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

Techniques d'enquête

Échantillonnage optimal en coût pour l'observation intégrée de populations différentes

par Piero Demetrio Falorsi, Paolo Righi et Pierre Lavallée

Date de diffusion : le 17 décembre 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, décembre 2019
Vol. 45, nº 3, pp. 519-547
Statistique Canada, nº 12-001-X au catalogue

Échantillonnage optimal en coût pour l'observation intégrée de populations différentes

Piero Demetrio Falorsi, Paolo Righi et Pierre Lavallée[1]

Résumé

Dans les études sociales ou économiques, il faut souvent adopter une vue d'ensemble de la société. Dans les études en agriculture par exemple, on peut établir un lien entre les caractéristiques des exploitations et les activités sociales des particuliers. On devrait donc étudier un phénomène en considérant les variables d'intérêt et en se reportant à cette fin à diverses populations cibles liées entre elles. Pour se renseigner sur un phénomène, on se doit de faire des observations en toute intégration, les unités d'une population devant être observées conjointement avec les unités liées d'une autre. Dans l'exemple de l'agriculture, cela veut dire qu'on devrait prélever un échantillon de ménages ruraux qui serait lié de quelque manière à l'échantillon d'exploitations à utiliser aux fins de l'étude.

Il existe plusieurs façons de prélever des échantillons intégrés. Nous analysons ici le problème de la définition d'une stratégie optimale d'échantillonnage dans cette optique. La solution proposée doit réduire le coût d'échantillonnage au minimum et satisfaire une précision préétablie de l'estimation des variables d'intérêt (dans l'une et/ou l'autre des populations) décrivant le phénomène. L'échantillonnage indirect dresse un cadre naturel pour un tel réglage, car les unités appartenant à une population peuvent être porteuses d'une information sur l'autre population visée par l'enquête.

Nous étudions ce problème selon divers contextes caractérisant l'information sur les liens disponibles à l'étape du plan de sondage, que les liens entre les unités soient connus à ce stade ou que l'information dont nous disposons sur ces mêmes liens laisse très nettement à désirer. Nous présentons ici une étude empirique de données agricoles pour un pays en développement. On peut y voir combien il est efficace de prendre en compte les probabilités d'inclusion au stade du plan de sondage à l'aide de l'information disponible (sur les liens en l'occurrence) et à quel point on peut ainsi nettement réduire les erreurs des estimations pour la population indirectement observée. Nous démontrons enfin la nécessité de disposer de bons modèles pour la prédiction des variables ou des liens inconnus.

Mots-clés : Enquêtes intégrées; répartition de l'échantillon; échantillonnage indirect.

1 Introduction

Il est souvent nécessaire dans les études sociales ou économiques d'observer plusieurs populations liées entre elles. Dans les études en agriculture par exemple, nous pouvons faire le lien entre les caractéristiques et le comportement des exploitations et les phénomènes liés non seulement aux exploitations elles-mêmes, mais aussi aux activités sociales des particuliers. Il faut donc procéder à une certaine intégration et ajouter l'étude de la population de ménages ruraux à l'étude de la population d'exploitations. En d'autres termes, si nous désirons nous renseigner sur un certain phénomène, nous devons faire des observations en toute intégration, d'où l'implication que les unités d'une population doivent être observées conjointement avec les unités correspondantes d'une autre. Dans l'exemple qui précède, cela veut dire qu'on devrait prélever un échantillon de ménages ruraux ayant un lien quelconque avec l'échantillon d'exploitations utilisé aux fins de l'étude.

1. Piero Demetrio Falorsi et Paolo Righi, Italian National Institute of Statistics. Courriel : falorsi@istat.it, parighi@istat.it; Pierre Lavallée, Courriel : plavall1962@gmail.com.

