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

Techniques d'enquête 41-1

Cadre généralisé pour la détermination des probabilités d'inclusion optimales dans les plans de sondage à un degré pour des enquêtes à plusieurs variables et plusieurs domaines

par Piero Demetrio Falorsi et Paolo Righi

Date de diffusion : le 29 juin 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 infostats@statcan.gc.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 « À propos de nous » > « Notre organisme » > « Offrir des services aux Canadiens »

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)

Techniques d'enquête, juin 2015
Vol. 41, n° 1, pp. 225-247
Statistique Canada, n° 12-001-X au catalogue

Cadre généralisé pour la détermination des probabilités d'inclusion optimales dans les plans de sondage à un degré pour des enquêtes à plusieurs variables et plusieurs domaines

Piero Demetrio Falorsi et Paolo Righi[1]

Résumé

L'article décrit un cadre généralisé de calcul des probabilités d'inclusion optimales dans divers contextes d'enquête dans lesquels il est requis de diffuser des estimations d'enquête d'une précision préétablie pour de multiples variables et domaines d'intérêt. Le cadre permet de définir des plans de sondage stratifiés classiques ou incomplets. Les probabilités d'inclusion optimales sont obtenues en minimisant les coûts au moyen d'un algorithme qui garantit l'établissement de bornes pour les erreurs d'échantillonnage au niveau du domaine, en supposant que les variables d'appartenance au domaine sont disponibles dans la base de sondage. Les variables cibles sont inconnues, mais peuvent être prédites au moyen de modèles de superpopulation appropriés. L'algorithme tient compte correctement de l'incertitude de ces modèles. Certaines expériences basées sur des données réelles montrent les propriétés empiriques de l'algorithme.

Mots-clés : Répartition optimale; stratification multidimensionnelle; estimations de domaine; échantillonnage équilibré.

1 Introduction

Les enquêtes menées dans le contexte de la statistique officielle produisent fréquemment un grand nombre d'estimations qui ont trait à différents paramètres d'intérêt ainsi qu'à des domaines d'estimation d'un niveau de détail très élevé. Lorsque des variables indicatrices de domaine sont disponibles pour chaque unité d'échantillonnage figurant dans la base de sondage, le concepteur du plan de sondage peut essayer de sélectionner un échantillon dans lequel la taille de chaque domaine est fixée. Dans ces conditions, il est possible d'obtenir des estimations directes pour chaque domaine et de contrôler les erreurs d'échantillonnage au niveau du domaine. Nous présentons ici un cadre *unifié* et *général* pour définir les *probabilités d'inclusion optimales* pour les *plans d'échantillonnage à un degré* lorsqu'on connaît les variables d'appartenance au domaine à l'étape de l'établissement du plan. Il pourrait s'agir du scénario le plus fréquent dans les enquêtes auprès des établissements et dans d'autres contextes d'enquête, comme les enquêtes agricoles ou les enquêtes sociales si les domaines sont de nature géographique (par exemple, type de municipalité, région, province, etc.). La progression croissante de l'intégration des données des registres administratifs et des bases de sondage pourrait aussi rendre l'approche présentée ici plus applicable aux enquêtes sociales. La proposition pourrait être utile pour la planification d'un sondage de deuxième phase optimal si l'on a recueilli les données sur les variables d'appartenance au domaine à la première phase.

Le problème de l'établissement de plans de sondage optimaux a été abordé dans certains articles récents. Gonzalez et Eltinge (2010) donnent un aperçu intéressant des approches en vue de définir des stratégies d'échantillonnage optimales. Le problème d'optimisation est habituellement traité dans le

1. Piero Demetrio Falorsi, FAO, Viale delle Terme di Caracalla, Roma. Courriel : piero.falorsi@fao.org; Paolo Righi, ISTAT Via C. Balbo 16, 00184 Roma. Courriel : parighi@istat.it.

contexte de l'échantillonnage stratifié avec taille d'échantillon fixe dans chaque strate. La répartition optimale sous échantillonnage stratifié pour une population univariée est bien décrite dans la littérature sur l'échantillonnage (Cochran 1977). Dans les cas multivariés, où plus d'une caractéristiques doivent être mesurées sur chaque unité échantillonnée, la répartition optimale pour les caractéristiques individuelles est de peu d'intérêt pratique, à moins que les diverses caractéristiques étudiées soient fortement corrélées. Il en est ainsi parce qu'une répartition optimale pour une caractéristique est généralement loin de l'être pour les autres. La multidimensionalité du problème mène à la définition d'une méthode de répartition de compromis (Khan, Mati et Ahsan 2010) associée à une perte de précision comparativement aux répartitions optimales individuelles. Plusieurs auteurs ont discuté de divers critères permettant d'obtenir une répartition de compromis réalisable – voir, par exemple, Kokan et Khan (1967), Chromy (1987), Bethel (1989), Falorsi et Righi (2008), Falorsi, Orsini et Righi (2006) et Choudhry, Rao et Hidiroglou (2012).

Récemment, certains articles ont porté sur la recherche des probabilités d'inclusion optimales sous échantillonnage équilibré (Tillé et Favre 2005; Chauvet, Bonnéry et Deville 2011), une classe générale de plans d'échantillonnage qui inclut les plans d'échantillonnage stratifiés comme cas particuliers. Plus précisément, Chauvet et coll. (2011) proposent l'adoption de l'algorithme du point fixe pour définir les probabilités d'inclusion optimales. Néanmoins, les articles susmentionnés n'abordent pas le cas où les variables d'équilibrage dépendent des probabilités d'inclusion et ne présentent qu'une solution partielle au problème dû au fait que la variance d'échantillonnage est une fonction **implicite** des probabilités d'inclusion. Choudhry et coll. (2012) propose un algorithme de répartition optimale pour les estimations de domaine sous échantillonnage stratifié (si les domaines d'estimation ne recoupent pas les strates). Leur algorithme représente un cas particulier de l'approche que nous proposons. Les conditions méthodologiques illustrées ici représentent une amélioration considérable par rapport à la version antérieure de la méthodologie décrite dans Falorsi et Righi (2008) qui ne tenait compte que du cas où les valeurs des variables d'intérêt étaient connues et où la mesure de la précision était exprimée par la variance sous le plan; en outre, la version antérieure ne tenait pas compte du fait que la variance sous le plan, bornée dans le problème d'optimisation, est une fonction implicite des probabilités d'inclusion. Le présent article porte sur le cas plus réaliste où les variables d'intérêt ne sont pas connues et doivent être estimées. En outre, il traite explicitement le problème découlant du fait que les variances anticipées sont des fonctions implicites des probabilités d'inclusion. Le nouvel algorithme d'optimisation peut être exécuté facilement, parce qu'il est fondé sur une décomposition générale de la mesure de la précision. Nous proposons un plan d'échantillonnage général qui englobe la plupart des plans d'échantillonnage à un degré adoptés dans les enquêtes réelles, par exemple l'échantillonnage aléatoire simple sans remise (EASSR), l'EASSR stratifié, l'échantillonnage PPT stratifié, les plans avec stratification incomplète, etc. Le cadre est fondé sur l'utilisation conjointe de *plans d'échantillonnage équilibrés* (Deville et Tillé 2004) qui, suivant les différentes définitions des équations d'équilibrage, représentent une vaste gamme de plans d'échantillonnage et de *modèles de superpopulation pour la prédiction* des valeurs inconnues des variables d'intérêt. La présentation de l'article est la suivante. À la section 2, nous exposons les définitions et la notation. À la section 3 et à la section 4, nous illustrons le plan d'échantillonnage et la variance anticipée. À la section 5, nous décrivons l'algorithme utilisé pour définir les probabilités d'inclusion optimales. À la section 6, nous illustrons les propriétés empiriques de l'algorithme au moyen de certaines expériences fondées sur des données réelles sur les entreprises. Enfin, à la section 7, nous présentons les conclusions.

2 Définitions et notation

À la présente section, nous exposons les concepts du ***domaine d'estimation*** et du ***domaine planifié*** qui jouent un rôle clé dans le cadre présenté ici.

Soit U la population de référence de N éléments et soit $U_d\ (d = 1,\ldots,D)$ un ***domaine d'estimation***, c'est-à-dire une sous-population générique de U contenant N_d éléments, pour laquelle des estimations distinctes doivent être calculées. Soit y_{rk} la valeur de la $r^{\text{e}}\ (r = 1,\ldots,R)$ variable d'intérêt attachée à la k^{e} unité de population et soit γ_{dk} l'indicateur d'appartenance au domaine pour l'unité k défini par

$$\gamma_{dk} = \begin{cases} 1 & \text{si } k \in U_d \\ 0 & \text{autrement} \end{cases}. \tag{2.1}$$

Nous supposons que les valeurs de γ_{dk} sont disponibles dans la base de sondage et que plus d'une valeur de $\gamma_{dk}\ (d = 1,\ldots,D)$ peut être égale à 1 pour chaque unité k; par conséquent, les domaines d'estimation peuvent se chevaucher.

Les paramètres d'intérêt sont les $D \times R$ totaux de domaine

$$t_{(dr)} = \sum_{k\in U} y_{rk}\gamma_{dk}\ (r = 1,\ldots,R; d = 1,\ldots,D). \tag{2.2}$$

Soit $p(\cdot)$ un plan d'échantillonnage sans remise à un degré et $\boldsymbol{\pi} = (\pi_1,\ldots,\pi_k,\ldots,\pi_N)'$ le vecteur de dimension N des probabilités d'inclusion. Soit s l'échantillon sélectionné avec la probabilité $p(s)$. Désignons par $U_h\ (h = 1,\ldots,H)$ la sous-population de taille $N_h = \sum_{k\in U_h}\delta_{hk}$ où $\delta_{hk} = 1$ si $k \in U_h$ et $\delta_{hk} = 0$ autrement.

Nous nous concentrons que les plans d'échantillonnage à taille fixe qui sont ceux qui satisfont

$$\sum_{k\in s}\boldsymbol{\delta}_k = \mathbf{n}, \tag{2.3}$$

où $\boldsymbol{\delta}_k = (\delta_{1k},\ldots,\delta_{hk},\ldots,\delta_{Hk})'$ et $\mathbf{n} = (n_1,\ldots,n_h,\ldots,n_H)'$ est le vecteur de nombres entiers définissant les tailles d'échantillon fixées au moment de l'établissement du plan d'échantillonnage. Puisque la taille d'échantillon n_h, qui correspond à U_h, ne varie pas d'une sélection d'échantillon à l'autre, la sous-population U_h sera appelée ***domaine planifié*** dans la suite de l'exposé. Une condition nécessaire, mais non suffisante, pour s'assurer que (2.3) soit satisfaite est que le vecteur $\boldsymbol{\pi}$ soit tel que

$$\sum_{k\in U}\pi_k\boldsymbol{\delta}_k = \mathbf{n}. \tag{2.4}$$

Dans notre configuration, les domaines planifiés peuvent se chevaucher; par conséquent, l'unité k peut posséder plus d'une valeur $\delta_{hk} = 1$ (pour $h = 1,\ldots,H$). Supposons que les valeurs de δ_{hk} sont connues et disponibles dans la base de sondage pour toutes les unités de la population. Supposons en outre que la matrice $(\boldsymbol{\delta}_1,\ldots,\boldsymbol{\delta}_k,\ldots,\boldsymbol{\delta}_N)'$ de dimensions $N \times H$ n'est pas singulière.

Les domaines planifiés et leur relation avec les domaines d'estimation jouent un rôle central dans notre cadre généralisé. Nous supposons que les domaines d'estimation peuvent être définis comme un agrégat de domaines planifiés complets, de sorte que la taille d'échantillon *prévue* dans le d^{e} domaine

d'estimation U_d, disons n_d, peut être obtenue sous forme d'un agrégat simple des tailles d'échantillon prévues des domaines planifiés inclus. Enfin, soit $\hat{t}_{(dr)}$ l'estimateur d'Horvitz-Thompson (HT) de $t_{(dr)}$ avec

$$\hat{t}_{(dr)} = \sum_{k \in s} \frac{1}{\pi_k} y_{rk} \gamma_{dk}. \tag{2.5}$$

Un exemple tiré des enquêtes-entreprises. Supposons que l'on doive calculer les estimations d'enquête séparément en considérant trois types de domaines, à savoir la *région* (20 modalités), l'*activité économique* (2 modalités : biens ou services) et la *taille de l'entreprise* (3 modalités : petite, moyenne ou grande). Autrement dit, il existe $D = 20 + 2 + 3 = 25$ domaines d'estimation chevauchants possibles. Les domaines planifiés peuvent être définis selon différentes options.

