricoles, les établissements et entreprises de commerce, etc.

On recommande de stratifier l'échantillon pour plusieurs raisons :

- on peut accroître la précision des estimations de population dans l'ensemble à un coût total fixe;
- on peut contrôler les tailles d'échantillon et la précision des estimations des strates, s'il y a lieu;
- on peut favoriser un équilibre de la charge de travail;
- on peut répartir les frais de déplacement entre les éléments d'enquête, si la stratification comporte une dimension géographique.

Lorsque la formation des strates est telle que la variabilité intrastrate est légère pour un jeu clé de variables, la stratification est jugée efficace, car elle mène à une meilleure précision des estimations par rapport aux autres plans de stratification.

La figure 2.1 présente la notation de base de notre exposé. Dans un échantillonnage stratifié, la population U se répartit en $H > 1$ sous-populations non vides appelées strates (Lohr, 2010), d'une taille $N_1, N_2, \ldots, N_h, \ldots, N_H$. Ces sous-populations ne se recouvrent pas et leur réunion forme la population entière :

$$N_1 + N_2 + \ldots + N_H = N. \tag{2.1}$$

$U = \{1, 2, \ldots, N\}$ – Ensemble des éléments de la population cible;

N – Nombre d'éléments ou taille de la population;

n – Nombre d'éléments ou taille de l'échantillon;

$H > 1$ – Nombre total de strates;

h – Indice de strate;

U_h – Ensemble des éléments de la strate satisfaisant $U_h \subset U$ et $U_h \neq \emptyset$;

N_h – Nombre d'éléments ou taille de population dans la strate h;

s_h – Ensemble des éléments échantillonnés dans la h^e strate – $s_h \subset U_h$;

n_h – Nombre d'éléments ou taille d'échantillon dans la strate h;

y_i – Valeur de la variable d'enquête pour l'élément de la population $i\,(i \in U)$;

x_i – Valeur de la variable de stratification pour l'élément de la population $i\,(i \in U)$;

$Y_h = \sum_{i \in U_h} y_i$ – Total de la variable d'enquête y dans la strate h;

$X_h = \sum_{i \in U_h} x_i$ – Total de la variable de stratification x dans la strate h;

$\bar{Y}_h = Y_h / N_h$ – Moyenne de population de la variable d'enquête y dans la strate h;

$\bar{X}_h = X_h / N_h$ – Moyenne de population de la variable de stratification x dans la strate h.

Figure 2.1 Notation utilisée dans l'étude.

Cochran (1977) énumère les facteurs influant sur l'efficacité d'un plan d'échantillonnage stratifié : choix de la (des) variable(s) de stratification; nombre de strates (H); délimitation des strates; taille totale de l'échantillon (n); répartition de l'échantillon total entre les strates; mode de sélection dans l'échantillonnage intrastrate. On définit les strates à l'aide d'une ou de plusieurs variables dont les valeurs sont connues pour chaque élément de la population. Dans la séquence, il y a sélection indépendante dans l'échantillonnage de chacune des H strates. Les tailles d'échantillon des strates sont telles que $n_1 + n_2 + \ldots + n_H = n$.

Il y a échantillonnage aléatoire simple stratifié (EASS) là où l'échantillonnage intrastrate se fait par sélection aléatoire simple sans remise. C'est le cas où l'estimateur de Horvitz-Thompson (HT) du total de la population (dans l'ensemble) $Y = \sum_{h=1}^{H} Y_h$ est donné par Cochran (1977) sous la forme suivante :

$$\hat{Y}_{\text{EASS}} = \sum_{h=1}^{H} N_h \bar{y}_h \tag{2.2}$$

où $\bar{y}_h = \sum_{i \in s_h} y_i / n_h$ est la moyenne d'échantillon des éléments échantillonnés dans la strate h.

La variance d'échantillonnage (var) et le coefficient de variation (CV) de l'estimateur $\hat{Y}_{\text{EASS}}$ se formulent respectivement ainsi :

$$\text{Var}\left(\hat{Y}_{\text{EASS}}\right) = \sum_{h=1}^{H} N_h \left(N_h - n_h\right) S_{hy}^2 / n_h \tag{2.3}$$

$$\text{CV}\left(\hat{Y}_{\text{EASS}}\right) = \sqrt{\text{Var}\left(\hat{Y}_{\text{EASS}}\right)} \Big/ Y \tag{2.4}$$

où $S_{hy}^2 = \sum_{i \in U_h} \left(y_i - \bar{Y}_h\right)^2 / \left(N_h - 1\right)$ est la variance de population de la variable d'enquête y dans la strate h.

On définit des quantités analogues de l'estimateur HT pour le total $X = \sum_{h=1}^{H} X_h$ de la variable de stratification x :

$$\hat{X}_{\text{EASS}} = \sum_{h=1}^{H} N_h \bar{x}_h \tag{2.5}$$

$$\text{Var}\left(\hat{X}_{\text{EASS}}\right) = \sum_{h=1}^{H} N_h \left(N_h - n_h\right) S_{hx}^2 / n_h \tag{2.6}$$

$$\text{CV}\left(\hat{X}_{\text{EASS}}\right) = \sqrt{\text{Var}\left(\hat{X}_{\text{EASS}}\right)} \Big/ X \tag{2.7}$$

où $\bar{x}_h = \sum_{i \in s_h} x_i / n_h$ est la moyenne d'échantillon de x pour les éléments échantillonnés dans la strate h et où $S_{hx}^2 = \sum_{i \in U_h} \left(x_i - \bar{X}_h\right)^2 / \left(N_h - 1\right)$ est la variance de population de la variable x dans la même strate h.

3 Le problème de stratification unidimensionnelle

Considérons le vecteur de population $X_U = \{x_1, x_2, \ldots, x_N\}$ correspondant à la variable de stratification x. Sans perte de généralité, nous posons que les éléments de population dans U sont ordonnés par la variable de stratification, de sorte que $x_1 \leq x_2 \leq \ldots \leq x_N$. Nous faisons intervenir les bornes de strate pour définir les H strates par la règle suivante :

1) $U_1 = \{i \in U \mid x_i \leq b_1\}$;

2) $U_h = \{i \in U \mid b_{h-1} < x_i \leq b_h\}$ pour $h = 2, 3, \ldots, H-1$;

3) $U_H = \{i \in U \mid b_{H-1} < x_i\}$.

Le problème de stratification se ramène à une détermination des points de démarcation, c'est-à-dire des bornes de strate $b_1 < b_2 < \ldots < b_h < \ldots < b_{H-1}$ avec minimisation de la variance (ou du CV d'une manière équivalente) de l'estimateur de $\hat{Y}_{\text{EASS}}$ total. Dans cette section, nous considérons que le nombre total de strates H est défini avant toute application des méthodes de stratification optimale examinées.

Dans la pratique, nous ne disposons pas des valeurs de la variable d'enquête y et, par conséquent, la variance n'est pas calculable dans l'expression (2.3). Une méthode courante consiste à minimiser plutôt la variance (ou le CV) de l'estimateur $\hat{X}_{\text{EASS}}$ pour le total de la variable de stratification x. Un certain nombre d'auteurs ont conçu des méthodes qui s'attachent à ce problème d'optimisation que nous appellerons désormais « problème de stratification unidimensionnelle ». Nous adoptons la même orientation ici.

Trouver les bornes qui minimisent la variance (2.6) ou le CV (2.7) représente un problème difficile tant en analyse qu'en calcul, et ce, parce que les tailles de population et d'échantillon en nombres entiers (N_h et n_h respectivement) dépendent non linéairement des bornes des strates. D'après de Moura Brito, Ochi, Montenegro et Maculan (2010a), le nombre de possibilités pour les bornes peut être trés large, car dépendent de N, H et du nombre de valeurs distinctes x dans la population.

Vu cette difficulté, on a conçu diverses méthodes dans les dernières décennies pour trouver des bornes optimales de strate, le but étant de dégager au moins des solutions correspondant à des minima locaux de bonne qualité.