L'observation intégrée de deux populations sous-entend que, si nous observons les variables de l'unité j de la première population U^A, nous devons observer les variables de toutes les unités qui, dans la seconde, U^B, sont liées à la j^e unité de U^A. Les liens entre les unités des deux populations obéissent à des règles, à des liens de dépendance conditionnelle ou à des relations créés formellement à de telles fins. Si nous reprenons l'exemple de l'agriculture, il s'agira fréquemment de populations statistiques différentes comme des exploitations, des ménages ruraux et des parcelles foncières, dont les unités sont liées entre elles. Un ménage peut se composer de travailleurs d'une exploitation et ceux-ci représentent le lien entre le ménage et l'exploitation. Une exploitation se compose de parcelles foncières qui sont le lien entre l'exploitation et la population de parcelles. L'observation intégrée de telles populations permet de mesurer les phénomènes d'ensemble appartenant au secteur agricole. Dans une exploitation par exemple, le niveau de scolarité de l'exploitant et la taille de la ferme, variables liées à la population d'exploitations, peuvent influer sur la productivité du sol (variable liée à la population statistique des parcelles) propre à cette exploitation. À son tour, cette productivité peut influer sur le risque de malnutrition des ménages (population des ménages ruraux) dont font partie les travailleurs de l'exploitation. Ainsi, l'observation de ces unités différentes en intégration nous renseigne sur les relations entre le niveau de scolarité, la productivité du sol et le risque de malnutrition. Si nous regardons seulement les agrégats, l'avantage qu'offre l'échantillonnage intégré est qu'on peut échantillonner dans la population U^B sans avoir une base de sondage à sa disposition pour ce faire.

Les études entreprise-établissement-travailleur sont un autre exemple concret d'usage possible de ce cadre méthodologique. Ainsi, le bien-être des ménages dont les membres travaillent dans des entreprises ayant une politique bien définie de responsabilité sociale peut différer du bien-être d'autres types de ménages et les enfants de ces premiers ménages peuvent obtenir de meilleurs résultats scolaires. Dans ce cas, l'observation intégrée permet d'étudier le comportement de sous-catégories de ménages définies par une variable observable dans la population d'entreprises.

On peut relever d'autres exemples dans les études sociodémographiques. Il est possible, par exemple, d'étudier le phénomène des enfants qui passent du temps dans deux ménages avec une observation intégrée de la population U^A de ménages et de la population U^B d'enfants.

En règle générale, cette observation en intégration peut servir à étudier des phénomènes dont les variables sont en corrélation, tout en appartenant à des populations statistiques différentes. Elle permet d'examiner les relations entre toutes les variables d'intérêt dans un même phénomène, bien que ces variables se rattachent à des populations différentes. L'observation indépendante de telles populations n'est pas l'observation de tout l'ensemble des variables d'intérêt liées, et il serait donc impossible d'étudier les relations unissant l'ensemble des variables décrivant le phénomène visé.

L'échantillonnage indirect (Lavallée, 2002, 2007) dresse un cadre naturel d'estimation des paramètres de deux populations cibles liées l'une à l'autre. Dans un tel cadre, les unités d'une population qui sont sélectionnées aux fins d'une enquête peuvent permettre de recueillir des renseignements sur une autre population par les liens entre les unités des deux. Ajoutons que l'échantillonnage indirect se prête à la

production de statistiques sur des populations pour lesquelles il n'existe pas de base de sondage. Dans ce contexte, le choix de la méthode d'échantillonnage présuppose que la population U^A est liée à la population d'intérêt U^B mais que seule la base de sondage de U^A est disponible. Nous prélevons alors un échantillon dans U^A et, par les liens entre les deux populations, observons un échantillon d'unités de U^B.