Option 1. Le domaine planifié unique U_h est défini par une intersection spécifique des catégories des domaines d'estimation. Dans ce cas, $H = 20 \times 2 \times 3 = 120$ domaines planifiés sont définis. Ils représentent une partition particulière de U. Les domaines planifiés ne se chevauchent pas et $\sum_h \delta_{hk} = 1$.

Option 2. Les domaines planifiés U_h coïncident avec les domaines d'estimation. Par conséquent, $H = D = 25$ et les $\boldsymbol{\delta}'_k$ sont définis comme des vecteurs contenant trois 1, de sorte que $\sum_h \delta_{hk} = 3$. Rappelons que les domaines planifiés se chevauchent.

Option 3. Les domaines planifiés U_h sont définis par *i*) la région selon l'activité économique et *ii*) l'activité économique selon la taille d'entreprise; alors, $H = (20 \times 2) + (2 \times 3) = 46$ avec $\sum_h \delta_{hk} = 2$.

D'autres relations intermédiaires entre les domaines d'estimation et les domaines planifiés sont possibles.

Soulignons que les domaines planifiés représentent le fondement pour la définition de classes plus générales de plans d'échantillonnage. Par exemple, les ***plans d'échantillonnage stratifiés*** requièrent que les domaines planifiés ne se chevauchent pas, car $\sum_h \delta_{hk} = 1$ et que chaque U_h est désignée comme étant une strate. Par conséquent, l'option 1 de l'exemple qui précède nous mène à définir un plan d'échantillonnage stratifié. En outre, les strates définies comme dans l'option 1 servent de fondement à ce que l'on appelle un « plan d'échantillonnage stratifié multimensionnel » (Winkler 2001).

Si $\sum_h \delta_{hk} > 1$, les tailles d'échantillon des domaines planifiés définies dans l'option 1 (strates) ne sont pas strictement contrôlées. Néanmoins, elles restent contrôlées à un niveau agrégé. Dans l'option 2 de l'exemple susmentionné, les tailles d'échantillon sont contrôlées uniquement pour les domaines d'estimation; par contre, dans l'option 3, les tailles d'échantillon sont contrôlées pour les sous-ensembles de deux partitions différentes, définies par *i*) la région selon l'activité économique et *ii*) l'activité économique selon la taille d'entreprise. En nous basant sur la définition de Winkler, nous désignons les plans utilisant ces types de domaines planifiés comme étant des ***plans d'échantillonnage stratifiés multidimensionnels incomplets*** **(ESI)**.

3 Échantillonnage

Soit $\mathbf{z}_k$ un vecteur de variables auxiliaires disponible pour toutes les unités $k \in U$. Un plan d'échantillonnage $p(s)$ est dit équilibré sur les variables auxiliaires si, et seulement si, il satisfait les ***équations d'équilibrage*** suivantes

$$\sum_{k \in s} \frac{\mathbf{z}_k}{\pi_k} = \sum_{k \in U} \mathbf{z}_k \tag{3.1}$$

pour chaque échantillon s tel que $p(s) > 0$ (Deville et Tillé 2004). Selon les variables auxiliaires et les probabilités d'inclusion, l'équation (3.1) peut être exactement ou approximativement satisfaite dans chaque échantillon possible; par conséquent, un plan d'échantillonnage équilibré n'existe pas toujours. En spécifiant

$$\mathbf{z}_k = \pi_k \boldsymbol{\delta}_k, \tag{3.2}$$

les équations (3.1) deviennent

$$\sum_{k \in s} \boldsymbol{\delta}_k = \sum_{k \in U} \pi_k \boldsymbol{\delta}_k. \tag{3.3}$$

Dans ce cas, les équations d'équilibrage stipulent que la taille d'échantillon réalisée dans chaque sous-population U_h est égale à la taille prévue. Dans différents contextes, Ernst (1989) et Deville et Tillé (2004; page 905 Section 7.3) ont prouvé que *i*) sous la spécification (3.2) et *ii*) si le vecteur des tailles prévues d'échantillon, données par $\mathbf{n} = \sum_{k \in U} \pi_k \boldsymbol{\delta}_k$, ne contient que des nombres entiers, alors un plan d'échantillonnage équilibré existe toujours. La spécification (3.2) définit des plans d'échantillonnage qui garantissent le respect de l'équation (2.4) sur laquelle nous souhaitons nous concentrer. Deville et Tillé (2004, pages 895 et 905), Deville et Tillé (2005, page 577) et Tillé (2006, page 168) ont montré que plusieurs plans d'échantillonnage habituels peuvent être considérés comme des cas particuliers de l'échantillonnage équilibré, en définissant de manière appropriée les vecteurs $\boldsymbol{\pi}$ et $\boldsymbol{\delta}_k$ de l'équation (3.2). Ces problèmes sont illustrés à la remarque 4.2 et à la section 6. Des échantillons équilibrés peuvent être tirés par la méthode du cube (Deville et Tillé 2004). Cette méthode facilite grandement la sélection sous des plans d'échantillonnage stratifiés incomplets en permettant de contourner les inconvénients de calcul des méthodes fondées sur des algorithmes de programmation linéaire (Lu et Sitter 2002). La méthode du cube satisfait exactement les équations (3.1) quand la spécification (3.2) est vérifiée et que $\mathbf{n}$ est un vecteur de nombres entiers. Dans les cas de l'EASSR et de l'EASSRS, on peut utiliser les méthodes classiques de sélection de l'échantillon, ainsi que la méthode du cube. Deville et Tillé (2005) proposent une approximation de la variance pour l'estimateur HT sous-échantillonnage équilibré

$$E_p \left(\hat{t}_{(dr)} - t_{(dr)}\right)^2 \cong [N/(N-H)]\left[\sum_{k \in U} (1/\pi_k - 1)\, \eta^2_{(dr)k}\right] \tag{3.4}$$

où E_p désigne l'espérance d'échantillon et

$$\eta_{(dr)k} = y_{rk}\gamma_{dk} - \pi_k \boldsymbol{\delta}'_k \left[\mathbf{A}(\boldsymbol{\pi})\right]^{-1} \sum_{j \in U} \pi_j \left(1/\pi_j - 1\right) \boldsymbol{\delta}_j y_{rk} \gamma_{dk} \tag{3.5}$$

avec

$$\mathbf{A}(\boldsymbol{\pi}) = \sum_{j \in U} \boldsymbol{\delta}_j \boldsymbol{\delta}'_j \pi_j (1 - \pi_j). \tag{3.6}$$

Les résultats de simulations donnés récemment dans Breidt et Chauvet (2011) confirment que l'équation (3.4) représente une bonne approximation de la variance d'échantillonnage quand les équations d'équilibrage sont satisfaites exactement. L'estimation de la variance est étudiée dans Deville et Tillé (2005).

4 Variance anticipée

Avant l'échantillonnage, les valeurs de y_{rk} ne sont pas connues et la variance exprimée par la formule (3.4) ne peut pas être utilisée pour planifier la précision de l'échantillonnage à la phase d'élaboration du plan. En pratique, il est nécessaire d'obtenir des valeurs substitutives ou de prédire les valeurs y_{rk} en se basant sur des modèles de superpopulation qui exploitent l'information auxiliaire. La disponibilité croissante d'information auxiliaire (obtenue par intégration des registres administratifs et des bases de sondage) facilite l'usage des prédictions. Sous inférence fondée sur un modèle, on suppose que les valeurs de y_{rk} sont la réalisation d'un modèle de superpopulation M. Le modèle que nous étudions est de la forme suivante :

$$\begin{cases} y_{rk} = f_r(\mathbf{x}_k; \boldsymbol{\beta}_r) + u_{rk} \\ E_M(u_{rk}) = 0 \ \ \forall k; \ E_M(u_{rk}^2) = \sigma_{rk}^2; \ E_M(u_{rk}, u_{rl}) = 0 \ \ \forall k \neq l \end{cases}, \tag{4.1}$$

où $\mathbf{x}_k$ est un vecteur de variables explicatives (disponibles dans la base de sondage), $\boldsymbol{\beta}_r$ est un vecteur de coefficients de régression et $f_r(\mathbf{x}_k; \boldsymbol{\beta}_r)$ est une fonction connue, u_{rk} est le terme d'erreur et $E_M(\cdot)$ désigne l'espérance sous le modèle. Les paramètres $\boldsymbol{\beta}_r$ et les variances σ_{rk}^2 sont supposés connus, quoiqu'en pratique ils sont habituellement estimés. Le modèle (4.1) est spécifique à une variable, et l'on peut utiliser différents modèles pour différentes variables sans créer de difficultés supplémentaires. Comme mesure de l'incertitude, nous considérons la *variance anticipée* (VA) (Isaki et Fuller 1982) :

$$\mathrm{VA}(\hat{t}_{(dr)}) = E_M E_p (\hat{t}_{(dr)} - t_{(dr)})^2. \tag{4.2}$$

Une expression générale pour la VA sous des modèles linéaires a été établie par Nedyalkova et Tillé (2008). Leur formulation s'obtient en considérant une fonction linéaire $f_r(\cdot)$ et un ensemble unique de variables auxiliaires, $\mathbf{x}_k$, utilisé à la fois pour la prédiction des valeurs de y et pour l'équilibrage de l'échantillon. Dans notre contexte, nous avons introduit $\mathbf{x}_k$ et $\mathbf{z}_k = \pi_k \boldsymbol{\delta}_k$, en soulignant que les variables auxiliaires peuvent être différentes pour la prédiction et l'équilibrage. Les variables $\mathbf{x}_k$ doivent être aussi prédictives de y_{rk} que possible, tandis que les variables $\mathbf{z}_k$ jouent un rôle instrumental dans le contrôle des tailles d'échantillon pour les sous-populations.

Dans le contexte considéré ici, en insérant la variance approximative (3.4) dans l'équation (4.2), nous obtenons l'expression approximative de la VA :

$$\mathrm{VAA}(\hat{t}_{(dr)}) = [N/(N-H)] \sum_{k \in U} (1/\pi_k - 1) E_M(\eta_{(dr)k}^2), \tag{4.3}$$

où les termes $\eta^2_{(dr)k}$ de (3.4) sont remplacés par $E_M\left(\eta^2_{(dr)k}\right)$. En définissant

$$\tilde{y}_{rk} = f_r\left(\mathbf{x}_k;\boldsymbol{\beta}_r\right), \tag{4.4}$$

nous pouvons reformuler l'équation (4.3) sous la forme

$$\mathrm{VAA}\left(\hat{t}_{(dr)}\right) = [N/(N-H)]\left[\sum_{k\in U}\frac{1}{\pi_k}\left(\tilde{y}^2_{rk}+\sigma^2_{rk}\right)\gamma_{dk} - \sum_{k\in U}\left(\tilde{y}^2_{rk}+\sigma^2_{rk}\right)\gamma_{dk} - \mathrm{VAA}_{3(dr)}\right], \tag{4.5}$$

où la troisième composante de variance de $\mathrm{VAA}\left(\hat{t}_{(dr)}\right)$ est

$$\begin{aligned}\mathrm{VAA}_{3(dr)} &= \sum_{k\in U}(1-\pi_k)\,a_{(dr)k}(\boldsymbol{\pi})\left[2\tilde{y}_{rk}\gamma_{dk} - \pi_k a_{(dr)k}(\boldsymbol{\pi})\right] \\ &+ \sum_{k\in U}(1-\pi_k)\left[2b_{(dr)k}(\boldsymbol{\pi}) - \pi_k c_{(dr)k}(\boldsymbol{\pi})\right]\end{aligned} \tag{4.6}$$

et $a_{(dr)k}(\boldsymbol{\pi})$, $b_{(dr)k}(\boldsymbol{\pi})$ et $c_{(dr)k}(\boldsymbol{\pi})$ sont des nombres réels définis respectivement par les équations (A1.4), (A1.7) et (A1.8) de l'annexe A1.