Dalenius (1951) s'est attaqué au problème dans le cas $H = 2$ en approchant la variance en (2.6) sans tenir compte de la correction de population finie, ce qui équivaut à supposer que l'échantillonnage intrastrate aurait été un échantillonnage aléatoire simple avec remise. La variance approximative à minimiser est alors donnée par :

$$\text{Var}\left(\hat{X}_{\text{EASS}}\right) \cong \sum_{h=1}^{H} N_h^2 S_{hx}^2 / n_h. \tag{3.1}$$

Dans une répartition de Neyman (Cochran, 1977) par la variable x et avec remplacement des tailles d'échantillon n_h en (3.1) par leurs valeurs théoriques $n_h = N_h S_{hx} \Big/ \sum_{k=1}^{H} N_k S_{kx}$, on obtient l'expression employée par Dalenius (1951) :

$$\text{Var}\left(\hat{X}_{\text{EASS}}\right) \cong \left(\sum_{h=1}^{H} N_h S_{hx}\right)^2 \Big/ n. \tag{3.2}$$

Dalenius et Hodges (1959) ont regardé le cas $H > 2$, et offert une solution analytique consistant à approcher la distribution de la variable x par son histogramme comportant un nombre modéré de classes. En prenant toujours la variance approximative et en posant une répartition de Neyman, Ekman (1959) a proposé une solution avec une approche géométrique pour trouver les bornes de strate. Hedlin (2000) a élargi la solution d'Ekman en retenant la variance initiale (2.6) comme la fonction à minimiser, ce qu'il a appelé la règle élargie d'Ekman.

Hidiroglou (1986) a avancé une approche qui spécifie d'avance la précision recherchée (CV) de l'estimateur du total et qui divise la population en deux strates $(H = 2)$ de sorte que la taille totale de l'échantillon n soit minimisée. Dans cette étude, la seconde strate est à tirage complet ou « à certitude », tous les éléments étant compris dans l'échantillon avec l'unité comme probabilité $(n_2 = N_2)$. Lavallée et Hidiroglou (1988) ont généralisé cette méthode au cas $H > 2,$ tout en conservant l'idée que la strate contenant le plus grand nombre d'unités de population doive être à échantillonnage complet. Ils adoptaient dans cette optique une répartition spéciale dite de puissance (Bankier, 1988). Plus récemment, Rivest (2002) a encore généralisé la méthode de Lavallée et Hidiroglou (1988) en considérant que le but est de minimiser la variance de l'estimateur d'un total pour une prévision de modélisation de la variable d'enquête y au lieu de la variable de stratification x.

Gunning et Horgan (2004) ont proposé ce qu'on appelle la méthode géométrique de définition des bornes de strate. Dans cette méthode, on pose que les CV de la variable de stratification x sont approximativement constants et que la distribution de la variable de stratification est approximativement uniforme dans chaque strate. Selon ces hypothèses, l'optimum des bornes de strate formerait une progression géométrique, menant de la sorte à une solution analytique très simple.

Keskintürk et Er (2007) ont proposé une technique d'optimisation globale relevant de ce qu'on appelle les algorithmes génétiques. Dans le même ordre d'idées, de Moura Brito et coll. (2017b) ont appliqué une autre technique d'optimisation globale appelée GRASP au problème de stratification. Dans notre propos, nous avons pris le même chemin que Keskintürk et Er (2007), mais en opérant un choix efficace d'algorithme génétique appelé algorithme génétique biaisé à clés aléatoires (BRKGA) que nous décrivons à la section 4.

Kozak (2004) a proposé une méthode dite de recherche aléatoire reprise par Kozak et Verma (2006), celle-ci étant alors comparée à la méthode géométrique de Gunning et Horgan (2004). Khan, Nand et Ahmad (2008) ont exploité les idées de la programmation dynamique pour concevoir un algorithme qui détermine les bornes de strate en considérant que la variable de stratification est en distribution triangulaire ou normale et que l'échantillonnage intrastrate est avec remise. De Moura Brito, Maculan, Lila et Montenegro (2010b) ont proposé un algorithme exact reposant sur la théorie des graphes et où on pose une répartition proportionnelle entre les strates.

Er (2011) a fait une comparaison d'efficacité entre des méthodes figurant dans la documentation spécialisée en prenant comme solution initiale la méthode géométrique de Gunning et Horgan (2004). Kozak (2014) a comparé sa technique de recherche aléatoire à l'algorithme génétique proposé par Keskintürk et Er (2007). Rao, Khan et Reddy (2014) ont mis au point une méthode qui traite simultanément les problèmes de détermination des bornes de strate et de répartition entre les strates. Leur algorithme repose sur l'hypothèse selon laquelle la variable de stratification suit une distribution de Pareto. Notre optique est plus générale, et nous ne supposons pas que la variable de taille obéit à une distribution en particulier.

4 Algorithme génétique biaisé à clés aléatoires

L'algorithme génétique biaisé à clés aléatoires (appelé BRKGA dans la suite de notre exposé), que proposent Gonçalves et Resende (2011), est une méthode métaheuristique qui a été appliquée à plusieurs problèmes d'optimisation. Voir Festa (2013) et Oliveira et coll. (2017), par exemple. Le principe sous-tendant cette méthode rappelle la théorie biologique de l'évolution des espèces.

L'algorithme prend une « population » initiale de solutions possibles au problème cible, laquelle vient d'un mécanisme aléatoire spécifié. Cette population évolue ensuite au gré des itérations en conservant les meilleures solutions disponibles à chaque itération (solutions retenues) et en remplaçant les solutions non retenues par des solutions produites par perturbation aléatoire et évoquant les croisements et les mutations des populations naturelles. Au fil des itérations, les solutions sont conservées ou évoluent selon la valeur de la fonction à optimiser.

Dans l'algorithme BRKGA, les solutions candidates sont codées, c'est-à-dire sont représentées par des vecteurs dont les éléments sont des nombres dans l'intervalle $(0; 1)$. Avec un vecteur observé, une procédure de décodage doit être appliquée. Cette procédure fait correspondre la valeur d'un vecteur à une solution possible du problème d'optimisation cible. C'est ce qui relie l'algorithme au problème d'optimisation précis à traiter. La figure 4.1 présente le pseudocode d'un algorithme BRKGA générique.

La démarche est décrite et illustrée en détail à la section 4.1 avec un exemple de problème de stratification unidimensionnelle et une description de toutes les étapes à la figure 4.1.

1) On génère la population initiale composée de p vecteurs aléatoires (clés) $\mathbf{v}$, où chaque valeur est tirée aléatoirement de la distribution uniforme $[0; 1]$.
2) On applique la procédure de décodage à chaque vecteur $\mathbf{v}$ de la population, ce qui donne p solutions possibles du problème d'optimisation.
3) On calcule la valeur de la fonction objective pour chaque solution dans la population.
4) On choisit les $p_e\,(1 < p_e < p)$ meilleures solutions (appelées solutions retenues) selon les valeurs de la fonction objective et les ajoute à la population à considérer à l'itération suivante.
5) On génère $p_m\,(1 < p_m < p)$ nouveaux vecteurs aléatoires comme à l'étape 1), ce qu'on appelle les *mutations*, et les ajoute à la population à considérer à l'itération suivante.
6) On génère les $(p - p_e - p_m)$ vecteurs restants appelés *croisements* pour compléter la population qui sera considérée à l'itération suivante, et ce, en croisant un des p_e vecteurs d'une solution retenue avec un des $(p - p_e)$ vecteurs d'une des solutions non retenues à la présente itération.
7) On fait des itérations à partir de l'étape 2) tant que les critères d'arrêt ne sont pas remplis.

Figure 4.1 Pseudocode pour un algorithme BRKGA.