Nous étudions ici le problème du plan de sondage pour l'observation intégrée de populations différentes. C'est pourquoi nous appliquons un plan d'échantillonnage indirect en nous attachant à la détermination des probabilités d'inclusion. Comme la somme de ces probabilités définit la taille d'échantillon prévue, nous nous trouvons à définir en gros le problème comme un problème de *répartition d'échantillon*. En fait, les deux problèmes (de détermination des probabilités d'inclusion et de répartition de l'échantillon) coïncident dans l'échantillonnage stratifié. Il est question dans un certain nombre d'ouvrages et d'articles du problème de répartition dans le cadre habituel de l'échantillonnage (direct). Si un seul paramètre cible doit être estimé pour l'ensemble de la population, on peut procéder à une répartition optimale d'échantillon stratifié (Cochran, 1977; Särndal, Swensson et Wretman, 1992). Plus précisément, une répartition optimale d'échantillon minimise la variance du total estimé dans les limites d'un budget ou, pour prendre l'inverse, on peut procéder à une répartition d'échantillon qui minimise le coût avec comme contrainte une erreur quelconque d'échantillonnage. Dans un scénario à plusieurs variables où on doit mesurer plusieurs caractéristiques de chaque unité de l'échantillon, une répartition optimale pour les diverses caractéristiques n'a guère d'utilité dans la pratique à moins que les caractéristiques visées ne soient hautement en corrélation. C'est qu'une répartition optimale pour une caractéristique peut être loin de l'être pour les autres. Le caractère multidimensionnel du problème mène aussi à un compromis en matière de méthode de répartition (Khan, Mati et Ahsan, 2010) avec une perte de précision par rapport à une répartition individuelle optimale. Plusieurs auteurs ont examiné divers critères d'obtention d'un compromis pratique en matière de répartition; voir, par exemple, Kokan et Khan (1967), Chromy (1987), Bethel (1989) et Choudhry, Rao et Hidiroglou (2012).

Falorsi et Righi (2015) dressent un cadre général pour le plan de sondage d'enquêtes à variables et à domaines multiples. Ici, nous généraliserons ce cadre un peu plus en direction d'une observation intégrée de deux populations. Nous étudierons différents scénarios quant au degré de connaissance des liens. Dans le premier, nous posons que les liens entre les populations sont connus à l'étape du plan de sondage; dans le deuxième, nous estimons les liens entre U^A et U^B à cette même étape; dans le troisième, nous ne connaissons rien des liens entre U^A et U^B, mais des variables auxiliaires pour U^A peuvent nous renseigner utilement sur U^B.

Nous présentons le contexte et les symboles à la section 2. Nous illustrons le problème d'optimisation de base et la façon de l'appliquer aux divers scénarios dans les sections 3 et 4. Nous présentons enfin des résultats empiriques à la section 5.

2 Contexte

Soit U^A et U^B deux populations cibles liées entre elles, U^A étant la population pour laquelle nous disposons d'une base de sondage et U^B la population pour laquelle une base de sondage peut ou non être

disponible. Dans l'exemple de l'agriculture, U^A est la population d'exploitations et U^B, la population de ménages ruraux. Soit s^A un échantillon prélevé sur U^A sans remise; la taille fixe d'échantillon est m^A et U^A compte M^A unités. Soit π_j^A la probabilité d'inclusion de la j^e unité dans U^A avec $\pi_j^A > 0$ et $\sum_{j \in U^A} \pi_j^A = m^A$ pour $\boldsymbol{\pi}^A = \left(\pi_1^A, \ldots, \pi_j^A, \ldots, \pi_{M^A}^A\right)'$. Nous désignons par $y_{j,v}$ la valeur de la v^e $(v = 1, \ldots, V)$ caractéristique de l'unité j et le total correspondant par Y_v^A.

Nous estimons le total Y_v^A avec l'estimateur de Horvitz-Thompson (HT),

$$\hat{Y}_v^A = \sum_{j \in s^A} w_j^A \, y_{j,v}, \tag{2.1}$$

où $w_j^A = 1/\pi_j^A$.