Remarque 4.1. L'expression (4.5) est une formule dont le calcul est laborieux mais, à toute fin pratique, ce calcul peut être simplifié au moyen d'une légère approximation à la hausse en posant que $b_{(dr)k}(\boldsymbol{\pi}) = c_{(dr)k}(\boldsymbol{\pi}) = 0$ dans (4.6). La preuve est donnée à l'annexe A3. Une approximation à la hausse est un choix prudent dans ces conditions, puisqu'il évite le risque de définir une taille d'échantillon insuffisante pour la précision attendue.

Remarque 4.2. Le plan EASSRS est obtenu si les domaines planifiés définissent une partition unique de la population (Option 1 de l'exemple à la section 2) et que le modèle (4.1) est spécifié de façon que les valeurs prédites soient $\tilde{y}_{rk} = \bar{Y}_{rh}$ avec $\sigma^2_{rk} = \sigma^2_{rh}$ (pour $k \in U_h$). La VAA devient

$$\mathrm{VAA}\left(\hat{t}_{(dr)}\right) = [N/(N-H)]\sum_{d=1}^{D}\sum_{h\in H_d}\sigma^2_{rh}N_h\left(N_h/n_h - 1\right), \tag{4.7}$$

où H_d est l'ensemble de domaines planifiés inclus dans U_d (voir l'annexe A4). Notons que l'expression (4.7) concorde avec le *résultat 2* de Nedyalkova et Tillé (2008), sauf pour le terme $N/(N-H)$. Si $[N/(N-H)](1/N_h) \approx 1/(N_h - 1)$, l'expression (4.7) approximerait la variance de l'estimation HT sous le plan EASSRS. Il est prouvé que l'approximation susmentionnée est vraie quand le nombre de domaines H reste petit comparativement à la taille globale de la population N, et que les tailles de domaine N_h sont grandes.

5 Détermination des probabilités d'inclusion optimales

Le vecteur des valeurs de π est déterminé en résolvant le problème d'optimisation suivant :

$$\begin{cases}\mathrm{Min}\left(\sum_{k\in U}\pi_k c_k\right) & \\ \mathrm{VAA}\left(\hat{t}_{(dr)}\right) \le \bar{V}_{(dr)} & (d=1,\ldots,D;\, r=1,\ldots,R), \\ 0 < \pi_k \le 1 & (k=1,\ldots,N)\end{cases} \tag{5.1}$$

où c_k est le coût de la collecte de l'information auprès de l'unité k et $\bar{V}_{(dr)}$ est un seuil de variance fixe correspondant à $\hat{t}_{(dr)}$. Le système (5.1) minimise le coût prévu en s'assurant que les variances anticipées soient bornées et que les probabilités d'inclusion soient comprises entre 0 et 1. Si toutes les valeurs de c_k sont des constantes égales à 1, le problème (5.1) minimise la taille d'échantillon. Nous notons que, dans le problème (5.1), les variances σ^2_{rk} figurant dans $\mathrm{VAA}(\hat{t}_{(dr)})$ sont traitées comme étant connues; en pratique, elles doivent être estimées. À la section 6, nous procédons à une évaluation empirique afin d'étudier la sensibilité de la taille d'échantillon globale en utilisant différentes valeurs estimées de σ^2_{rk}.

Pour résoudre (5.1), nous réarrangeons les contraintes d'inégalité afin d'obtenir

$$\sum_{k\in U} \frac{\left(\tilde{y}^2_{rk} + \sigma^2_{rk}\right)\gamma_{dk}}{\pi_k} \le \frac{N-H}{N}\bar{V}_{(dr)} + \sum_{k\in U}\left(\tilde{y}^2_{rk} + \sigma^2_{rk}\right)\gamma_{dk} + \mathrm{VAA}_{3(dr)}. \qquad (5.2)$$

En fixant de manière appropriée les valeurs de $\mathrm{VAA}_{3(dr)}$, le problème d'optimisation devient un problème linéaire convexe séparé (PLCS) classique (Boyd et Vandenberghe 2004). La figure 5.1 illustre le diagramme de cheminement de l'algorithme (un logiciel prototype dans lequel est mis en œuvre l'algorithme est disponible à l'adresse http://www.istat.it/it/strumenti/metodi-e-software/software), qui est structuré en deux boucles emboîtées : la **boucle externe** (BE) et la **boucle interne** (BI). Les deux boucles sont mises à jour en suivant un schéma d'algorithme *du point fixe*. La convergence sous certaines approximations est démontrée à l'annexe A2.

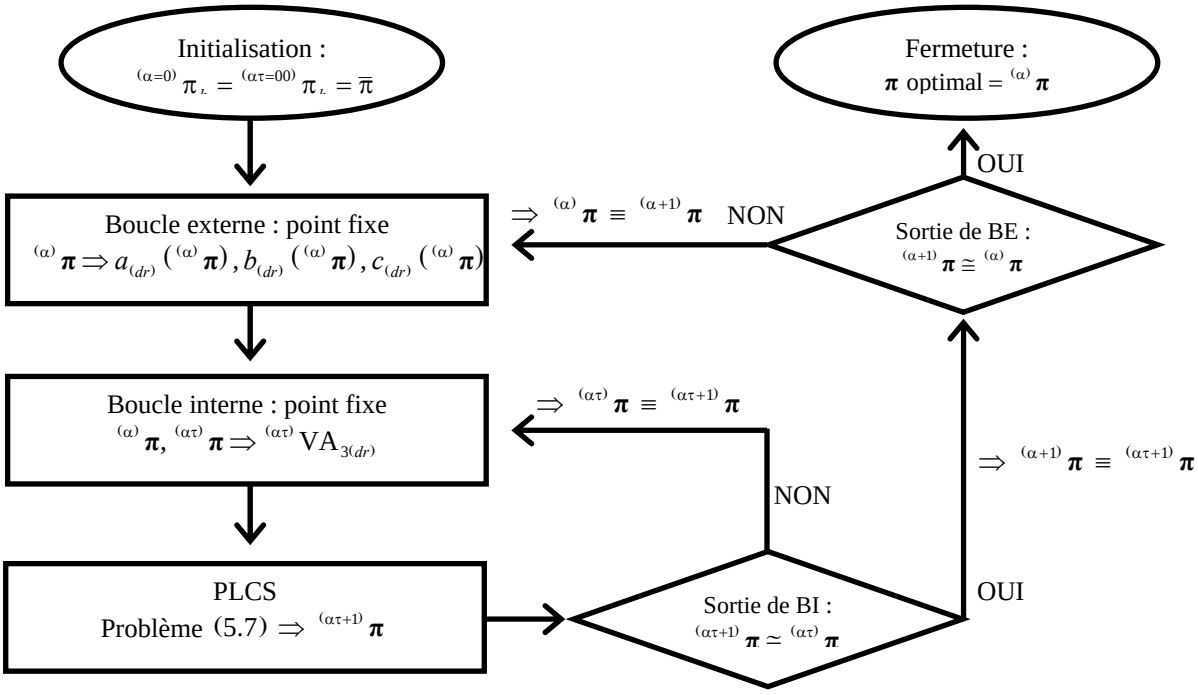

Figure 5.1 Diagramme de cheminement de l'algorithme

Initialisation. À l'itération $\alpha = 0$ de la BE, fixer $^{(\alpha=0)}\boldsymbol{\pi} = \left\{{}^{(\alpha=0)}\pi_k = \bar{\pi}; k = 1,\ldots,N\right\}$ avec $0 < \bar{\pi} \le 1$. Un choix raisonnable est $\bar{\pi} = 0,5$. À l'itération $\tau = 0$ de la boucle interne, fixer $^{(\alpha\tau=0)}\boldsymbol{\pi} = {}^{(\alpha)}\boldsymbol{\pi}$. Fixer le vecteur de dimension N, $\boldsymbol{\varepsilon}$, de faibles valeurs positives.

Boucle externe

- ***Fixation des valeurs pour la boucle interne***. Conformément aux expressions (A1.4), (A1.7) et (A1.8) données à l'annexe A1, les valeurs scalaires réelles suivantes sont calculées

$$a_{(dr)k}\left({}^{(\alpha)}\boldsymbol{\pi}\right) = \boldsymbol{\delta}_k' \left[\mathbf{A}\left({}^{(\alpha)}\boldsymbol{\pi}\right)\right]^{-1} \sum_{j\in U} \boldsymbol{\delta}_j \tilde{y}_{rj} \gamma_{dj} \left(1 - {}^{(\alpha)}\pi_j\right), \tag{5.3}$$

$$b_{(dr)k}\left({}^{(\alpha)}\boldsymbol{\pi}\right) = \boldsymbol{\delta}_k' \left[\mathbf{A}\left({}^{(\alpha)}\boldsymbol{\pi}\right)\right]^{-1} \boldsymbol{\delta}_k \sigma_{rk}^2 \gamma_{dk} \left(1 - {}^{(\alpha)}\pi_k\right), \tag{5.4}$$

$$c_{(dr)k}\left({}^{(\alpha)}\boldsymbol{\pi}\right) = \pi_k^2 \boldsymbol{\delta}_k' \left[\mathbf{A}\left({}^{(\alpha)}\boldsymbol{\pi}\right)\right]^{-1} \left[\sum_{j\in U} \boldsymbol{\delta}_j \boldsymbol{\delta}_j' \sigma_{rj}^2 \gamma_{dj} \left(1 - {}^{(\alpha)}\pi_j\right)^2\right] \left[\mathbf{A}\left({}^{(\alpha)}\boldsymbol{\pi}\right)\right]^{-1} \boldsymbol{\delta}_k. \tag{5.5}$$

- ***Lancement de la boucle interne***. La boucle interne est exécutée jusqu'à la convergence.
- ***Mise à jour ou sortie.*** Si le vecteur ${}^{(\alpha+1)}\boldsymbol{\pi}$ est tel que $\left|{}^{(\alpha+1)}\boldsymbol{\pi} - {}^{(\alpha)}\boldsymbol{\pi}\right| > \boldsymbol{\varepsilon}$, alors la boucle externe est itérée en mettant à jour le vecteur ${}^{(\alpha)}\boldsymbol{\pi}$ avec ${}^{(\alpha+1)}\boldsymbol{\pi}$. Si $\left|{}^{(\alpha+1)}\boldsymbol{\pi} - {}^{(\alpha)}\boldsymbol{\pi}\right| \leq \boldsymbol{\varepsilon}$, alors la bouche externe se ferme et ${}^{(\alpha)}\boldsymbol{\pi}$ représente la solution donnant les valeurs optimales du problème donné par le système (5.1).

Boucle interne

- ***Fixation des valeurs pour le PLCS***. Les valeurs suivantes sont calculées :

$$\begin{aligned} {}^{(\alpha\tau)}\mathrm{VAA}_{3(dr)} &= \sum_{k\in U} \left(1 - {}^{(\alpha\tau)}\pi_k\right) a_{(dr)k}\left({}^{(\alpha)}\boldsymbol{\pi}\right) \left[2\tilde{y}_{rk}\gamma_{dk} - {}^{(\alpha\tau)}\pi_k a_{(dr)k}\left({}^{(\alpha)}\boldsymbol{\pi}\right)\right] \\ &+ \sum_{k\in U} \left(1 - {}^{(\alpha\tau)}\pi_k\right) \left[2b_{(dr)k}\left({}^{(\alpha)}\boldsymbol{\pi}\right) - {}^{(\alpha\tau)}\pi_k c_{(dr)k}\left({}^{(\alpha)}\boldsymbol{\pi}\right)\right]. \end{aligned} \tag{5.6}$$

conformément à l'expression (A1.7) à l'annexe A1.