4.1 Algorithme BRKGA pour le problème de stratification unidimensionnelle

On considère d'abord le vecteur de population $X_U = \{x_1, x_2, \ldots, x_N\}$ et calcule l'ensemble $C = \{c_1, c_2, \ldots, c_K\}$ contenant les K valeurs distinctes de x observées dans la population. Si

$X_U = \{1, 3, 3, 5, 6, 7, 7, 7, 8, 9, 10, 10, 11\}$, par exemple, $C = \{1, 3, 5, 6, 7, 8, 9, 10, 11\}$. Si $K > 100$, nous calculons les dix percentiles supérieurs de x pour dégager l'ensemble $Q = \{q_{90}, q_{91}, \ldots, q_{99}, q_{100}\}$. Si $K \leq 100$, nous calculons les percentiles choisis de x pour dégager l'ensemble $Q = \{q_5, q_{10}, \ldots, q_{95}, q_{100}\}$. Nous avons retenu le point de démarcation de 100 pour K après une certaine expérimentation initiale de notre méthode avec quelques-unes des populations considérées dans l'expérience numérique que nous décrivons à la section 5. Les définitions autres de l'ensemble Q aident à diversifier le jeu de solutions possibles issu de l'algorithme BRKGA.

Dans l'application de l'algorithme au problème de stratification unidimensionnelle, chaque solution est représentée par un vecteur $\mathbf{v} = \{v_1, \ldots, v_H\}$ à H positions où les $H - 1$ premières positions contiennent des valeurs entre 0 et 1 et où la position H reçoit la valeur d'un percentile de la distribution de la variable de stratification x.

Nous prenons ensuite $x_{\min}$ comme valeur la plus petite de C et v_H comme élément choisi au hasard dans Q. À la première itération, nous tirons les valeurs des $H - 1$ premières positions de chaque vecteur $\mathbf{v}$ indépendamment de la distribution uniforme $[0; 1]$.

La procédure de décodage permettant de dégager de chaque vecteur $\mathbf{v}$ généré une solution du problème de stratification unidimensionnelle se définit ainsi :

$$b_h = x_{\min} + v_h (v_H - x_{\min}) \quad \text{pour} \quad h = 1, \ldots, H - 1. \tag{4.1}$$

Une fois obtenues les $H - 1$ premières valeurs pour b_h, celles-ci doivent être mises par ordre croissant de sorte que les éléments du vecteur résultant $\mathbf{b} = \left(b_{(1)}, b_{(2)}, \ldots, b_{(H-1)}\right)$ forment les bornes de solution pour le vecteur $\mathbf{v}$ correspondant, $b_{(h)}$ étant la statistique de h^{e} ordre des valeurs $b_1, \ldots, b_{H-1}$ calculées en (4.1).

Pour citer un exemple de décodage, supposons que $H = 4$, $x_{\min} = 10$, $K = 300$, $Q = \{200; 215; 280{,}5; 300; 318; 400; 425; 478; 500; 510\}$. Considérons aussi le vecteur $\mathbf{v} = (0{,}48; 0{,}35; 0{,}20)$ généré comme nous l'avons décrit. Il s'ensuit que $b_1 = 10 + 0{,}48 \times (200 - 10)$ que $b_2 = 10 + 0{,}35 \times (200 - 10)$ et que $b_3 = 10 + 0{,}20 \times (200 - 10)$. Après tri, on obtient alors $\mathbf{b} = (48; 76{,}5; 101{,}2)$.

Le vecteur $\mathbf{b}$ étant donné, les valeurs de N_h et S_{hx}^2 s'obtiennent facilement pour chacune des H strates. Nous dégageons les valeurs des tailles d'échantillon n_h pour les diverses strates en appliquant la méthode de répartition optimale proposée par de Moura Brito et coll. (2015). Nous calculons ainsi les tailles d'échantillon n_h de manière à minimiser une somme pondérée des variances (ou des CV) des estimateurs des totaux de m variables d'enquête, la taille totale d'échantillon n étant fixe.

Comme nous prenons ici comme cible de la minimisation la variance de l'estimateur pour le total de la variable de stratification x, nous posons $m = 1$ et utilisons la formulation (D) qui vient de de Moura Brito et coll. (2015) pour résoudre le problème de répartition optimale unidimensionnelle avec l'équation (2.6) comme variance à minimiser. À noter que cette méthode donne l'optimum global pour le problème de répartition.

Nous poursuivons avec l'algorithme selon la figure 4.1 en générant un ensemble initial de p vecteurs $\mathbf{v}$. À l'étape 2, nous décodons chacun de ces vecteurs $\mathbf{v}$ pour dégager une solution possible $\mathbf{b}$ du problème

de stratification optimale. À l'étape 3, nous obtenons la répartition optimale correspondant à $\mathbf{b}$ et calculons la valeur de la fonction objective. Nous exécutons ensuite les étapes 4 à 6 pour trouver la population suivante de solutions possibles et reprenons la procédure jusqu'à ce que les critères d'arrêt soient remplis. À l'étape 4, nous dégageons les p_e solutions retenues et les ajoutons à la population suivante. À l'étape 5, nous produisons p_m mutations et les ajoutons à la population suivante. À l'étape 6, nous produisons $(p - p_e - p_m)$ croisements à l'aide de l'opérateur de « croisement uniforme » proposé par Spears et De Jong (1991) pour tirer un nouveau vecteur $\mathbf{v}$ d'une des p_e solutions retenues et d'une des $(p - p_e - p_m)$ solutions non retenues actuelles. Nous procédons ainsi : une fois choisis les deux vecteurs ($\mathbf{v}_e$ et $\mathbf{v}_n$) à croiser, nous générons un vecteur auxiliaire à clés aléatoires ($\mathbf{v}_a$) avec des tirages indépendants de la distribution uniforme $[0; 1]$. Soit $r_c > 0{,}5$ une probabilité préspécifiée qu'une valeur soit copiée du vecteur retenu $\mathbf{v}_e$. Nous formons alors le vecteur croisé $\mathbf{v}_c$ en en tirant les valeurs de $\mathbf{v}_e$ aux positions où la valeur correspondante dans $\mathbf{v}_a$ est moindre que r_c (ce qui équivaut à 0,7 dans l'exemple de la figure 4.2) et de $\mathbf{v}_n$ à toutes les autres positions.

Pour produire chacun des $(p - p_e - p_m)$ vecteurs de la génération suivante, l'algorithme choisit un vecteur v_e au hasard (par la fonction d'*échantillon* en R) dans les p_e vecteurs retenus et un autre vecteur v_n dans les $p - p_e$ vecteurs non retenus et il croise les vecteurs ainsi obtenus. La sélection des vecteurs à partir des deux sous-ensembles se fait avec remise, ce qui implique que, individuellement, des vecteurs retenus ou non retenus peuvent être sélectionnés pour être croisés plus d'une fois.

Vecteurs\positions	1	2	3
$\mathbf{v}_e$	**0,31**	**0,77**	**0,65**
$\mathbf{v}_n$	**0,26**	**0,18**	**0,36**
$\mathbf{v}_a$	**0,58**	**0,89**	**0,11**
$\mathbf{v}_c$	**0,31**	**0,18**	**0,65**

Figure 4.2 Croisement uniforme avec $r_c = 0{,}7$.

Prenons maintenant un exemple avec $H = 4$, $x_{\min} = 10$, $K = 300$, $p = 8$, $p_e = 3$, $p_m = 3$, $r_c = 0{,}7$ et $Q = \{200; 215; 280{,}5; 300; 318; 400; 425; 478; 500; 510\}$. La figure 4.3 illustre l'application de toutes les étapes de l'algorithme au problème de stratification unidimensionnelle pour deux itérations consécutives de cet algorithme.

Nous avons mis en œuvre dans le package *stratbr* en *R* disponible à partir de CRAN (voir de Moura Brito et coll., 2017a) l'approche BRKGA ici décrite du problème de stratification optimale unidimensionnelle. Le package a permis d'obtenir tous les résultats présentés à la section 5.