Nombreux sont les plans de sondage qui, dans la pratique, définissent des domaines planifiés qui sont des sous-populations à tailles d'échantillon fixes avant prélèvement de l'échantillon. Désignons par $U_h^A \, (h = 1, \ldots, H)$ le domaine planifié de taille $M_h^A = \sum_{j \in U^A} d_{j(h)}$, où $d_{j(h)} = 1$ si $j \in U_h^A$ et où $d_{j(h)} = 0$ dans les autres cas. Posons que les valeurs $d_{j(h)}$ sont connues et disponibles dans la base de sondage pour toutes les unités de population. Les bases de sondage à taille fixe sont celles qui satisfont la relation

$$\sum_{j \in s^A} \mathbf{d}_j = \mathbf{m}^A,$$

où $\mathbf{d}_j = \left(d_{j(1)}, \ldots, d_{j(h)}, \ldots, d_{j(H)}\right)'$ et où $\mathbf{m}^A = \left(m_1^A, \ldots, m_h^A, \ldots, m_H^A\right)'$ est le vecteur de nombres entiers définissant les tailles fixes d'échantillon à l'étape du plan de sondage avec $\sum_{j \in U^A} d_{j(h)} \pi_j^A = m_h^A$. Dans ce cadre, les domaines planifiés peuvent être en chevauchement et, par conséquent, l'unité j peut avoir plus d'une valeur $d_{j(h)} = 1$ (pour $h = 1, \ldots, H$). Plusieurs plans de sondage usuels à taille fixe peuvent être considérés comme cas d'espèce. Un exemple bien connu est l'échantillonnage aléatoire simple stratifié sans remise (EASSSR), plan de sondage dont les strates sont les domaines planifiés et où chaque vecteur $\mathbf{d}_j$ compte $H - 1$ éléments égaux à zéro et un élément égal à l'unité, d'où l'implication que chaque unité j peut appartenir à un et un seul domaine planifié. Ajoutons que, dans un tel plan de sondage, toutes les unités de la strate U_h^A ont une probabilité uniforme d'inclusion donnée par $\pi_j^A = \frac{m_h^A}{M_h^A}$ pour $j \in U_h^A$. Si chaque vecteur $\mathbf{d}_j$ a $H - 1$ éléments à zéro et un élément à l'unité et que les valeurs π_j^A peuvent être différentes dans la strate, nous avons là un plan d'échantillonnage stratifié sans remise à tailles d'échantillon fixes et à probabilités variables dans chaque strate. D'après la définition de Winkler (2001), si $\sum_{h=1}^{H} d_{j(h)} > 1$, nous avons là un plan de sondage à stratification multidimensionnelle incomplète.

Posons que la matrice $\mathbf{D} = \left(\mathbf{d}_1, \ldots, \mathbf{d}_j, \ldots, \mathbf{d}_{M^A}\right)'$ de taille $M^A \times H$ est non singulière. En appliquant ce cadre de plan de sondage général, Deville et Tillé (2005) ont proposé une expression approchée de la variance pour $\hat{Y}_v^A$ selon la théorie d'échantillonnage de Poisson sous la forme

$$V\left(\hat{Y}_v^A \,\middle|\, \mathbf{m}^A\right) \cong \left[M^A/(M^A - H)\right] \sum_{j \in U^A} \left(\frac{1}{\pi_j^A} - 1\right) \eta_{j,v}^2, \tag{2.2}$$

où $\hat{Y}_v^A \left| \mathbf{m}^A \right.$ est l'estimateur HT en fonction d'un plan de sondage général à taille d'échantillon fixe avec $\mathbf{m}^A$ unités liées au vecteur $\boldsymbol{\pi}^A$,

$$\eta_{j,v} = y_{j,v} - \pi_j^A \mathbf{d}_j' \boldsymbol{\beta}_v, \tag{2.3}$$

et

$$\boldsymbol{\beta}_v = \boldsymbol{\Delta}^{-1} \sum_{j \in U^A} \pi_j^A \left(\frac{1}{\pi_j^A} - 1 \right) \mathbf{d}_j y_{j,v} \tag{2.4}$$

avec

$$\boldsymbol{\Delta} = \sum_{j \in U^A} \mathbf{d}_j \mathbf{d}_j' \pi_j^A \left(1 - \pi_j^A\right). \tag{2.5}$$