- ***Résolution du PLCS***. En considérant que les valeurs de ${}^{(a\tau)}\mathrm{VAA}_{3(dr)}$ sont fixes, ${}^{(\alpha\tau+1)}\boldsymbol{\pi}$ s'obtient en résolvant, au moyen d'un algorithme standard pour un PLCS classique, le problème d'optimisation suivant :

$$\begin{cases} \mathrm{Min}\left(\sum_{k\in U} {}^{(\alpha\tau+1)}\pi_k c_k\right) \\ \sum_{k\in U} \dfrac{\left(\tilde{y}_{rk}^2 + \sigma_{rk}^2\right)\gamma_{dk}}{{}^{(\alpha\tau+1)}\pi_k} \leq \dfrac{N-H}{N}\bar{V}_{(dr)} + \sum_{k\in U}\left(\tilde{y}_{rk}^2 + \sigma_{rk}^2\right)\gamma_{dk} + {}^{(\alpha\tau)}\mathrm{VAA}_{3(dr)}. \\ 0 < {}^{(\alpha\tau+1)}\pi_k \leq 1 \qquad (k = 1,\ldots,N) \end{cases} \tag{5.7}$$

- ***Mise à jour ou sortie.*** Si le vecteur ${}^{(\alpha\tau+1)}\boldsymbol{\pi}$ est tel que $\left|{}^{(\alpha\tau+1)}\boldsymbol{\pi} - {}^{(\alpha\tau)}\boldsymbol{\pi}\right| > \boldsymbol{\varepsilon}$, alors la boucle interne est itérée en mettant à jour le vecteur ${}^{(\alpha\tau)}\boldsymbol{\pi}$ avec ${}^{(\alpha\tau+1)}\boldsymbol{\pi}$. Si $\left|{}^{(\alpha\tau+1)}\boldsymbol{\pi} - {}^{(\alpha\tau)}\boldsymbol{\pi}\right| \leq \boldsymbol{\varepsilon}$, alors la boucle interne se ferme et le vecteur mis à jour ${}^{(\alpha+1)}\boldsymbol{\pi}$ pour la boucle externe est donnée par ${}^{(\alpha\tau+1)}\boldsymbol{\pi}$.

Remarque 5.1. Le problème du système (5.7) peut être résolu par l'algorithme proposé dans Falorsi et Righi (2008, section 3.1) qui représente une légère modification de l'algorithme de Chromy (1987), élaboré au départ pour la répartition optimale multivariée sous des plans EASSRS et mis en œuvre dans des outils logiciels standard (voir par exemple le logiciel Mauss-R disponible à l'adresse : http://www3.istat.it/strumenti/metodi/software/campione/mauss_r/). Ou bien, le PLCS peut être traité en se servant de la procédure NLP de SAS comme l'ont proposé Choudhry et coll. (2012).

Remarque 5.2. L'algorithme fait la distinction entre le vecteur ${}^{(\alpha)}\pi_k$ (mis à jour dans la boucle externe) et le vecteur ${}^{(\alpha\tau)}\pi_k$ (mis à jour dans la boucle interne). L'innovation de l'algorithme proposé tient précisément à cette particularité. Si cette distinction entre les probabilités d'inclusion n'est pas faite, c'est-à-dire si ${}^{(\alpha\tau)}\boldsymbol{\pi} = {}^{(\alpha)}\boldsymbol{\pi}$, nous avons observé dans plusieurs expériences que les solutions itérées du PLCS pour chaque boucle externe ne convergent pas vers un point stationnaire.

Remarque 5.3. Après la phase d'optimisation, dans laquelle le vecteur $\boldsymbol{\pi}$ est défini comme étant la solution du problème du système (5.1), une *phase de calage* est exécutée (Falorsi et Righi 2008) afin d'obtenir les probabilités d'inclusion calées, ${}_{\text{cal}}\pi_k$, qui modifient marginalement le vecteur $\boldsymbol{\pi}$ *optimal* afin de satisfaire $\sum_{k\in U} {}_{\text{cal}}\pi_k\boldsymbol{\delta}_k = \mathbf{n}$, où $\mathbf{n}$ est un vecteur de nombres entiers. L'utilisation de l'algorithme d'ajustement proportionnel itératif généralisé (Dykstra et Wollan 1987) permet de s'assurer que toutes les probabilités d'inclusion calées sont comprises dans l'intervalle $(0,1]$.

6 Évaluations empiriques

Plusieurs simulations ont été exécutées sur des ensembles de données réelles et de données simulées pour étudier les propriétés empiriques de la stratégie d'échantillonnage proposée. Ici, nous montrons les résultats obtenus pour un seul exercice portant sur des données réelles se rapportant à la population d'entreprises de 1999 dont le nombre d'employés était compris entre 1 et 99 et qui appartenaient au secteur des Activités informatiques (code à deux chiffres de la *Nomenclature statistique des activités économiques dans la Communauté européenne, Rév. 1*, dont l'acronyme est NACE). Nous avons effectué trois expériences. L'expérience (a) avait pour but de vérifier si la répartition obtenue au moyen de l'algorithme proposé convergeait vers la solution de l'algorithme de Chromy sous le plan EASSRS. L'expérience (b) visait à comparer les tailles d'échantillon du plan EASSRS classique avec celles du plan d'échantillonnage stratifié incomplet (ESI), dans lequel les strates définies par classification croisée étaient des sous-populations non planifiées; cette expérience consistait à étudier le risque de fardeau statistique dû à la sélection répétée lors de différentes éditions de l'enquête. Enfin, l'expérience (c) avait pour objet de mesurer les discordances entre le coefficient de variation (CV) prévu calculé par l'algorithme et le CV empirique obtenu par une simulation Monte Carlo.

Dans les trois expériences, les valeurs de c_k ont été fixées uniformément à 1. La variance anticipée obtenue conformément à l'approximation proposée à la remarque 4.1 a également été calculée.

La taille de la population choisie pour les expériences était de $N = 10\,392$ entreprises. Les domaines d'intérêt définissaient deux partitions de la population cible, à savoir la *région géographique*, avec 20 domaines marginaux (DOM1), et le *groupe d'activités économiques* (code à 3 chiffres de la NACE

avec 6 groupes distincts) *selon la classe de taille* (définie en fonction du nombre d'employés : $1 = 1-4; 2 = 5-9; 3 = 10-19;\ \ 4 = 20-99$), avec 24 domaines marginaux (DOM2). Le nombre global de domaines marginaux était égal à 44, tandis que le nombre de strates formées par classification croisée ou de strates multidimensionnelles ayant une taille de population non nulle était de 360. La valeur modale de la distribution des tailles de population était de 1, et 29,17 % des strates formées par classification croisée ne contenaient au plus que 2 unités. Ce type de strate représente un problème critique dans le contexte des approches d'échantillonnage stratifiées classiques. En effet, pour calculer des estimations de variance sans biais, ces strates doivent être à tirage complet (afin qu'elles ne contribuent pas à la variance des estimations), alors que la règle de répartition exigerait un moins grand nombre d'unités et, en général, un nombre non entier d'unités échantillonnées. Le *coût de la main-d'œuvre* et la *valeur ajoutée* étaient les variables d'intérêt pour lesquelles les données sont fournies par une source administrative pour chaque unité de la population. Habituellement, les deux variables ont une distribution fortement asymétrique.

Pour toutes les études empiriques, les estimations cibles étaient les 88 totaux au niveau du domaine (2 variables fois 44 domaines marginaux). Dans chaque expérience, les probabilités d'inclusion ont été déterminées en fixant la variance $\bar{V}_{(dr)} = (0{,}1 t_{(dr)})^2$ dans (5.1), ce qui équivaut à fixer à 10 % le niveau accepté maximal du CV en pourcentage des estimations au niveau du domaine.

Étude empirique (a). La première expérience tenait compte de la partition DOM1. Ces domaines représentaient à la fois les domaines *planifiés* et les domaines *d'estimation*. Puisque les domaines planifiés définissaient une partition de la population d'intérêt, ils pouvaient également être considérés comme des strates dans les plans d'échantillonnage classiques. Le modèle de travail prédictif était donné par

$$\begin{cases} y_{rk} = \alpha_d + u_{rk} \ \ \forall k \in U_d \ (d = 1, \ldots, 20) \\ E_M(u_{rk}) = 0,\ E_M\left(u_{rk}^2\right) = \sigma_{rd}^2 \ \ \forall k \in U_d;\ \ E_M(u_{rk}, u_{rl}) = 0 \ \ \forall k \neq l \end{cases}, \tag{6.1}$$

où α_d est un effet fixe et les variances dans la superpopulation σ_{rd}^2 étaient estimées au moyen de la variance résiduelle du modèle prédictif dans chaque région. L'algorithme proposé à la section 5 a été exécuté en utilisant trois valeurs initiales distinctes des probabilités d'inclusion $\bar{\pi}$, égales à 0,01, 0,50 et 0,99, respectivement. Les valeurs initiales des probabilités d'inclusion n'avaient aucune incidence sur la solution finale, mais celle-ci était obtenue à la suite d'un nombre différent d'itérations. Nous constatons que le nombre global de boucles internes était de 17 pour $\bar{\pi} = 0{,}01$. La convergence a été obtenue avec 13 boucles internes pour $\bar{\pi} = 0{,}50$; 14 boucles internes ont été nécessaires pour $\bar{\pi} = 0{,}99$. Cependant, après la neuvième itération, les trois tailles d'échantillon étaient relativement similaires (figure 6.1). Dans l'expérience, les tailles d'échantillon globales étaient de 3 105 pour la répartition de Chromy servant de référence et de 3 110 pour la méthode proposée ici. Cependant, les différences entre les deux tailles d'échantillonnage au niveau du domaine étaient des nombres fractionnaires qui étaient toujours inférieurs à 1, et la différence relative absolue la plus importante était inférieure à 0,3 %. Cela met en relief le fait que l'algorithme proposé définit en fait les mêmes tailles d'échantillon de domaine que celles calculées pour la répartition de référence. En ce qui concerne la convergence, les valeurs initiales des probabilités d'inclusion n'ont aucune incidence sur la solution finale, quoique celle-ci soit obtenue moyennant des nombres différents d'itérations.

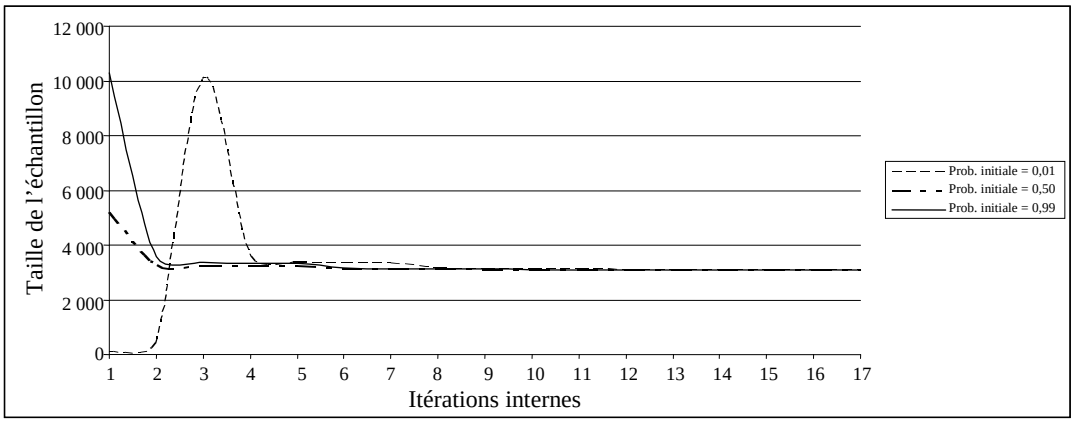

Figure 6.1 Convergence de l'algorithme avec différentes probabilités d'inclusion initiales dans l'étude empirique (a)

Des résultats similaires ont été obtenus quand les domaines d'intérêt étaient définis par la partition DOM2.

Études empiriques (b). Soit U_{d_1} une région particulière $(d_1 = 1,\ldots,20)$ de DOM1, et soit U_{d_2} (avec $d_2 = 1,\ldots,24$) un groupe d'activités économiques particulier selon la classe de taille d'entreprise de la partition DOM2. Nous avons utilisé deux modèles de prédiction, M_1 et M_2. En se référant à la notation des modèles ANOVA, M_1 est le modèle saturé donné par

$$\begin{cases} y_{rk} = \alpha_{d_1} + \lambda_{d_2} + (\alpha\lambda)_{d_1 d_2} + u_{rk} \;\; \forall k \in U_{d_1} \cap U_{d_2} \\ E_M(u_{rk}) = 0,\; E_M\left(u_{rk}^2\right) = \sigma^2_{r(d_1 d_2)} \;\; \forall k \in U_{d_1} \cap U_{d_2};\;\; E_M(u_{rk}, u_{rl}) = 0 \;\; \forall k \neq l \end{cases}, \qquad (6.2)$$

dans lequel α_{d_1} et λ_{d_2} sont les effets principaux, reliés aux domaines U_{d_1} et U_{d_2}, respectivement, et où $(\alpha\lambda)_{d_1 d_2}$ est l'effet d'interaction. Les variances de modèle $\sigma^2_{r(d_1 d_2)}$ ont été estimées par la méthode des moindres carrés ordinaires en calculant les variances des termes résiduels au niveau $U_{d_1} \cap U_{d_2}$. Le modèle M_2 est identique au modèle M_1 sans le facteur d'interaction. Le tableau 6.1 montre la qualité de l'ajustement des deux modèles.