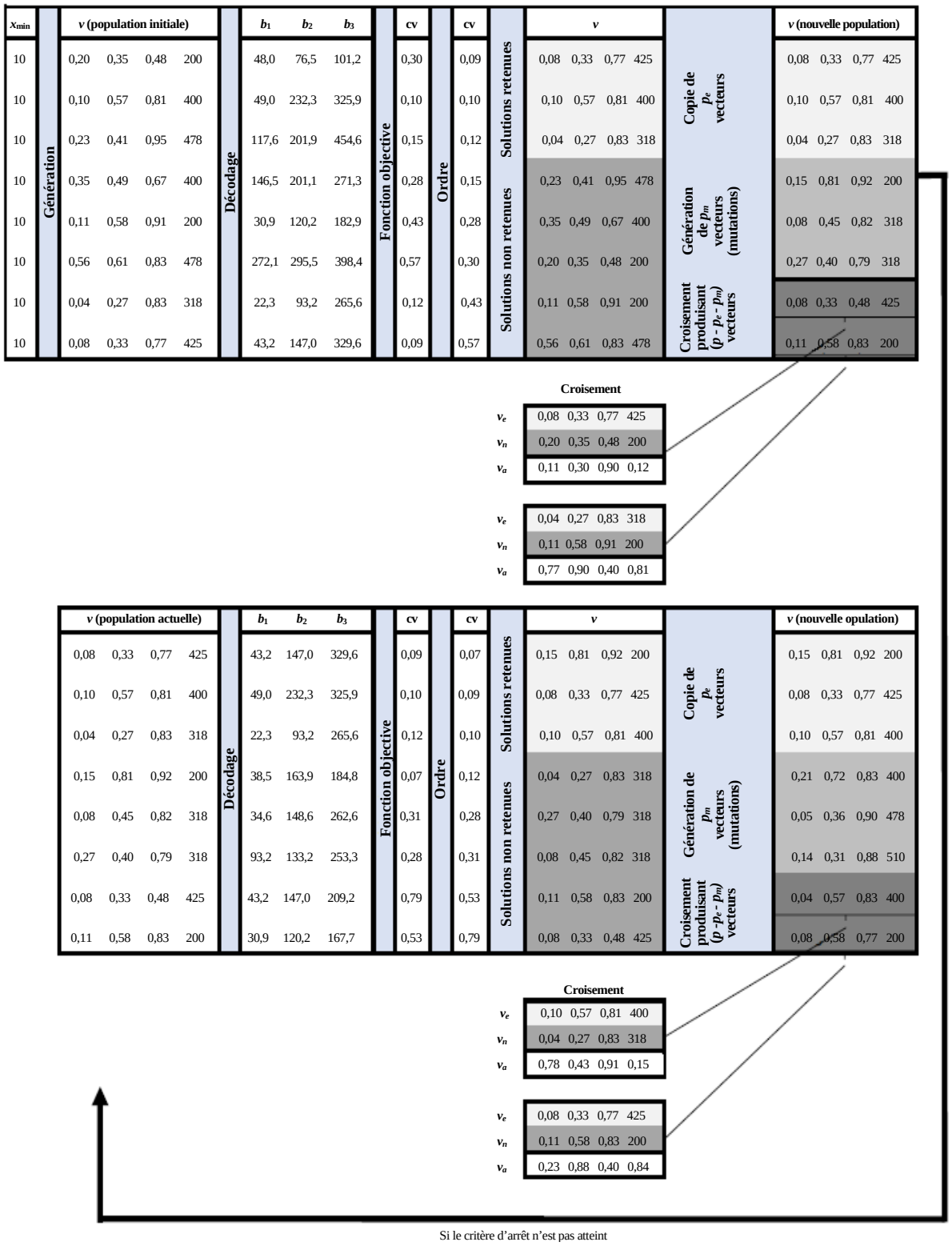

Figure 4.3 Illustration de la méthode BRKGA de stratification optimale.

5 Résultats du calcul

Dans cette section, nous présentons les résultats de l'application de six méthodes au problème de stratification (Dalenius et Hodges (DH), méthode géométrique (GH), Kozak (KO), algorithme génétique de Keskintürk et Er (KE), GRASP (GR) et nouvelle méthode BRKGA décrite à la section 4 (BR)). Nous avons effectué toute cette expérience avec la version 3.3.1 de R. On peut trouver les méthodes DH, GH et KO dans le package *stratification* en R de Baillargeon et Rivest (2014) (version 2.2-5). Nous employons dans ce cas la méthode de répartition d'échantillon de Neyman. La méthode KE figure dans le package en *R GA4Stratification* de Er, Keskintürk et Daly (2010) (version 1.0). Avec cette méthode, le maximum d'itérations considérées était de 10 000 et les valeurs des autres paramètres requis étaient celles qu'indiquent Keskintürk et Er (2007), à savoir $p = 35$ solutions candidates dans chaque population, un taux de mutation de 15 % et une répartition d'échantillon aussi fondée sur l'algorithme génétique. Les auteurs ont appliqué les méthodes GR et BR en R et le code en question figure dans le package *stratbr* de de Moura Brito et coll. (2017a) (version 1.2) disponible par le réseau CRAN.

Dans le cas de la méthode BR, nous avons examiné $p = 100$ solutions candidates à chaque itération avec 20 % de solutions retenues $(p_e = 20)$ et 30 % de mutations $(p_m = 30)$. La probabilité de copier un gène du vecteur retenu était fixée à $r_c = 0{,}6$ et le nombre total d'itérations, à 1 500. Aux fins de la répartition d'échantillon, nous avons combiné les méthodes BR et GR à la formulation proposée par de Moura Brito et coll. (2015) qui figure dans le package *MultAlloc* en R et qui est aussi disponible par le réseau CRAN.

Dans une comparaison d'efficacité relative, nous avons appliqué ces méthodes à 27 populations. Certaines de celles-ci figurent dans les packages *stratification* et *GA4Stratification* en R; elles ont auparavant servi à certaines études comparatives comme celles de Keskintürk et Er (2007), Er (2011) et de Moura Brito et coll. (2017b). On trouvera à l'annexe A une brève description de toutes ces populations avec des précisions sur les variables considérées comme la « variable x » dans chaque population. Le tableau 5.1 présente certaines descriptions sommaires des populations en question.

Les 27 populations traitées forment ici un ensemble très hétérogène et leur taille totale varie de quelques centaines d'éléments (ME84 et P75 avec $N = 18\,570$ sont les plus petites tailles) à plusieurs milliers (Coffee avec $N = 18\,570$ est la taille la plus grande). On constate aussi une forte variation (de $K = 51$ pour Kozak1 à $K = 5\,453$ pour Kozak3) du nombre K de valeurs distinctes de la variable de stratification, qui est la mesure de taille la plus importante pour l'efficacité de notre algorithme d'optimisation. Notons enfin une ample variation de l'asymétrie des distributions de la variable x entre des valeurs modestes en sens négatif (-0,70 pour Beta103) et une valeur appréciable (40,04 pour CensoCO).

Nous avons fait tous les calculs du volet computationnel de notre expérience en R avec un ordinateur de 24 Go de mémoire vive et 8 processeurs de 3,40 GHz (I7). Tirant parti de l'architecture multicœur des ordinateurs modernes, nous avons employé le package *snowfall* en R pour un traitement parallèle de l'algorithme BRKGA. Précisons que, à chaque itération, la procédure de décodage produit un jeu de

solutions pour les bornes. Ces bornes sont ensuite transmises au package *MultAlloc* pour une répartition optimale permettant d'obtenir les tailles d'échantillon des diverses strates, puis de calculer la fonction objective de variance. Comme la formulation de cet exercice d'optimisation globale influe directement sur le temps de calcul à cette étape, nous avons mis la répartition et le calcul de la fonction objective en traitement parallèle.