La variance en (2.2) ressemble à l'expression de la variance pour l'estimateur HT dans un plan d'échantillonnage de Poisson, mais elle fait intervenir les résidus $\eta_{j,v}$ au lieu de la valeur initiale $y_{j,v}$. Dans la pratique, il s'agit là de l'expression approchée de la variance avec $H = 1$ pour le plan d'échantillonnage conditionnel de Poisson (comme il est introduit dans Deville et Tillé, 2005). Dans un tel plan de sondage, on échantillonne sans remise jusqu'à obtention d'une taille d'échantillon donnée.

Pour préciser le degré d'approximation de (2.2), considérons le plan de sondage EASSSR. En appliquant l'expression en (2.2), nous obtenons

$$V\left(\hat{Y}_v^A \left| \mathbf{m}^A \right.\right) = \left[M^A / (M^A - H)\right] \sum_{h=1}^{H} \sigma_{v,h}^2 M_h^A \left(\frac{M_h^A}{m_h^A} - 1 \right),$$

où $\sigma_{v,h}^2$ est la variance de plan de sondage des valeurs $y_{j,v}$ dans la strate U_h^A (voir l'annexe 4 dans Falorsi et Righi, 2015). Cette expression approchée donne de bons résultats lorsque le nombre de domaines H demeure petit à comparer à la taille globale de population M^A.

Soit M^B, N^B, U_i^B et M_i^B le nombre d'unités dans U^B, le nombre de grappes dans U^B, la i^e grappe de U^B avec $\bigcup_{i=1}^{N^B} U_i^B = U^B$ et le nombre d'unités dans la i^e grappe U_i^B respectivement. Nous désignons par $y_{ik,r}$ la valeur de la r^e $(r = 1, \ldots, R)$ caractéristique de la k^e unité de la i^e grappe de U^B. Nous désignons le total de population pour l'ensemble des $y_{ik,r}$ par

$$Y_r^B = \sum_{i=1}^{N^B} \sum_{k=1}^{M_i^B} y_{ik,r}.$$

Soit $l_{j,ik}$ une variable indicatrice de l'existence d'un lien : $l_{j,ik} = 1$ s'il existe un lien entre la j^e unité dans U^A et la k^e unité dans U_i^B et $l_{j,ik} = 0$ dans les autres cas.

Supposons que nous procédons à un échantillonnage indirect : si l'unité $j \in U^A$ est incluse dans s^A, toutes les grappes U_i^B, pour lesquelles $L_{j,i}^B = \sum_{k=1}^{M_i^B} l_{j,ik} > 0$, sont observées $\left(\text{c'est-à-dire les } y_{ik,r}\right)$ dans l'échantillon indirect de la population U^B. Soit n^B la taille de l'échantillon de grappes de la population U^B que dégage cet échantillonnage indirect. Nous estimons Y_r^B avec l'estimateur fondé sur la théorie de la méthode généralisée de partage des poids (MGPP) de Lavallée (2002, 2007) :

$$\hat{Y}_r^B = \sum_{i=1}^{n^B} y_{i,r} w_i^B, \tag{2.6}$$

où

$$y_{i,r} = \sum_{k=1}^{M_i^B} y_{ik,r}$$

et

$$w_i^B = \sum_{j \in s^A} w_j^A \tilde{L}_{j,i}^B$$

avec

$$\tilde{L}_{j,i}^B = \frac{L_{j,i}^B}{L_i^B}$$

et

$$L_i^B = \sum_{j=1}^{M^A} L_{j,i}^B .$$

Le théorème à la section 3 de Lavallée (2002, 2007) dit que (2.6) donne un estimateur sans biais pour Y_r^B à condition que tous les liens $l_{j,ik}$ puissent être dûment identifiés et avec $L_i^B > 0$ pour tous les $i \in U^B$. En définissant