Tableau 6.1
Qualité de l'ajustement des modèles utilisés pour la prédiction

Modèle	Qualité de l'ajustement R^2 %	
	Coût de la main-d'œuvre	Valeur ajoutée
Modèle M_1 (expression 6.2)	68,1	64,1
Modèle M_2 (expression 6.2 sans les interactions)	65,1	61,0

Dans le cas du modèle M_1, nous avons considéré trois répartitions différentes pour l'EASSRS : *i*) aucune contrainte de taille d'échantillon de strate n'est imposée; *ii*) au moins une unité échantillonnée par strate est requise (pour obtenir des estimations ponctuelles sans biais); *iii*) au moins deux unités

échantillonnées par strate sont requises (pour obtenir des estimations de variance sans biais) pour toutes les strates ayant une taille de population de deux entreprises ou plus. Les deux premières répartitions sont plutôt théoriques, puisque dans toutes les enquêtes-entreprises réalisées par l'Institut national de statistique de l'Italie, la sélection d'au moins deux unités par strate est requise. Les résultats de l'expérience sont présentés plus bas au tableau 6.2. Seuls les résultats pour le cas où les probabilités d'inclusion initiales étaient égales à $\bar{\pi} = 0{,}50$ sont examinés ici; des tailles d'échantillon identiques ont été obtenues pour les autres valeurs initiales des probabilités d'inclusion, avec un processus de convergence un peu plus lent. Les trois plans EASSRS comptaient 716,6, 944 et 1 042 unités d'échantillonnage, respectivement. Le plan d'échantillonnage stratifié incomplet (ESI) a donné 936 unités pour le modèle M_1, tandis qu'il a donné 991 unités pour le modèle M_2. Le meilleur résultat donné par le modèle M_1 comparativement au modèle M_2 tenait au fait que son ajustement était meilleur. Enfin, les plans ESI ont aidé à aborder la question du fardeau statistique des entreprises répondantes. En effet, si l'on suppose que les probabilités d'inclusion restent fixes pour les différentes éditions de l'enquête, leurs distributions peuvent être utilisées pour évaluer le fardeau statistique dans les enquêtes répétées. Le tableau 6.2 montre que le nombre d'entreprises sélectionnées avec certitude lors de chaque édition de l'enquête était de 175 pour le troisième plan EASSRS, tandis que 30 et 40 entreprises ont été sélectionnées avec certitude sous le premier et le deuxième plan ESI, respectivement. L'analyse des tailles (mesurées par l'effectif) des entreprises incluses dans l'échantillon avec certitude montre que, dans le cas du troisième plan EASSRS, la taille moyenne était égale à 20,6. Dans certains cas, des entreprises comptant deux employés étaient incluses dans l'échantillon sélectionné avec certitude. Inversement, nous constatons que dans le cas du premier et du deuxième plan ESI, la taille minimale des entreprises était de 17 et 16 employés, respectivement, et que la taille moyenne était supérieure à 40 unités.

Tableau 6.2
Tailles d'échantillon et répartition des entreprises incluses avec certitude dans l'échantillon, pour différents plans d'échantillonnage

Plan d'échantillonnage		Taille de l'échantillon	Entreprises sélectionnées avec certitude		
			Nombre	Nombre d'employés	
				Moyen	Minimum
Stratifié classique avec le modèle M_1	Pas de contrainte de taille d'échantillon de strate	716,6	10	47,0	23,0
	Au moins une unité échantillonnée par strate	944,0	119	24,0	2,0
	Au moins deux unités échantillonnées par strate	1 042,0	175	20,6	2,0
Échantillonnage stratifié incomplet avec le modèle M_1		936,0	30	50,1	17,0
Échantillonnage stratifié incomplet avec le modèle M_2 sans interactions		991,0	40	42,9	16,0

Enfin, pour évaluer la sensibilité de la solution, nous avons répété l'expérience artificiellement et modifié les valeurs de $\tilde{y}_{rk}$ et $\tilde{\sigma}^2_{rk}$ dans le problème d'optimisation (5.1). En particulier, nous avons augmenté les valeurs prédites de $\tilde{\sigma}^2_{rk}$ de 20 % et 120 % respectivement, et diminué de 20 % les valeurs de $\tilde{y}_{rk}$ prédites par le modèle M_1. Comme prévu, les tailles d'échantillon ont augmenté, mais le plan EASSRS avec au moins une unité échantillonnée par strate et le premier plan ESI ont défini approximativement les mêmes tailles d'échantillon (tableau 6.3).

Tableau 6.3
Tailles d'échantillon avec valeurs prévues modifiées des prédictions du modèle (4.1)

Plan d'échantillonnage		Taille de l'échantillon		
		$\tilde{\sigma}^2_{rk}$ augmenté de 20 %	$\tilde{\sigma}^2_{rk}$ augmenté de 120 %	$\tilde{y}_{rk}$ diminué de 20 %
EASSRS avec modèle M_1	Aucune contrainte de taille d'échantillon de strate	821,0	1 269,0	993,8
	Au moins une unité échantillonnée par strate	1 035,0	1 472,0	1 206,0
	Au moins deux unités échantillonnées par strate	1 125,0	1 536,0	1 283,0
Plan ESI avec modèle M_1		1 039,7	1 460,9	1 207,5

Étude empirique (c). Nous avons utilisé le modèle de prédiction linéaire hétéroscédastique M_3 :

$$\begin{cases} y_{rk} = \alpha_r + \varphi_r x_k + u_{rk} \\ E_M(u_{rk}) = 0, \;\; E_M(u_{rk}^2) = \sigma_{rk}^2 = \sigma_r^2 x_k \;\; \forall k \in U; \; E_M(\varepsilon_{rk}, \varepsilon_{rl}) = 0 \;\; \forall k \neq l \end{cases}, \qquad (6.3)$$

où x_k est le *nombre d'employés* dans la k^{e} entreprise, et α_r et φ_r sont les paramètres de régression. Notons que le nombre d'employés est disponible dans la base de sondage en Italie.

Nous avons calculé deux estimations différentes de la variance du modèle :
a) $\tilde{\sigma}^2_{rk} = 1/N_{(X=x_k)} \sum_{k \in U_{(X=x_k)}} (y_{rk} - \mathrm{A}_r - \mathrm{F}_r x_k)^2$ et b) $\tilde{\sigma}^2_{rk} = \tilde{\sigma}^2_r x_k$, dans lesquelles $\tilde{\sigma}^2_r = 1/(N-2) \sum_{k \in U} [(y_{rk} - \mathrm{A}_r - \mathrm{F}_r x_k)/x_k]^2$, où $U_{(X=x)}$ est la population d'entreprises, de taille $N_{(X=x)}$, pour laquelle la variable X prend la valeur x; A_r et F_r sont les estimations de α_r et φ_r, respectivement, par les moindres carrés pondérés pour la population dénombrée complète. La somme des variances de modèle obtenue par la méthode (a) était plus faible que celle obtenue par la méthode (b). Cela a été reflété par les tailles d'échantillon calculées. La première répartition définit une taille d'échantillon global de 927 unités, tandis que la deuxième répartition définit une taille d'échantillon de 951. Nous avons tiré successivement 1 000 échantillons pour chacune des répartitions et avons calculé les ratios $\mathrm{RCV}(\hat{t}_{(dr)}) = \mathrm{CVP}(\hat{t}_{(dr)})/\mathrm{CVS}(\hat{t}_{(dr)})$, avec $\mathrm{CVP}(\hat{t}_{(dr)}) = \left[\sqrt{\mathrm{VAA}(\hat{t}_{(dr)})}/\hat{t}_{(dr)}\right]100$ représentant le CV prévu (%) et

$$\mathrm{CVS}(\hat{t}_{(dr)}) = 100 \sqrt{(1/I)\left[\sum_{i=1}^{I} \hat{t}^i_{(dr)} - (1/I)\sum_{i=1}^{I} \hat{t}^i_{(dr)}\right]^2} \Big/ (1/I)\sum_{i=1}^{I} \hat{t}^i_{(dr)}$$

représentant le CV simulé (ou empirique), obtenu comme résultat de la simulation, en désignant par $\hat{t}^i_{(dr)}$ l'estimation HT dans la i^{e} itération et $I = 1\,000$. Par souci de concision, seuls les principaux résultats de la répartition (b) sont présentés à la figure 6.2 pour DOM1 et DOM2, respectivement, pour les deux variables d'intérêt. En examinant la figure de gauche, nous remarquons que la simulation produit généralement un CV plus petit que le CV prévu, ce qui donne un ratio RCV plus grand que 1 pour les deux variables. Une exception a lieu, pour la valeur ajoutée dans un domaine de DOM1.

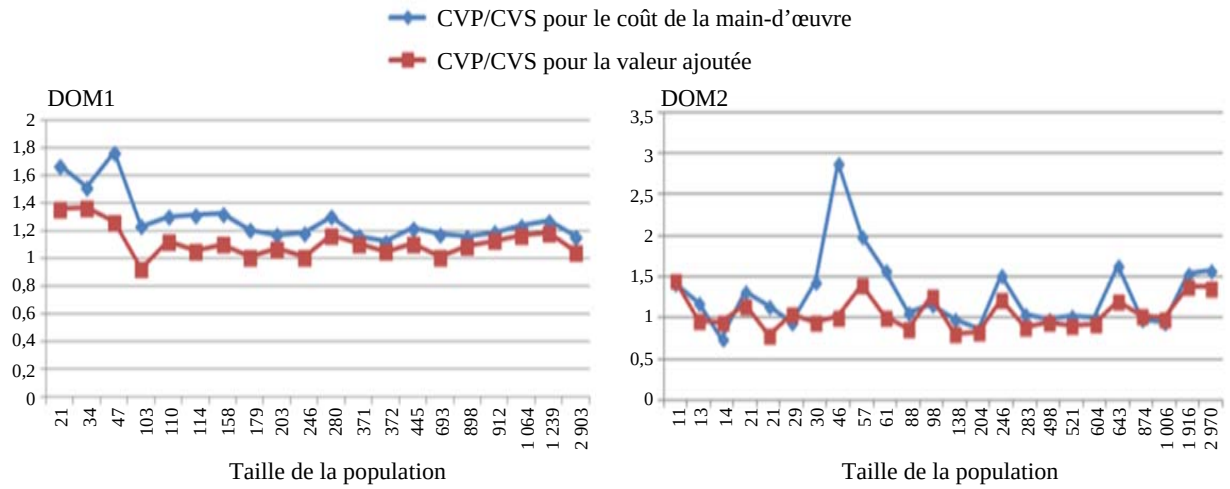

Figure 6.2 RCV selon la taille de la population pour le coût de la main-d'œuvre et la valeur ajoutée

La valeur de RCV inférieure à 1 peut être expliquée par l'augmentation des tailles d'échantillon de domaine en raison de l'étape de calage. Nous constatons qu'en général, ces divergences sont observées dans des domaines dont la taille de population est petite; donc, l'étape de calage peut avoir un effet non négligeable. La figure de droite présente des données empiriques plus articulées et conflictuelles. Premièrement, nous constatons que les RCV sont souvent plus grands que 1 ou très proches de 1. Néanmoins, dans trois domaines, la variable de valeur ajoutée possède un CV simulé égal à 11,5 %, 12,0 % et 12,3 %, respectivement. Dans ces cas rares, et certains autres (coût de la main-d'œuvre dans deux domaines), les divergences sont en harmonie avec les constatations de Deville et Tillé (2005) quant aux propriétés empiriques de l'approximation de la variance pour l'échantillonnage équilibré.