Tableau 5.1
Tableau récapitulatif de la variable de stratification pour les 27 populations

Populations	N	K	Minimum	Maximum	Asymétrie
AgrMinas	844	226	5,00	47 800,00	7,32
BeefFarms	430	353	50,00	24 250,00	4,56
Beta103	1 000	1 000	357,98	985,96	-0,70
CensoCO	9 977	79	1,00	911,00	40,04
Chi5	1 000	1 000	0,06	23,43	1,40
Café	18 570	538	0,01	13 212,00	19,69
Débiteurs	3 369	1 129	40,00	28 000,00	6,44
HHinctot	16 025	224	1,00	6 900,00	2,71
Iso2004	487	487	6,36	1 044,66	10,03
Kozak1	4 000	51	72,00	3,00	1,40
Kozak3	2 000	581	2 793,00	6,00	3,55
Kozak4	10 000	5 453	74 400,00	62,00	4,20
ME84	284	264	173,00	47 074,00	8,64
EMCD	2 000	2 000	1,41	4 863,66	8,61
P100e10	1 000	1 000	73,56	127,32	-0,03
P75	284	68	4,00	671,00	8,43
Pop500	500	261	0,01	47 841,42	21,53
Pop800	800	402	0,01	4 735,10	22,13
pop1076	1 076	88	5,00	1 643,00	13,23
pop1616	1 616	165	5,00	2 618,00	11,09
pop2911	2 911	247	5,00	2 497,00	11,50
REV84	284	277	347,00	59 877,00	7,83
SugarCaneFarms	338	101	18,00	280,00	2,26
Swiss	2 896	881	0,00	3 634,00	2,73
USbanks	357	200	70,00	977,00	2,07
UScities	1 038	116	10,00	198,00	2,87
UScolleges	677	576	200,00	9 623,00	2,45

Nota : N est la taille totale de population et K est le nombre de valeurs uniques de la variable de stratification.

Les six méthodes de l'expérience numérique ont été appliquées à chacune des 27 populations; le nombre H de strates était de 3, 4, 5 et 6. Nous avons employé ces valeurs puisqu'elles revenaient souvent dans les applications et dans des études comparatives semblables de la documentation spécialisée comme celles de Er (2011) et de Gunning et Horgan (2004). Nous avons négligé les valeurs supérieures de H, car le gain d'efficacité serait modeste avec $H > 6$. Nous avons pris comme taille d'échantillon $n = 100$ (à coût fixe) comme dans les expériences numériques de Er (2011) et Kozak et Verma (2006).

Pour évaluer l'efficacité des méthodes, nous avons calculé les CV de l'estimateur du total de la variable de stratification x pour chaque population et chaque nombre de strates, ce qui a donné $27 \times 4 = 108$

scénarios pour chaque méthode. Nous avons obtenu les CV à l'équation (2.7) et multiplié ces valeurs par 100 pour les mettre sous forme de pourcentage. Le tableau 5.2 présente les CV des six méthodes. Les cases ombrées correspondent aux méthodes représentant la meilleure solution (CV minimal) dans chacun des 108 scénarios. Les « sans objet » dans ces tableaux sont les cas où nous ne pouvions obtenir de solutions à cause de problèmes avec la méthode de stratification ou la répartition correspondante.

Si nous analysons les résultats au tableau 5.2 et, en particulier, les cases ombrées, il ressort que BR est d'un excellent rendement si on compare cette méthode aux cinq rivales. Cette perception est renforcée par les courbes de la figure 5.1 où la méthode BR est comparée à toutes ses rivales. Les points au-dessus de la droite représentent les scénarios où la méthode mise en comparaison est d'un moindre rendement que la méthode BR. À considérer ces courbes, il est clair que les trois méthodes les plus performantes sont GR, KO et BR.

Le tableau 5.3 indique en pourcentage le nombre de fois que chaque méthode produit la meilleure solution sur les 108 scénarios. Les deux méthodes BR et KO sont d'un rendement supérieur aux autres méthodes et se retrouvent à égalité à plusieurs reprises pour la meilleure solution. La méthode DH a produit la meilleure solution dans seulement 3 des 108 scénarios et la méthode GH ne le fait jamais.

Ajoutons que la méthode géométrique GH donne non seulement des CV élevés, mais souvent aussi des solutions impossibles où les bornes de strate mènent à des répartitions où les tailles d'échantillon sont supérieures aux tailles de population correspondantes. Cette méthode a quelquefois pour effet de répartir la population en laissant très peu d'éléments dans certaines strates. D'après Gunning et Horgan (2004) et comme le signalent Keskintürk et Er (2007), comme l'étendue des intervalles s'accroît géométriquement, la méthode GH ne donne pas de bons résultats avec de faibles valeurs de la variable de stratification, puisque certaines strates sont alors étroites. Cette méthode est inapplicable de surcroît lorsque la valeur la plus basse de la variable de stratification est zéro.

Pour la plupart des populations, la méthode KE a produit des CV proches de ceux des méthodes KO, GR et BR, qui sont les plus efficaces en temps de calcul. Nous avons observé une forte variation des temps de calcul entre les méthodes. La méthode KE était la pire à ce critère avec des temps bien supérieurs à ceux des méthodes rivales. Par ailleurs, la méthode KO avait les calculs les plus rapides et offrait fréquemment la meilleure précision possible (CV les plus bas). La méthode BR présentait un temps de calcul intermédiaire entre ceux des méthodes KO et KE.

Le graphique à la figure 5.2 présente en pourcentage les fois que chacune des méthodes BR, KO, KE et GR produit la meilleure solution selon le nombre de strates. On y voit un net avantage pour la méthode BR comparativement aux méthodes KE et GR. Comparée à la KO, la BR a un meilleur rendement avec $H = 3$ et $H = 6$, la KO l'emporte cependant sur le BR avec $H = 4$ et $H = 5$. La GR était aussi bonne que la KO avec $H = 3$ et $H = 6$, mais l'était moins que la BR et KO avec $H = 4$ et $H = 5$. La KE était clairement la perdante dans cette analyse pour tout nombre H de strates.

Nous avons en outre étudié les associations entre le rendement et d'autres facteurs possibles comme l'asymétrie ou la taille (N ou K) des populations, mais sans en arriver à des associations significatives dans notre ensemble limité de populations.

Tableau 5.2
CV de l'estimateur du total de la variable de stratification selon les scénarios

Populations	H	CV_{DH}	CV_{GH}	CV_{KO}	CV_{KE}	CV_{GR}	CV_{BR}
AgrMinas	3	4,158	7,187	4,050	4,089	4,050	4,050
	4	2,714	4,965	2,643	2,811	2,645	2,645
	5	2,325	3,828	1,945	2,262	1,945	1,945
	6	1,821	2,975	1,593	1,932	1,580	1,580
BeefFarms	3	2,758	2,491	1,875	2,086	1,875	1,875
	4	1,853	1,825	1,188	1,557	1,188	1,188
	5	1,455	1,369	0,902	1,280	0,902	0,902
	6	1,148	1,167	0,726	0,990	0,726	0,726
Beta103	3	0,561	0,810	0,560	0,560	0,559	0,559
	4	0,413	0,579	0,410	0,408	0,410	0,410
	5	0,337	0,500	0,329	0,329	0,329	0,329
	6	0,280	0,418	0,276	0,275	0,277	0,276
CensoCO	3	NA	4,839	4,334	4,336	4,334	4,334
	4	NA	4,388	3,078	3,062	3,078	3,078
	5	NA	NA	2,401	2,435	2,401	2,401
	6	NA	NA	1,949	1,956	1,943	1,943
Chi5	3	2,522	4,217	2,502	2,489	2,502	2,502
	4	1,897	3,199	1,889	1,881	1,889	1,889
	5	1,518	2,875	1,515	1,538	1,515	1,515
	6	1,258	NA	1,248	1,251	1,248	1,248
Café	3	10,049	12,598	6,906	6,876	6,906	6,906
	4	NA	10,450	4,996	5,027	4,996	4,996
	5	NA	8,124	3,877	3,939	3,877	3,877
	6	NA	6,756	3,176	3,477	3,176	3,176
Débiteurs	3	5,626	6,150	5,554	5,554	5,554	5,554
	4	4,098	4,387	4,049	4,049	4,049	4,049
	5	3,163	3,595	3,131	3,131	3,131	3,131
	6	2,639	2,897	2,562	2,562	2,562	2,562
HHinctot	3	3,206	5,106	3,184	3,184	3,184	3,184
	4	2,436	4,542	2,429	2,430	2,429	2,429
	5	1,993	4,225	1,973	1,979	1,973	1,973
	6	1,676	3,794	1,629	1,629	1,629	1,629
Iso2004	3	2,716	3,330	1,894	1,894	1,894	1,894
	4	2,059	2,154	1,206	1,206	1,207	1,207
	5	1,616	1,839	0,908	0,908	0,909	0,909
	6	1,380	NA	0,702	0,703	0,704	0,703
Kozak1	3	1,695	2,432	1,695	1,695	1,695	1,695
	4	1,305	2,020	1,301	1,301	1,301	1,301
	5	1,051	1,705	1,050	1,052	1,050	1,050
	6	0,904	1,402	0,890	0,917	0,890	0,890
Kozak3	3	3,673	5,049	3,663	3,659	3,663	3,663
	4	2,733	3,980	2,723	2,724	2,723	2,723
	5	2,208	3,199	2,178	2,231	2,178	2,178
	6	1,823	2,733	1,817	1,827	1,819	1,817
Kozak4	3	4,263	5,811	4,257	4,239	4,257	4,257
	4	3,219	4,696	3,204	3,193	3,205	3,204
	5	2,606	3,873	2,589	2,587	2,591	2,589
	6	2,168	3,236	2,155	2,155	2,157	2,158
ME84	3	1,703	2,527	1,296	1,296	1,296	1,296
	4	1,402	1,642	0,870	0,870	0,870	0,870
	5	1,050	1,549	0,661	0,661	0,661	0,661
	6	0,907	1,213	0,521	0,577	0,521	0,521