$$z_{j,r} = \sum_{i=1}^{N^B} \tilde{L}_{j,i}^B \, y_{i,r}, \tag{2.7}$$

on peut exprimer l'estimateur en (2.6) comme étant l'estimateur habituel de Horvitz-Thompson sur les valeurs z renvoyant à la population U^A,

$$\hat{Y}_r^B = \sum_{s^A} z_{j,r} w_j^A. \tag{2.8}$$

Ainsi, la variance $V\left(\hat{Y}_r^B\right)$ de $\hat{Y}_r^B$ peut s'exprimer comme la variance de l'estimateur HT sur la population U^A. La variance approchée de $\hat{Y}_r^B$ pour les plans de sondage à taille fixe est donnée par

$$V\left(\hat{Y}_r^B \,\middle|\, \mathbf{m}^A\right) \cong \left[M^A/(M^A - H)\right] \sum_{j \in U^A} \left(\frac{1}{\pi_j^A} - 1\right) \eta_{j,r}^2, \tag{2.9}$$

où $\hat{Y}_r^B \,\middle|\, \mathbf{m}^A$ est l'estimateur HT en fonction d'un plan de sondage général à taille fixe d'échantillon avec $\mathbf{m}^A$ unités et le vecteur lié $\boldsymbol{\pi}^A$,

$$\eta_{j,r} = z_{j,r} - \pi_j^A \mathbf{d}_j' \boldsymbol{\beta}_r$$

avec

$$\boldsymbol{\beta}_r = \boldsymbol{\Delta}^{-1} \sum_{j \in U^A} \pi_j^A \left(\frac{1}{\pi_j^A} - 1 \right) \mathbf{d}_j \; z_{j,r}.$$

Remarque 2.1. Comme prolongement intéressant du cadre qui précède et utile dans le cas des études intégrées, il y a le cas d'un total calculé par croisement d'une variable de la population U^A et d'une variable de la population U^B. Pour illustrer, soit y_v une variable de U^A à C modalités et $y_{j,v(c)}$ une variable dichotomique où $y_{j,v(c)} = 1$ si l'unité j est caractérisée par la modalité $c\,(c = 1, \ldots, C)$ de y_v, et où $y_{j,v(c)} = 0$ dans les autres cas. De plus, soit y_r une variable de U^B à G modalités et $y_{ik,r(g)}$ une variable dichotomique où $y_{ik,r(g)} = 1$ si l'unité k de la grappe i est caractérisée par la modalité $g\,(g = 1, \ldots, G)$ de y_r et où $y_{ik,r(g)} = 0$ dans les autres cas. Le nombre total d'unités de U^B caractérisées par la modalité $g\,(g = 1, \ldots, G)$ de y_r et liées aux unités de la population U^A caractérisées par la modalité c de la variable y_v peut se définir comme

$$Y_{(c,g)}^B = \sum_{j=1}^{N^A} \sum_{i=1}^{N^B} \sum_{k=1}^{M_i^B} l_{j,ik} y_{j,v(c)} \, y_{ik,r(g)} = \sum_{i=1}^{N^B} y_{i,(c,g)},$$

où $y_{i,(c,g)} = \sum_{j=1}^{N^A} \sum_{k=1}^{M_i^B} l_{j,ik} \; y_{j,v(c)} y_{ik,r(g)}$.

À titre d'exemple, prenons le cas illustré en introduction d'une analyse intégrée portant sur la productivité des exploitations et la malnutrition des ménages et posons que $Y_{(c,g)}^B$ représente le total de membres ayant un problème de malnutrition dans les ménages de travailleurs des exploitations caractérisées par une haute productivité. Dans ce cas, $y_{j,v(c)}$ prend la valeur 1 si la productivité de l'exploitation j est élevée et $y_{ik,r(g)}$ prend la même valeur unité si le membre k du ménage i connaît un problème de malnutrition.