7 Conclusion

L'article décrit une nouvelle approche en vue de déterminer les probabilités d'inclusion optimales dans divers contextes d'enquête caractérisés par la nécessité de diffuser des estimations d'enquête d'une précision préétablie, pour de multiples variables et domaines d'intérêt.

La principale contribution de l'article a trait au calcul pratique de ces probabilités au moyen d'un nouvel algorithme, qui convient pour un plan d'échantillonnage multidimensionnel général dans lequel l'échantillonnage stratifié classique représente un cas particulier. L'approche proposée, l'algorithme et le calcul final sont orientés domaine et variable.

Dans notre cadre, les variables indicatrices d'appartenance à un domaine sont supposées connues, tandis que les variables d'intérêt sont inconnues. La procédure est alors appliquée aux valeurs prédites des caractéristiques d'intérêt au moyen d'un modèle de superpopulation, et l'algorithme permet de tenir compte de l'incertitude du modèle; cela reflète le fait que les valeurs des variables d'intérêt sont inconnues. En utilisant la variance anticipée comme mesure de la précision de l'estimateur, cette approche

permet de contourner les limites des algorithmes standard utilisés pour la répartition des échantillons, dans lesquels les variables d'intérêt dictant la solution sont supposées connues.

L'algorithme proposé exploite une procédure standard, mais présente certaines innovations en matière de calcul qui pourraient être utiles pour faire face à la complexité qui découle du fait que les variances anticipées sont des fonctions implicites des probabilités d'inclusion. L'algorithme a été testé sur des données simulées et des données d'enquête réelles afin d'évaluer sa performance et ses propriétés. Les résultats d'un petit ensemble d'expériences sont présentés ici. Ils confirment une amélioration, en ce qui concerne l'efficacité, de la stratégie d'échantillonnage. Une généralisation naturelle du cas examiné ici peut être élaborée en considérant que les indicateurs de domaine et d'autres variables indépendantes quantitatives sont connus à l'étape de l'élaboration du plan d'échantillonnage. Nous notons que la variance anticipée en ne tenant compte que des indicateurs de domaine est plus grande que la variance anticipée de ce cas plus général. Donc, notre solution représente une borne supérieure (et d'une certaine robustesse) de la solution à la phase de l'élaboration du plan. En outre, la solution algorithmique peut être adaptée facilement à cette situation plus générale.

Remerciements

La présente étude a été financée par le partenariat de la Stratégie mondiale pour l'amélioration des statistiques agricoles et rurales : http://www.fao.org/economic/ess/ess-capacity/strategie-mondiale/fr/.

Annexe A1

VA de l'estimateur HT

Considérons le résidu $\eta_{(dr)k}$ tel qu'il est exprimé par l'équation (3.5), et remplaçons le terme y_{rk} par $\tilde{y}_{rk} + u_{rk}$, ce qui nous donne

$$\eta_{(dr)k} = (\tilde{y}_{rk} + u_{rk})\gamma_{dk} - \pi_k \boldsymbol{\delta}'_k [\mathbf{A}(\boldsymbol{\pi})]^{-1} \sum_{j\in U} \pi_j \boldsymbol{\delta}_j (\tilde{y}_{rj} + u_{rj})\gamma_{dj} (1/\pi_j - 1). \quad \text{(A1.1)}$$

Les moindres prédictions pondérées de $\tilde{y}_{rk}\gamma_{dk}$ et $u_{rk}\gamma_{dk}$, avec les prédicteurs $\pi_k \boldsymbol{\delta}_k$ et les pondérations $1/\pi_k - 1$, sont

$$\hat{\tilde{y}}_{(dr)k} = \pi_k a_{(dr)k} \quad \text{(A1.2)}$$

et

$$\hat{u}_{(dr)k} = \pi_k \boldsymbol{\delta}'_k [\mathbf{A}(\boldsymbol{\pi})]^{-1} \sum_{j\in U} \pi_j \boldsymbol{\delta}_j u_{rj}\gamma_{dj} (1/\pi_j - 1), \quad \text{(A1.3)}$$

avec

$$a_{(dr)k}(\boldsymbol{\pi}) = \boldsymbol{\delta}'_k [\mathbf{A}(\boldsymbol{\pi})]^{-1} \sum_{j\in U} \pi_j \boldsymbol{\delta}_j \tilde{y}_{rj}\gamma_{dk} (1/\pi_j - 1). \quad \text{(A1.4)}$$

En utilisant les formules (A1.2) et (A1.3), l'expression (A1.1) peut être reformulée sous la forme $\eta_{(dr)k} = (\tilde{y}_{rk} + u_{rk})\gamma_{dk} - [\hat{\tilde{y}}_{(dr)k} + \hat{u}_{(dr)k}]$. Par conséquent, l'espérance sous le modèle de $\eta^2_{(dr)k}$ est

$$E_M\left(\eta_{(dr)k}^2\right) = \left(\tilde{y}_{rk}\gamma_{dk} - \hat{\tilde{y}}_{(dr)k}\right)^2 + E_M\left[\left(u_{rk}\gamma_{dk} - \hat{u}_{(dr)k}\right)^2\right] + \text{termes de moyenne nulle,} \quad \text{(A1.5)}$$

car $E_M\left(u_{rk}\right) = 0$. En outre,

$$E_M\left[\left(u_{rk}\gamma_{dk} - \hat{u}_{(dr)k}\right)^2\right] = \sigma_{rk}^2\gamma_{dk} + E_M\left(\hat{u}_{(dr)k}\right)^2 - 2E_M\left(u_{rk}\gamma_{dk}, \hat{u}_{(dr)k}\right), \quad \text{(A1.6)}$$

où $E_M\left(u_{rk}\gamma_{dk}\hat{u}_{(dr)k}\right) = \pi_k b_{(dr)k}\left(\boldsymbol{\pi}\right)$ et $E_M\left(\hat{u}_{(dr)k}\right)^2 = \pi_k^2 c_{(dr)k}\left(\boldsymbol{\pi}\right)$, avec

$$b_{(dr)k}\left(\boldsymbol{\pi}\right) = \boldsymbol{\delta}_k'\left[\mathbf{A}\left(\boldsymbol{\pi}\right)\right]^{-1}\boldsymbol{\delta}_k\sigma_{rk}^2\gamma_{dk}\left(1-\pi_k\right) \quad \text{(A1.7)}$$

et

$$c_{(dr)k}\left(\boldsymbol{\pi}\right) = \boldsymbol{\delta}_k'\left[\mathbf{A}\left(\boldsymbol{\pi}\right)\right]^{-1}\left[\sum_{j\in U}\boldsymbol{\delta}_j\boldsymbol{\delta}_j'\sigma_{rj}^2\gamma_{dj}\left(1-\pi_j\right)^2\right]\left[\mathbf{A}\left(\boldsymbol{\pi}\right)\right]^{-1}\boldsymbol{\delta}_k. \quad \text{(A1.8)}$$

L'expression (4.5) est obtenue facilement en insérant les expressions provenant de (A1.2) à (A1.8) dans l'équation (4.3).

Annexe A2

Convergence de l'algorithme

Le problème d'optimisation (5.1) est résolu par deux *itérations du point fixe* emboîtées. Étant donné un vecteur $\mathbf{x}$ de dimension q inconnu, l'itération du point fixe choisit une valeur supposée initiale ${}^{(0)}\mathbf{x}$. Puis, l'algorithme calcule des itérés subséquents selon ${}^{(\tau+1)}\mathbf{x} = \mathbf{g}\left({}^{(\tau)}\mathbf{x}\right)$, avec $\tau = 1, 2, \ldots$, où $\mathbf{g}(\cdot)$ est un système de q équations de mise à jour. La fonction multivariée $\mathbf{g}$ possède un point fixe dans un domaine $Q \subseteq \Re^q$ si $\mathbf{g}$ applique Q dans Q. Soit $J_{\mathbf{g}}\left(\mathbf{x}\right)$ la matrice jacobéenne de la dérivée partielle première de $\mathbf{g}$ évaluée à $\mathbf{x}$. S'il existe une constante $\rho < 1$ telle que, dans une norme matricielle naturelle, $\left\|J_{\mathbf{g}}\left(\mathbf{x}\right)\right\| \leq \rho, \mathbf{x} \in Q$, $\mathbf{g}$ possède un point fixe unique $\mathbf{x}^* \in Q$, et l'itération du point fixe est garantie de converger vers $\mathbf{x}^*$ pour toute valeur supposée initiale choisie dans Q. En ce qui concerne l'algorithme proposé, la convergence de la boucle interne (BI) et de la boucle externe (BE) est obtenue quand les termes ${}^{(\alpha\tau)}\mathrm{VAA}_{3(dr)}$ convergent vers le point fixe. Cela signifie que les vecteurs ${}^{(\alpha)}\boldsymbol{\pi}$ et ${}^{(\alpha\tau)}\boldsymbol{\pi}$ ne changent pas dans les itérations de la BE et de la BI. Dans la démonstration qui suit, nous considérons la méthode proposée par Chromy (1987) pour résoudre le PLCS du système (5.7), et nous formulons certaines hypothèses raisonnables, à savoir : 1) $\hat{u}_{(dr)k} \cong 0$; 2) $[N/(N-H)] \cong 1$; 3) $\hat{\tilde{y}}_{rk} \cong \tilde{y}_{rk}$; 4) ${}^{(\alpha)}\pi_k \cong {}^{(\alpha\tau)}\Delta\,{}^{(\alpha\tau)}\pi_k$ avec $0 < {}^{(\alpha\tau)}\Delta \leq 1$; 5) $c_k \cong \bar{c}$. L'hypothèse (1) correspond à l'approximation à la hausse de la variance anticipée, donnée à la remarque 4.1, et implique que $b_{(dr)k}\left({}^{(\alpha)}\boldsymbol{\pi}\right) = c_{(dr)k}\left({}^{(\alpha)}\boldsymbol{\pi}\right) = 0$. L'hypothèse (3) implique que $a_{(dr)k}\left({}^{(\alpha)}\boldsymbol{\pi}\right)\tilde{y}_{rk}\gamma_{dk} \cong \tilde{y}_{rk}^2\gamma_{dk}\,/\,{}^{(\alpha)}\pi_k$. L'hypothèse (4) énonce que la structure des probabilités d'inclusion demeure à peu près constante dans les différentes itérations de la BI. L'hypothèse devient raisonnable compte tenu du fait que l'équation de mise à jour A2.2 qui suit (d'une probabilité d'inclusion donnée) est essentiellement déterminée par le seuil de variance qui requiert la taille d'échantillon la plus grande. Il est plausible d'émettre l'hypothèse que ce seuil demeure plus ou moins le même dans les itérations de la BI subséquentes d'une BE donnée.