Tableau 5.2 (suite)
CV de l'estimateur du total de la variable de stratification selon les scénarios

Populations	H	CV_{DH}	CV_{GH}	CV_{KO}	CV_{KE}	CV_{GR}	CV_{BR}
EMCD	3	4,363	5,829	4,167	4,167	4,167	4,167
	4	3,406	5,259	2,960	2,960	2,961	2,960
	5	2,498	4,015	2,297	2,485	2,297	2,297
	6	2,167	3,445	1,836	1,836	1,838	1,836
P100e10	3	0,375	0,444	0,373	0,371	0,373	0,373
	4	0,295	0,346	0,294	0,294	0,294	0,294
	5	0,236	0,288	0,236	0,236	0,236	0,236
	6	0,198	0,242	0,196	0,198	0,196	0,196
P75	3	1,635	2,592	1,459	1,459	1,459	1,459
	4	1,415	1,798	0,966	0,966	0,966	0,966
	5	1,047	1,563	0,829	0,835	0,713	0,713
	6	0,896	1,250	0,769	0,553	0,552	0,552
pop1076	3	4,597	3,715	2,437	2,775	2,437	2,437
	4	NA	2,853	1,624	2,164	1,624	1,624
	5	NA	2,168	1,204	1,869	1,203	1,203
	6	NA	1,827	0,953	1,549	0,951	0,951
pop1616	3	4,989	4,318	3,898	3,921	3,898	3,898
	4	3,823	3,267	2,564	2,716	2,564	2,564
	5	3,187	2,508	1,882	2,183	1,882	1,882
	6	NA	2,050	1,527	1,962	1,496	1,496
pop2911	3	5,925	5,935	5,605	5,569	5,605	5,605
	4	4,070	3,992	3,807	3,807	3,807	3,807
	5	3,262	3,183	2,918	2,943	2,918	2,918
	6	2,632	2,649	2,281	2,418	2,281	2,281
Pop500	3	NA	0,678	0,092	0,127	0,092	0,092
	4	NA	0,178	0,059	0,082	0,060	0,060
	5	NA	0,194	0,043	0,059	0,045	0,046
	6	NA	0,117	0,033	0,046	0,036	0,037
Pop800	3	NA	3,133	1,555	2,448	1,555	1,555
	4	NA	2,755	0,996	1,511	0,996	0,996
	5	NA	1,620	0,701	1,261	0,702	0,702
	6	NA	1,436	0,546	0,823	0,550	0,548
REV84	3	1,901	2,777	1,614	1,776	1,614	1,614
	4	1,500	1,975	1,120	1,120	1,120	1,120
	5	1,235	1,700	0,835	0,836	0,835	0,835
	6	0,881	1,315	0,666	0,666	0,667	0,666
SugarCaneFarms	3	1,640	1,929	1,627	1,628	1,627	1,627
	4	1,152	1,440	1,118	1,122	1,118	1,118
	5	0,912	1,186	0,839	0,858	0,839	0,839
	6	0,707	1,041	0,691	0,732	0,682	0,682
Swiss	3	3,726	NA	3,682	3,683	3,690	3,682
	4	2,830	NA	2,781	2,781	2,787	2,781
	5	2,246	NA	2,227	2,549	2,232	2,228
	6	1,905	NA	1,860	1,880	1,864	1,860
USbanks	3	1,861	1,843	1,802	1,802	1,802	1,802
	4	1,364	1,417	1,270	1,270	1,270	1,270
	5	1,118	1,079	0,861	0,861	0,861	0,861
	6	0,794	0,850	0,718	0,710	0,710	0,710
UScities	3	2,738	2,705	2,655	2,687	2,655	2,655
	4	1,972	1,951	1,927	1,934	1,927	1,927
	5	1,483	1,451	1,436	1,437	1,436	1,436
	6	1,260	1,305	1,228	1,214	1,209	1,209
UScolleges	3	2,928	3,169	2,749	2,749	2,749	2,749
	4	2,106	2,185	2,018	2,018	2,018	2,018
	5	1,707	1,838	1,606	1,607	1,607	1,606
	6	1,486	1,488	1,323	1,323	1,323	1,323