On peut tirer directement l'estimateur MGPP de $Y_{(c,g)}^B$ de l'expression en (2.8) à l'aide de la variable transformée $z_{j,r} = \sum_{i=1}^{N^B} \tilde{L}_{j,i}^B \; y_{i,(c,g)}$.

3 Énoncé du problème

Vu le cadre qui précède, nous désirons trouver le vecteur $\boldsymbol{\pi}^A = \left(\pi_1^A, \ldots, \pi_j^A, \ldots, \pi_{M^A}^A\right)'$ des probabilités d'inclusion qui minimise le coût d'enquête prévu bornant les variances d'échantillonnage $V\left(\hat{Y}_v^A \,\middle|\, \mathbf{m}^A\right)$ $(v = 1, \ldots, V)$ et $V\left(\hat{Y}_r^B \,\middle|\, \mathbf{m}^A\right)$ $(r = 1, \ldots, R)$ selon des contraintes données de variance :

$$\begin{cases} \min \sum_{j \in U^A} c_j \; \pi_j^A \\ V\left(\hat{Y}_v^A \,\middle|\, \mathbf{m}^A\right) \leq V_v^* \quad \forall v = 1, \ldots, V \\ V\left(\hat{Y}_r^B \,\middle|\, \mathbf{m}^A\right) \leq V_r^* \quad \forall r = 1, \ldots, R \\ 0 < \pi_j^A \leq 1 \quad \forall j = 1, \ldots, M^A \end{cases} \tag{3.1}$$

où $\mathbf{m}^A$ est donné par le vecteur $\boldsymbol{\pi}^A$ minimisant la fonction de coût, où $V_v^*\,(v = 1, \ldots, V)$ et V_r^* $(r = 1, \ldots, R)$ sont les seuils de variance fixés par l'auteur du plan de sondage et où c_j est le coût variable

d'observation de l'unité j dans la population U^A et des unités liées $L_j^A = \sum_{i=1}^{N^B} L_{j,i}^B$ dans la population U^B. En d'autres termes, nous recherchons les probabilités de sélection optimale qui minimiseront la variance des estimations établies tant pour U^A que pour U^B. Dans l'exemple de l'agriculture, cela se traduirait par le calcul de probabilités de sélection optimale pour les estimations de la population d'exploitations et de la population de ménages ruraux selon la contrainte de précision spécifiée.

Une expression raisonnable de c_j est

$$c_j = f_c\left(C^A, L_j^A; C^B\right), \tag{3.2}$$

où f_c est une fonction non décroissante monotone connue, où C^A est le coût par unité d'observation de la population U^A et où C^B est le coût d'observation par unité élémentaire de la population U^B. Brewer et Gregoire (2009) proposent une vaste analyse des différentes formes que prend la fonction de coût.

Le problème de minimisation en (3.1) est une généralisation de la méthode d'optimisation à une variable sous contrainte de précision (Cochran, 1977). Avec ce problème, nous posons que toutes les valeurs $y_{j,v}$, $y_{i,r}$, L_j^A, $L_{j,i}^B$, L_i^B, $\boldsymbol{\beta}_v$ et $\boldsymbol{\beta}_r$ sont connues, auquel cas le problème en (3.1) devient un classique problème convexe linéaire séparé ou PCLS (Boyd et Vanderberg, 2004), lequel peut se résoudre par l'algorithme que propose Chromy (1987). Celui-ci a été conçu à l'origine pour une répartition optimale à plusieurs variables dans un plan de sondage EASSSR 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/.). Autre possibilité, on peut résoudre le PCLS par la procédure NLP en SAS que proposent Choudhry et coll. (2012). Les vecteurs $\boldsymbol{\beta}_v$ et $\boldsymbol{\beta}_r$ dépendent du vecteur $\boldsymbol{\pi}^A$. Falorsi et Righi (2015) définissent un nouvel algorithme de recherche d'une solution optimale compte tenu de la dépendance entre $\boldsymbol{\beta}_v$ et $\boldsymbol{\beta}_r$ avec le vecteur optimal $\boldsymbol{\pi}^A$.