Preuve de la convergence de la boucle interne. En reformulant l'expression (4.6) conformément aux hypothèses (1) à (4),

$$ {}^{(\alpha\tau+1)}\mathrm{VAA}_{3(dr)} = \sum_{k\in U}\left[\left(\frac{1}{{}^{(\alpha\tau+1)}\pi_k}-1\right)\left(2\frac{\tilde{y}_{rk}^2\gamma_{dk}}{{}^{(\alpha\tau+1)}\Delta}-\frac{\tilde{y}_{rk}^2\gamma_{dk}}{{}^{(\alpha\tau+1)}\Delta^2}\right)\right]. \tag{A2.1}$$

En considérant que, dans le problème (5.7), les valeurs de ${}^{(a\tau)}\mathrm{VAA}_{3(dr)}$ sont fixes, chaque valeur du vecteur ${}^{(\alpha\tau+1)}\boldsymbol{\pi}$ s'obtient comme une solution du PLCS avec l'algorithme de Chromy. Désignons par $\alpha\tau v^*$ l'itération de l'algorithme de Chromy durant laquelle il converge, où ${}^{(\alpha\tau v^*+1)}\boldsymbol{\pi} \cong {}^{(\alpha\tau v^*)}\boldsymbol{\pi}$. Alors, la BI met à jour la probabilité générique conformément à l'expression

$$ {}^{(\alpha\tau+1)}\pi_k = \left[\sum_{(dr)} {}^{(\alpha\tau v^*+1)}\phi_{(dr)}\frac{\left(\tilde{y}_{rk}^2+\sigma_{rk}^2\right)\gamma_{dk}}{\bar{c}}\right]^{1/2}, \tag{A2.2}$$

où le deuxième terme du membre de droite représente la formule de mise à jour de l'algorithme de Chromy, et $\sum_{(dr)}$ représente $\sum_{d=1}^{D}\sum_{r=1}^{R}$, et ${}^{(\alpha\tau v^*+1)}\phi_{(dr)}$ est le multiplicateur de Lagrange généralisé, où

$$\begin{aligned} {}^{(\alpha\tau v^*+1)}\phi_{(dr)} &= {}^{(\alpha\tau v^*)}\phi_{(dr)}\left[\frac{{}^{(\alpha\tau v^*)}V_{(dr)}}{\ddot{V}_{(dr)}+{}^{(\alpha\tau)}\mathrm{VAA}_{3(dr)}}\right]^2, \\ {}^{(\alpha\tau v^*)}V_{(dr)} &= \sum_{k\in U}\frac{\left(\tilde{y}_{rk}^2+\sigma_{rk}^2\right)\gamma_{dk}}{{}^{(\alpha\tau v^*)}\pi_k} \end{aligned} \tag{A2.3}$$

et

$$\ddot{V}_{(dr)} = \bar{V}_{(dr)} + \sum_{k\in U}\left(\tilde{y}_{rk}^2+\sigma_{rk}^2\right)\gamma_{dk}.$$

La théorie de Kuhn-Tucker énonce que ${}^{(\alpha\tau v^*)}\phi_{(dr)}\left[{}^{(\alpha\tau v^*)}V_{(dr)} - \left(\ddot{V}_{(dr)}+{}^{(\alpha\tau)}\mathrm{AV}_{3(dr)}\right)\right] = 0$; par conséquent, ${}^{(\alpha\tau v^*+1)}\phi_{(dr)} = {}^{(\alpha\tau v^*)}\phi_{(dr)}$ et ${}^{(\alpha\tau v^*+1)}\phi_{(dr)} > 0$ si et seulement si ${}^{(\alpha\tau v^*)}V_{(dr)}/\left(\ddot{V}_{(dr)}+{}^{(\alpha\tau)}\mathrm{AV}_{3(dr)}\right) = 1$. Chromy affirme que peu de ${}^{(\alpha\tau v^*)}\phi_{(dr)}$ (pour $r = 1,\ldots,R; d = 1,\ldots,D$) sont plus grands que zéro, et que dans la plupart des cas, une seule valeur est strictement positive. En notant ${}^{(\alpha\tau)}\mathbf{VAA}_3 = \left({}^{(\alpha\tau)}\mathrm{VAA}_{3(11)},\ldots,{}^{(\alpha\tau)}\mathrm{VAA}_{3(1R)},\ldots,{}^{(\alpha\tau)}\mathrm{VAA}_{3(DR)}\right)'$, nous définissons ${}^{(\alpha\tau+1)}\mathbf{VAA}_3 = \mathbf{g}\left({}^{(\alpha\tau)}\mathbf{VAA}_3\right)$ comme étant le système de $D\times R$ équations de mise à jour, où l'équation $(\overline{dr})$ générique du système

$$\begin{aligned} g_{(\overline{dr})}\left({}^{(\alpha\tau)}\mathbf{VAA_3}\right) &\cong \sum_{k\in U}\left(2\frac{\tilde{y}_{rk}^2\gamma_{dk}}{{}^{(\alpha\tau+1)}\Delta}-\frac{\tilde{y}_{rk}^2\gamma_{dk}}{{}^{(\alpha\tau+1)}\Delta^2}\right) \\ &\times \left\{\left[\sum_{(dr)} {}^{(\alpha\tau v^*)}\phi_{(dr)}\left[\frac{{}^{(\alpha\tau v^*)}V_{(dr)}}{\ddot{V}_{(dr)}+{}^{(\alpha\tau)}\mathrm{VAA}_{3(dr)}}\right]^2\frac{\left(\tilde{y}_{rk}^2+\sigma_{rk}^2\right)\gamma_{dk}}{\bar{c}}\right]^{-1/2}-1\right\}, \end{aligned} \tag{A2.4}$$

s'obtient en insérant l'expression (A2.2) dans (A2.1). Si l'on obtient la convergence, alors dans la dernière itération, ${}^{(\alpha\tau+1)}\mathbf{VAA}_3 \cong {}^{(\alpha\tau)}\mathbf{VAA}_3$. La fonction de l'équation (A2.4) est continue et dérivable. En outre,

elle s'applique sur l'intervalle des valeurs possibles de $\mathrm{VAA}_{3(dr)}$. Alors, la BI converge si la condition qui suit est satisfaite :

$$\left\| J_{\mathbf{g}} \left(\mathbf{VAA}_3 \right) \right\| \leq 1. \tag{A2.5}$$

La matrice jacobienne est semi-définie positive, et un résultat bien connu énonce que $\text{trace}\left(J_{\mathbf{g}} J_{\mathbf{g}}'\right) \leq \text{trace}\left(J_{\mathbf{g}}\right)^2$. En considérant la norme de Frobenius $\left\|J_{\mathbf{g}}\right\|_F = \sqrt{\text{trace}\left(J_{\mathbf{g}} J_{\mathbf{g}}'\right)}$, elle devient $\left\|J_{\mathbf{g}}\right\|_F \leq \text{trace}\left(J_{\mathbf{g}}\right)$. Donc, nous pouvons tenir compte de la trace de la matrice jacobienne pour vérifier la condition (A2.5). Soit $g'_{(\overline{dr})} = \partial g_{(\overline{dr})} \left({}^{(\alpha\tau-1)}\mathbf{VAA}_{3(dr)} \big/ \partial\, {}^{(\alpha\tau-1)}\mathrm{VAA}_{3(\overline{dr})} \right)$ l'élément $(\overline{dr})$ de la diagonale de $J_{\mathbf{g}}\left(\mathbf{VAA}_3\right)$. En utilisant la condition de Kuhn-Tucker ${}^{(\alpha\tau v^*)}V_{(dr)} \big/ \left(\ddot{V}_{(dr)} + {}^{(\alpha\tau)}\mathrm{AV}_{3(dr)} \right) = 1$,

$$\begin{aligned} g'_{(\overline{dr})} &= \sum\nolimits_{k\in U} \left(2\frac{\tilde{y}_{\overline{rk}}^2 \gamma_{\overline{dk}}}{{}^{(\alpha\tau+1)}\Delta} - \frac{\tilde{y}_{\overline{rk}}^2 \gamma_{\overline{dk}}}{{}^{(\alpha\tau+1)}\Delta^2} \right) \left[\sum\nolimits_{(dr)} {}^{(\alpha\tau v^*)}\phi_{(dr)} \frac{\left(\tilde{y}_{rk}^2 + \sigma_{rk}^2\right)\gamma_{dk}}{\overline{c}} \right]^{-3/2} \\ &\times {}^{(\alpha\tau v^*)}\phi_{(\overline{dr})} \frac{1}{{}^{(\alpha\tau v^*)}V_{(\overline{dr})}} \frac{\left(\tilde{y}_{\overline{rk}}^2 + \sigma_{\overline{rk}}^2\right)\gamma_{\overline{dk}}}{\overline{c}}. \end{aligned}$$

Puisque dans de nombreux cas, ${}^{(\alpha\tau v^*)}\phi_{(\overline{dr})} = 0$ (Chromy 1987), l'élément $g'_{(\overline{dr})}$ respectif est nul. Quand ${}^{(\alpha\tau v^*)}\phi_{(\overline{dr})} > 0$, alors

$$g'_{(\overline{dr})} \leq \sum\nolimits_{k\in U} \left(2\frac{\tilde{y}_{\overline{rk}}^2 \gamma_{\overline{dk}}}{{}^{(\alpha\tau+1)}\Delta} - \frac{\tilde{y}_{\overline{rk}}^2 \gamma_{\overline{dk}}}{{}^{(\alpha\tau+1)}\Delta^2} \right) \left[{}^{(\alpha\tau v^*)}\phi_{(\overline{dr})} \frac{\left(\tilde{y}_{\overline{rk}}^2 + \sigma_{\overline{rk}}^2\right)\gamma_{\overline{dk}}}{\overline{c}} \right]^{-3/2} \times {}^{(\alpha\tau v^*)}\phi_{(\overline{dr})} \frac{1}{{}^{(\alpha\tau v^*)}V_{(\overline{dr})}} \frac{\left(\tilde{y}_{\overline{rk}}^2 + \sigma_{\overline{rk}}^2\right)\gamma_{\overline{dk}}}{\overline{c}}$$

$$= \sum\nolimits_{k\in U} \left(2\frac{\tilde{y}_{\overline{rk}}^2 \gamma_{\overline{dk}}}{{}^{(\alpha\tau+1)}\Delta} - \frac{\tilde{y}_{\overline{rk}}^2 \gamma_{\overline{dk}}}{{}^{(\alpha\tau+1)}\Delta^2} \right) \frac{1}{\sqrt{{}^{(\alpha\tau v^*)}\phi_{(\overline{dr})} \frac{\left(\tilde{y}_{\overline{rk}}^2 + \sigma_{\overline{rk}}^2\right)\gamma_{\overline{dk}}}{\overline{c}}\, {}^{(\alpha\tau v^*)}V_{(\overline{dr})}}}$$

$$\leq \sum\nolimits_{k\in U} \frac{\frac{\tilde{y}_{\overline{rk}}\gamma_{\overline{dk}}}{{}^{(\alpha\tau+1)}\Delta}\left(2 - \frac{1}{{}^{(\alpha\tau+1)}\Delta}\right)}{\sqrt{\overline{c}\, {}^{(\alpha\tau v^*)}\phi_{(\overline{dr})}\gamma_{\overline{dk}}\, {}^{(\alpha\tau v^*)}V_{(\overline{dr})}}} << 1.$$

Par conséquent, la $\text{trace}\left(J_{\mathbf{g}}\right)$ doit être inférieure à 1.

Preuve de la convergence de la boucle externe. Soit ${}^{(\alpha\tau+1)}\boldsymbol{\pi}$ la solution du problème de point fixe de la BI; alors, la BE met à jour le vecteur ${}^{(\alpha)}\boldsymbol{\pi}$ avec ${}^{(\alpha+1)}\boldsymbol{\pi} = {}^{(\alpha\tau+1)}\boldsymbol{\pi}$. Sous les conditions (1), (2) et (3),

$${}^{(\alpha+1)}\mathrm{VAA}_{3(dr)} = \sum\nolimits_{k\in U} \left(\frac{1}{{}^{(\alpha\tau+1)}\pi_k} - 1 \right) \tilde{y}_{rk}^2 \gamma_{dk}. \tag{A2.6}$$

En insérant l'expression (A2.2) dans la formule (A2.6) quand la BI converge, le système de $D \times R$ équations de mise à jour de ${}^{(\alpha+1)}\mathbf{VAA}_3$ est donné par ${}^{(\alpha+1)}\mathbf{VAA}_3 = \mathbf{j}\left({}^{(\alpha\tau)}\mathbf{VAA}_3\right)$, où l'équation générique de $\mathbf{j}$ est

$$
\begin{aligned}
{}^{(\alpha+1)}\mathrm{VAA}_{3(dr)} &= j_{(\bar{d}\bar{r})}\left({}^{(\alpha\tau)}\mathbf{VAA}_3\right) \\
&= \sum\nolimits_{k\in U} \tilde{y}_{\bar{r}k}^2 \gamma_{\bar{d}k} \left(\left[\sum\nolimits_{(dr)} {}^{(\alpha\tau v^*)}\phi_{(dr)} \left[\frac{{}^{(\alpha\tau v^*)}V_{(dr)}}{\ddot{V}_{(\bar{d}\bar{r})} + {}^{(\alpha\tau)}\mathrm{VAA}_{3(\bar{d}\bar{r})}} \right]^2 \frac{\left(\tilde{y}_{rk}^2 + \sigma_{rk}^2\right)\gamma_{dk}}{\bar{c}} \right]^{-1/2} - 1 \right).
\end{aligned}
\tag{A2.7}
$$