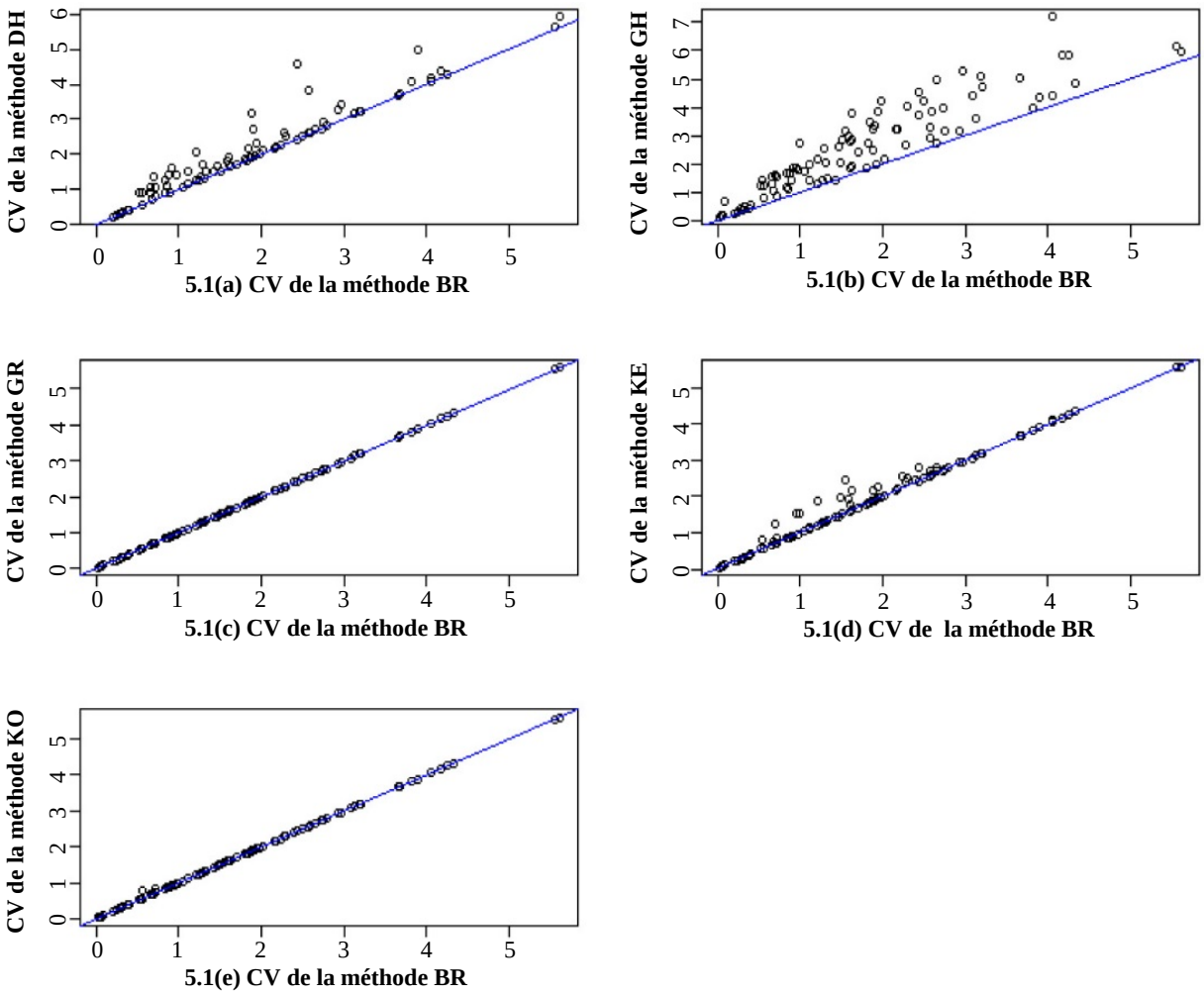

Figure 5.1 Comparaison des CV des estimateurs du total dans les diverses méthodes de stratification pour l'ensemble des populations et les nombres de strates (H).

Tableau 5.3
Pourcentage de fois que la méthode a produit la meilleure solution

Méthode	Nombre de fois en %
DH	2,8
GH	0,0
KE	42,6
GR	71,3
KO	78,7
BR	78,7

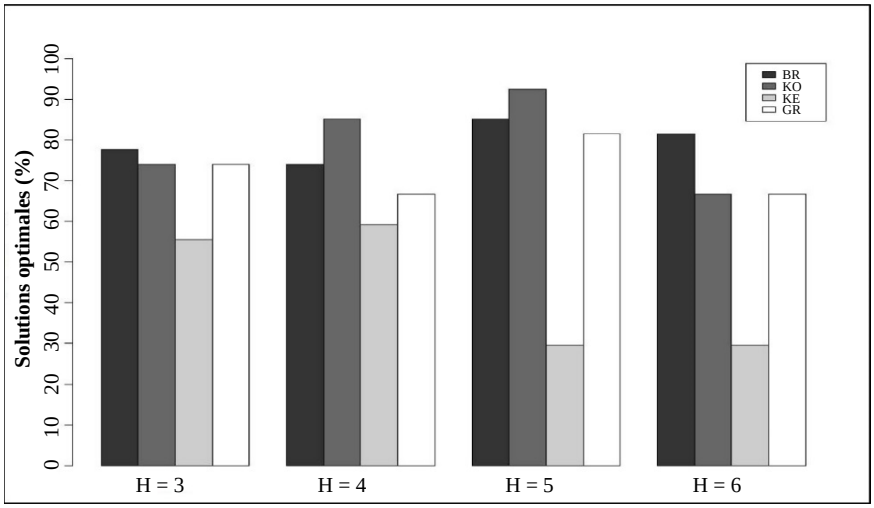

Figure 5.2 Pourcentage de meilleures solutions produites par méthode et nombre de strates (H).

6 Conclusions

Comme nous l'avons mentionné, l'échantillonnage stratifié est très important comme plan de sondage, puisqu'il aide à améliorer la précision des estimations d'enquête pour une taille donnée d'échantillon ou un budget d'enquête. Cette constatation vaut particulièrement pour les populations asymétriques ou hétérogènes qui sont souvent caractéristiques des enquêtes auprès des entreprises ou des établissements. Les gains que peut apporter la stratification dépendent fortement de la délimitation des strates et de la répartition de l'échantillon entre ces strates pour une variable de stratification et une méthode de sélection d'échantillon déterminées.

Nous avons présenté une nouvelle méthode d'optimisation applicable à ce problème de stratification avec l'algorithme génétique biaisé à clés aléatoires (BRKGA). Dans notre approche (appelée BR), nous combinons l'algorithme de détermination des bornes de strate à la formulation proposée par de Moura Brito et coll. (2015) pour une répartition optimale de l'échantillon, laquelle se révèle efficace en temps de calcul dans le cas des grandes populations (N élevé).

Les résultats présentés de la comparaison de cette méthode avec les cinq méthodes rivales considérées semblent indiquer que la méthode BR constitue un bon moyen de traitement des problèmes de stratification et de répartition dans la pratique.

Il serait facile de généraliser notre approche aux cas où la variable de stratification x n'est pas « mesurée », mais récapitule plutôt un certain nombre de covariables sous la forme d'une variable y prédite. Il en va de même de la généralisation à deux variables numériques x ou plus, ce qu'on peut aisément accomplir en changeant la fonction de décodage servant à tirer les solutions possibles de l'algorithme BRKGA avec le package *stratbr* en R (voir de Moura Brito et coll., 2017a).

Dans des futurs travaux, nous nous emploierons à concevoir et à évaluer d'autres procédures de décodage à utiliser avec la méthode BR pour la production de solutions d'une qualité supérieure à celles que nous obtenons avec la procédure de décodage examinée ici. Dans cette recherche, nous tenterons de résoudre en

même temps le problème de minimisation de la taille totale d'échantillon en fonction d'une précision recherchée, comme l'ont fait Lavallée et Hidiroglou (1988). Disons enfin que, dans de nouveaux travaux empiriques, nous pourrions varier les tailles d'échantillon pour les diverses populations à l'étude comme l'ont fait Kozak (2004) et Gunning et Horgan (2004).

Annexe A

Tableau A1
Description des 27 populations considérées dans l'expérience numérique

Population	Description
AgrMinas	Production agricole des municipalités de l'État de Minas Gerais au Brésil selon le recensement agricole de 2006. Variable de stratification : superficie ensemencée.
BeefFarms	Élevages australiens pour la boucherie stratifiés en sept régions industrielles selon Chambers et Dunstan (1986). Variable de stratification : taille des élevages.
Beta103	Population en simulation tirée d'une distribution bêta avec les paramètres $a = 10$ et $b = 3$ selon Keskintürk et Er (2007).
Chi5	Population en simulation tirée d'une distribution khi-carré avec $df = 5$ selon Keskintürk et Er (2007).
Café	Plantations de café de l'État de Paraná au Brésil dans le recensement agricole de 1996 selon de Moura Brito et coll. (2015). Variable de stratification : nombre de caféiers.
CensoCO	Données du recensement des écoles de 2012 au Brésil pour la région centre-ouest. Variable de stratification : nombre de salles de classe.
Débiteurs	Population de débiteurs d'une entreprise irlandaise selon Er (2011). Variable de stratification : passif déclaré par les débiteurs irlandais.
HHinctot	Population de valeurs brutes de revenu familial (avant impôt sur le revenu) dans l'Enquête sur les dépenses des familles de 2001 de Statistique Canada selon Er (2011).
Iso2004	Données obtenues par la Chambre industrielle d'Istanbul sur les ventes nettes de 487 entreprises industrielles de Turquie parmi les 500 entreprises les plus importantes en 2004 d'après Keskintürk et Er (2007). Variable de stratification : ventes nettes.
Kozak1, Kozak3, Kozak4	Populations considérées par Kozak et Verma (2006). Variable de stratification : formule $X = \exp(Z)$, où Z est une réalisation d'une variable aléatoire normale.
ME84	Données de Särndal, Swensson et Wretman (1992) selon Er (2011). Variable de stratification : nombre d'employés municipaux en 1984.
EMCD	Population en simulation tirée de l'Enquête mensuelle sur le commerce de détail de Statistique Canada selon Er (2011). Variable de stratification : mesure de taille employée pour les détaillants canadiens dans cette enquête de Statistique Canada; on crée cette mesure en combinant une information d'enquête indépendante à trois variables administratives des déclarations de revenu des sociétés.
P75	Population en milliers de 284 municipalités suédoises en 1975 selon Er (2011). Variable de stratification : population en milliers.
P100e10	Population en simulation tirée d'une distribution normale avec $\mu = 100$ et $\sigma = 10$ selon Keskintürk et Er (2007).
pop1076	Population extraite de l'enquête annuelle sur la fabrication au Brésil selon de Moura Brito et coll. (2017b). Variable de stratification : nombre d'employés.
pop1616	Population extraite de l'enquête annuelle sur la fabrication au Brésil selon de Moura Brito et coll. (2017b). Variable de stratification : nombre d'employés.
pop2911	Population extraite de l'enquête annuelle sur la fabrication au Brésil selon de Moura Brito et coll. (2017b). Variable de stratification : nombre d'employés.
Pop500	Population $N = 500$ en simulation tirée de la distribution log normale $X = e^z$ avec Z normal, $\mu = 4$ et $\sigma^2 = 2{,}7$ selon Hedlin (2000).
Pop800	Population $N = 800$ en simulation tirée de la distribution log normale $X = e^z$ avec Z normal, $\mu = 4$ et $\sigma^2 = 2{,}7$ selon Hedlin (2000).
REV84	Valeur des bâtiments en millions de couronnes suédoises dans 284 municipalités de Suède en 1984 selon Er (2011). Variable de stratification : produit de la fiscalité municipale en 1985.
SugarCaneFarms	Plantations de canne à sucre en Australie selon Chambers et Dunstan (1986). Variable de stratification : récolte totale de canne à sucre.
USbanks	Actif en millions de dollars américains des grandes banques commerciales nord-américaines selon Er (2011). Variable de stratification : ressources en millions de dollars des grandes banques commerciales américaines.
UScities	Population en milliers des villes nord-américaines en 1940 selon Er (2011). Variable de stratification : population en milliers.
UScolleges	Nombre d'étudiants en quatrième année dans les facultés américaines en 1952-1953 selon Er (2011). Variable de stratification : nombre d'étudiants.
Swiss	Données sur les municipalités de Suisse en 2003 avec le package SamplingStrata en R. Variable de stratification : superficie en culture.

Bibliographie

Baillargeon, S., et Rivest, L.-P. (2014). Stratification: Univariate stratification of survey populations. Package R version 2.2-5. http://CRAN.R-project.org/package=stratification.

Bankier, M.D. (1988). Power allocations: Determining sample sizes for sub-national areas. *The American Statistician*, 42, 174-177.

Chambers, R., et Dunstan, R. (1986). Estimating distribution functions from survey data. *Biometrika*, 73, 3, 597-604.

Cochran, W. (1977). *Sampling Techniques, 3rd Ed*. New York: John Wiley & Sons, Inc.

Dalenius, T. (1951). The problem of optimum stratification. *Scandinavian Actuarial Journal*, 1-2, 133-148.

Dalenius, T., et Hodges, J. (1959). Minimum variance stratification. *Journal of the American Statistical Association*, 285, 54, 88-101.

De Moura Brito, J.A.M., do Nascimento Silva, P.L. et da Veiga, T.M. (2017a). Stratbr: Optimal Stratification in Stratified Sampling. Package R version 1.2. https://CRAN.R-project.org/package=stratbr.

De Moura Brito, J.A.M., do Nascimento Silva, P.L., Silva Semaan, G. et Maculan, N. (2015). Application des formulaires de la programmation en nombres entiers à la répartition optimale dans l'échantillonnage stratifié. *Techniques d'enquête*, 41, 2, 451-467. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2015002/article/14249-fra.pdf.

De Moura Brito, J.A.M., Maculan, N., Lila, M. et Montenegro, F. (2010b). An exact algorithm for the stratification problem with proportional allocation. *Optimization Letters*, 4, 185-195.

De Moura Brito, J.A.M., Ochi, L., Montenegro, F. et Maculan, N. (2010a). An iterative local search approach applied to the optimal stratification problem. *International Transactions in Operational Research*, 17, 6, 753-764.

De Moura Brito, J.A.M., Silva Semaan, G., Fadel, A. et Brito, L.R. (2017b). An optimization approach applied to the optimal stratification problem. *Communications in Statistics: Simulation and Computation*, 46, 4419-4451.

Ekman, G. (1959). An approximation useful in univariate stratification. *The Annals of Mathematical Statistics*, 30, 1, 219-229.

Er, S. (2011). Comparison of the efficiency of the various algorithms in stratified sampling when the initial solutions are determined with geometric method. *International Journal of Statistics and Applications*, 1, 1, 1-10.

Er, S., Keskintürk, T. et Daly, C. (2010). GA4Stratification: A genetic algorithm approach to determine stratum boundaries and sample sizes of each stratum in stratified sampling. Package R version 1.0. http://CRAN.R-project.org/package=stratification.

Festa, P. (2013). A biased random-key genetic algorithm for data clustering. *SI:BIOCOMP, Math. Biosci.*, 245, 1, 76-85.

Gonçalves, J.F., et Resende, M.G.C. (2004). An evolutionary algorithm for manufacturing cell formation. *Comput. Ind. Eng*, 47, 247-273.

Gonçalves, J.F., et Resende, M. (2011). Biased random-key genetic algorithms for combinatorial optimization. *Journal of Heuristics*, 17, 487-525.

Gonçalves, J.F., Mendes, J.J.M. et Resende, M.G.C. (2005). A hybrid genetic algorithm for the job shop scheduling problem. *Eur. J. Oper. Res*, 167, 77-95.

Gunning, P., et Horgan, J.M. (2004). Un nouvel algorithme pour la construction de bornes de stratification dans les populations asymétriques. *Techniques d'enquête*, 30, 2, 177-185. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2004002/article/7749-fra.pdf.

Hedlin, D. (2000). A procedure for stratification by an extended Ekman rule. *Journal of Official Statistics*, 16, 15-29.

Hidiroglou, M.A. (1986). The construction of a self-representing stratum of large units in survey design. *The American Statistician*, 1, 40, 27-31.

Hidiroglou, M.A., et Kozak, M. (2017). Stratification of skewed populations: A comparison of optimisation-based versus approximate methods. *Revue Internationale de Statistique*, https://doi.org/10.1111/insr.12230.

Keskintürk, T., et Er, S. (2007). A genetic algorithm approach to determine stratum boundaries and sample sizes of each stratum in stratified sampling. *Computational Statistics & Data Analysis*, 52, 53-67.

Khan, M.G.M., Nand, N. et Ahmad, N. (2008). Détermination des bornes optimales de strate au moyen de la programmation dynamique. *Techniques d'enquête*, 34, 2, 227-236. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2008002/article/10761-fra.pdf.

Kozak, M. (2004). Optimal stratification using random search method in agricultural surveys. *Statistics in Transition*, 6, 5, 797-806.

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

Kozak, M. (2014). Comparison of random search method and genetic algorithm for stratification. *Communications in Statistics – Simulation and Computation*, 43, 2, 249-253.

Kozak, M., et Verma, M.R. (2006). Approche de la stratification par une méthode géométrique et par optimisation : une comparaison de l'efficacité. *Techniques d'enquête*, 32, 2, 177-183. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2006002/article/9550-fra.pdf.

Lavallée, P., et Hidiroglou, M.A. (1988). Sur la stratification de populations asymétriques. *Techniques d'enquête*, 14, 1, 35-45. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1988001/article/14602-fra.pdf.

Lohr, S. (2010). *Sampling: Design and Analysis, 2nd Ed*. Washington: Duxbury Press.

Oliveira, R.M., Chaves, A.A. et Lorena, L.A.N. (2017). A comparison of two hybrid methods for constrained clustering problems. *Applied Soft Computing*, 54, 256-266.

Rao, D.K., Khan, M.G.M. et Reddy, K.G. (2014). Optimum stratification of a skewed population. *International Journal of Mathematical, Computational, Physical and Quantum Engineering*, 8, 3, 497-500.

Rivest, L.-P. (2002). Une généralisation de l'algorithme de Lavallée et Hidiroglou pour la stratification dans les enquêtes auprès des entreprises. *Techniques d'enquête*, 28, 2, 207-214. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2002002/article/6432-fra.pdf.

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

Spears, W., et De Jong, K. (1991). On the virtues of parameterized uniform crossover. Dans *Proceedings of the Fourth International Conference on Genetic Algorithms*, 230-236.