En notant que ${}^{(\alpha)}\mathbf{VAA}_3 = {}^{(\alpha\tau=0)}\mathbf{VAA}_3$, le système **j** peut être exprimé sous une forme récursive

$$
{}^{(\alpha+1)}\mathbf{VAA}_3 \cong \mathbf{j}\left(\mathbf{g}\left({}^{(\alpha\tau-1)}\mathbf{VAA}_3\right)\right) = \mathbf{j}\left(\mathbf{g}\left(\mathbf{g}\left(.....\mathbf{g}\left({}^{(\alpha\tau=0)}\mathbf{VAA}_3\right)\right)\right)\right) = \mathbf{f}\left({}^{(\alpha)}\mathbf{VAA}_3\right),
$$

avec $\mathbf{f}(\cdot) = \mathbf{j}(\mathbf{g}(\mathbf{g}(.....\mathbf{g}(\cdot))))$ en tant que système de $D \times R$ équations de mise à jour de ${}^{(\alpha+1)}\mathrm{VAA}_3$, par rapport aux valeurs antérieures de la BE, ${}^{(\alpha)}\mathbf{VAA}_3$. Pour démontrer la convergence de la BE, il est nécessaire de démontrer que la norme jacobienne $\|J_{\mathbf{f}}(\mathbf{VAA}_3)\|$ est inférieure à 1. En utilisant les résultats classiques de l'algèbre matricielle,

$$
\|J_{\mathbf{f}}(\mathbf{VAA}_3)\| \le \left\|J_{\mathbf{j}}\left({}^{(\alpha\tau)}\mathbf{VAA}_3\right)\right\| \times \left\|J_{\mathbf{g}}\left({}^{(\alpha\tau-1)}\mathbf{VAA}_3\right)\right\| \times \ldots \times \left\|J_{\mathbf{g}}\left({}^{(\alpha\tau=0)}\mathbf{VAA}_3\right)\right\|,
$$

où la norme générique $\|J_{\mathbf{g}}(\cdot)\|$ est inférieure à 1 (voir la preuve de convergence de la BI). Soit $j'_{(\bar{d}\bar{r})}$ l'élément $(\bar{d}\bar{r})$ de la diagonale de $J_{\mathbf{j}}\left({}^{(\alpha\tau)}\mathbf{VAA}_3\right)$. Il est donné par

$$
\begin{aligned}
j'_{(\bar{d}\bar{r})} &= \sum\nolimits_{k\in U} \tilde{y}_{\bar{r}k}^2 \gamma_{\bar{d}k} \left[\sum\nolimits_{(dr)} {}^{(\alpha\tau v^*)}\phi_{(dr)} \frac{\left(\tilde{y}_{rk}^2 + \sigma_{rk}^2\right)\gamma_{dk}}{\bar{c}} \right]^{-3/2} \\
&\times {}^{(\alpha\tau v^*)}\phi_{(\bar{d}\bar{r})} \frac{1}{{}^{(\alpha\tau v^*)}V_{(\bar{d}\bar{r})}} \frac{\left(\tilde{y}_{\bar{r}k}^2 + \sigma_{\bar{r}k}^2\right)\gamma_{\bar{d}k}}{\bar{c}}.
\end{aligned}
\tag{A2.8}
$$

Par conséquent, nous avons

$$
\begin{aligned}
j'_{(\bar{d}\bar{r})} &\le \sum\nolimits_{k\in U} \tilde{y}_{\bar{r}k}^2 \gamma_{\bar{d}k} \left[{}^{(\alpha\tau v^*)}\phi_{(\bar{d}\bar{r})} \frac{\left(\tilde{y}_{\bar{r}k}^2 + \sigma_{\bar{r}k}^2\right)\gamma_{\bar{d}k}}{\bar{c}} \right]^{-3/2} {}^{(\alpha\tau v^*)}\phi_{(\bar{d}\bar{r})} \frac{1}{{}^{(\alpha\tau v^*)}V_{(\bar{d}\bar{r})}} \frac{\left(\tilde{y}_{\bar{r}k}^2 + \sigma_{\bar{r}k}^2\right)\gamma_{\bar{d}k}}{\bar{c}} \\
&= \frac{1}{{}^{(\alpha\tau v^*)}V_{(\bar{d}\bar{r})}} \sum\nolimits_{k\in U} \tilde{y}_{\bar{r}k}^2 \gamma_{\bar{d}k} \left[{}^{(\alpha\tau v^*)}\phi_{(\bar{d}\bar{r})} \frac{\left(\tilde{y}_{\bar{r}k}^2 + \sigma_{\bar{r}k}^2\right)\gamma_{\bar{d}k}}{\bar{c}} \right]^{-1/2}.
\end{aligned}
$$

L'inégalité qui suit est vérifiée

$$
j'_{(\bar{d}\bar{r})} < \frac{\sum_{k\in U} \tilde{y}_{\bar{r}k}\gamma_{\bar{d}k}}{\sqrt{\bar{c}\,{}^{(\alpha\tau v^*)}\phi_{(\bar{d}\bar{r})}}\,{}^{(\alpha\tau v^*)}V_{(\bar{d}\bar{r})}} << 1.
$$

Donc, la norme $\left\|J_{\mathbf{j}}\left({}^{(\alpha\tau)}\mathbf{VAA}_3\right)\right\| < 1$, et par conséquent la BE converge.

Annexe A3

Preuve que l'approximation de la remarque 4.1 est à la hausse

Puisque $\hat{u}_{(dr)k}$ est la prédiction par les moindres carrés pondérés de $u_{rk}\gamma_{dk}$, en utilisant une valeur différente de $\hat{u}_{(dr)k}$, telle que $\hat{u}_{(dr)k} = 0$, nous obtenons

$$\sum_{k\in U}(1/\pi_k - 1)\,E_M\left[(u_{rk}\gamma_{dk} - \hat{u}_{(dr)k})^2\right] \leq \sum_{k\in U}(1/\pi_k - 1)\,E_M\left[(u_{rk}\gamma_{dk} - 0)^2\right],$$

où $E_M\left[(u_{rk}\gamma_{dk} - 0)^2\right] = \sigma^2_{rk}\gamma_{dk}$. En remplaçant les termes $E_M\left[(u_{rk}\gamma_{dk} - \hat{u}_{(dr)k})^2\right]$ par $\sigma^2_{rk}\gamma_{dk}$ dans l'expression (A1.5), la VAA (4.3) est surestimée. L'approximation $\hat{u}_{(dr)k} = 0$ implique que $b_{(dr)k}(\boldsymbol{\pi}) = c_{(dr)k}(\boldsymbol{\pi}) = 0$. Enfin, nous soulignons que, dans la plupart des cas, la hausse est légère, puisque les $\hat{u}_{(dr)k}$ sont obtenus au moyen des variables $\mathbf{z}_k$ qui ont généralement un pouvoir prédictif très faible pour les valeurs de $u_{rk}\gamma_{dk}$ (voir la section 4). Dans ces situations, $\hat{u}_{(dr)k} \cong (1/N)\sum_{k\in U} u_{rk}\gamma_{dk} \cong 0$. Donc $E_M(u_{rk}\gamma_{dk}\hat{u}_{(dr)k}) \cong 0$ et $E_M(\hat{u}_{(dr)k})^2 \cong 0$.

Annexe A4

Preuve de l'expression (4.7)

Dans ce cas, chaque vecteur $\boldsymbol{\delta}_k$ contient $H - 1$ éléments nuls et 1 élément égal à 1 (correspondant à la population planifiée à laquelle l'unité k appartient). Étant donné les valeurs d'entrée, la procédure d'optimisation $\pi_k = \pi_h$ pour $k \in U_h$. Sous l'hypothèse susmentionnée, $[\mathbf{A}(\boldsymbol{\pi})]^{-1}$ est une matrice diagonale dont le hh^{e} élément est donné par $[\mathbf{A}_{hh}(\boldsymbol{\pi})]^{-1} = \left[N_h\pi_h^2(1/\pi_h - 1)\right]^{-1}$. En considérant que $\tilde{y}_{rk} = \bar{Y}_{rh}$, les expressions (A1.2) et (A1.3) peuvent être reformulées, respectivement, sous la forme

$$\hat{\bar{y}}_{(dr)k} = \pi_h\boldsymbol{\delta}'_k[\mathbf{A}(\boldsymbol{\pi})]^{-1}N_h\pi_h(1/\pi_h - 1)\bar{Y}_{rh} = \bar{Y}_{rh}. \tag{A4.1}$$

$$\hat{u}_{(dr)k} = \pi_h\boldsymbol{\delta}'_k[\mathbf{A}(\boldsymbol{\pi})]^{-1}\pi_h(1/\pi_h - 1)\sum_{j\in U}u_{rj} = (\pi_h N_h)^{-1}\sum_{j\in U_h}u_{rj}, \tag{A4.2}$$

mais $\sum_{j\in U_h} u_{rj} = 0$ en tant que somme des résidus d'un modèle de régression.

En utilisant les formules (A4.1) et (A4.2), l'expression (4.5) est donnée par

$$\begin{aligned}\text{VAA}(\hat{t}_{(dr)}) &= [N/(N-H)]\sum_h\left(\frac{1}{\pi_h} - 1\right)\sum_{k\in U_h}E_M(u_{rk}\gamma_{dk})^2 \\ &= [N/(N-H)]\sum_{d=1}^{D}\sum_{h\in H_d}\sigma^2_{rh}N_h(N_h/n_h - 1),\end{aligned}$$

puisque que $\pi_h = n_h/N_h$, et l'expression (4.7) peut être obtenue.

Bibliographie

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

Boyd, S., et Vanderberg, L. (2004). *Convex Optimization*. Cambridge University Press.

Breidt, F.J., et Chauvet, G. (2011). Improved variance estimation for balanced samples drawn via the cube method. *Journal of Statistical Planning and Inference*, 141, 479-487.

Chauvet, G., Bonnéry, D. et Deville, J.-C. (2011). Optimal inclusion probabilities for balanced sampling. *Journal of Statistical Planning and Inference*, 141, 984-994.

Choudhry, G.H., Rao, J.N.K. et Hidiroglou, M.A. (2012). À propos de la répartition de l'échantillon pour une estimation sur domaine efficace. *Techniques d'enquête*, 18, 1, 25-32.

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*. New York : John Wiley & Sons, Inc.

Deville, J.-C., et Tillé, Y. (2004). Efficient balanced sampling: The cube method. *Biometrika*, 91, 893-912.

Deville, J.-C., et Tillé, Y. (2005). Variance approximation under balanced sampling, *Journal of Statistical Planning and Inference*, 128, 569-591.

Dykstra R. et Wollan P. (1987). Finding I-projections subject to a finite set of linear inequality constraints, *Applied Statistics*, 36, 377-383.

Ernst, L.R. (1989). Further applications of linear programming to sampling problems. *Proceedings of the Survey Research Methods Section,* American Statistical Association, 625-631.

Falorsi, P.D., et Righi, P. (2008). Une approche d'échantillonnage équilibré pour des plans de sondage à stratification multidimensionnelle pour l'estimation pour petits domaines. *Techniques d'enquête*, 34, 2, 247-259.

Falorsi, P.D., Orsini, D. et Righi, P. (2006). Balanced and coordinated sampling designs for small domain estimation. *Statistics in Transition*, 7, 1173-1198.

Gonzalez, J.M., et Eltinge, J.L. (2010). Optimal survey design: A review. *Section on Survey Research Methods – JSM 2010*, Octobre.

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

Khan, M.G.M., Mati, T. et Ahsan, M.J. (2010). An optimal multivariate stratified sampling design using auxiliary information: An integer solution using goal programming approach. *Journal of Official Statistics*, 26, 695-708.

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

Lu, W., et Sitter, R.R. (2002). Méthode pratique de stratification multiple par programmation linéaire. *Techniques d'enquête*, 28, 2, 215-224.

Nedyalkova, D., et Tillé, Y. (2008). Optimal sampling and estimation strategies under the linear model. *Biometrika*, 95, 521-537.

Tillé, Y. (2006). *Sampling Algorithms*. Springer-Verlag, New York.

Tillé, Y., et Favre, A.-C. (2005). Optimal allocation in balanced sampling. *Statistics and Probability Letters*, 74, 31-37.

Winkler, W.E. (2001). Multi-way survey stratification and sampling. *Research Report Series*, Statistics #2001-01. Statistical Research Division U.S. Bureau of the Census Washington D.C. 20233.