4 Contextes informatifs et problème d'optimisation

Comme ils sont évoqués en (3.1), les problèmes d'optimisation sont largement théoriques, car on se doit de connaître les valeurs des variables d'intérêt dans les deux populations U^A et U^B, ainsi que les valeurs des liens qui existent entre les unités de l'une et de l'autre. Présentons maintenant trois autres contextes bien concrets avec une quantité variable d'information. Commençons par deux contextes à information très riche et enchaînons avec le troisième à information très pauvre. Ce dernier contexte est le plus courant, bien que les deux premiers gagnent en vraisemblance avec la disponibilité croissante de registres administratifs et de logiciels statistiques pour l'intégration des données.

Contexte 1. Nous disposons de bases de sondage pour U^A et U^B. Toutes les valeurs L_j^A, $L_{j,i}^B$ et L_i^B sont connues et les valeurs de $y_{j,v}$, $y_{i,r}$ sont inconnues, mais peuvent se prédire par des modèles appropriés de superpopulation.

Il peut s'agir d'un scénario réaliste dans des pays comme les pays scandinaves qui ont des systèmes de registres bien établis (Wallgren et Wallgren, 2014) où les unités d'un registre statistique comportent des identificateurs uniques de bonne qualité, d'où la possibilité de reconnaître une même unité dans l'ensemble des registres. En agriculture par exemple, cela permet de lier chaque exploitation à un ou plusieurs ménages ruraux et chaque ménage rural à une ou plusieurs exploitations.

Les modèles *de travail* que nous étudions peuvent s'exprimer sous les formes suivantes :

Niveau de l'unité **Niveau de la grappe**

$$\begin{cases} y_{j,v} = \tilde{y}_{j,v} + u_{j,v} = f_v\left(\mathbf{x}_j; \boldsymbol{\varphi}_v\right) + u_{j,v} \\ E_{M_v}\left(u_{j,v}\right) = 0, \;\; E_{M_v}\left(u_{j,v}^2\right) = \sigma_{j,v}^2, \;\; \forall j \\ E_{M_v}\left(u_{j,v}, u_{l,v}\right) = 0, \;\; \forall j \neq l \end{cases} \quad , \quad \begin{cases} y_{i,r} = \tilde{y}_{i,r} + u_{i,r} = f_r\left(\mathbf{x}_i; \boldsymbol{\varphi}_r\right) + u_{i,r} \\ E_{M_r}\left(u_{i,r}\right) = 0, \;\; E_{M_r}\left(u_{i,r}^2\right) = \sigma_{i,r}^2, \;\; \forall i \\ E_{M_r}\left(u_{i,r}, u_{i',r}\right) = 0, \;\; \forall i \neq i' \end{cases} \tag{4.1}$$

où, si nous omettons les indices pour simplifier, les $\mathbf{x}$ sont les vecteurs de prédicteurs (disponibles dans les deux bases de sondage), où les $\boldsymbol{\varphi}$ sont des vecteurs des coefficients de régression et les $f(\mathbf{x}; \boldsymbol{\varphi})$ des fonctions connues, où les u sont les termes d'erreur, où $\tilde{y}$ sont les valeurs prédites et où enfin $E_M(\cdot)$ désigne les espérances selon les modèles. Les prédicteurs $\mathbf{x}$ peuvent varier dans les modèles selon qu'il s'agit de l'unité ou de la grappe. Nous supposons que les paramètres des modèles sont connus, bien que, dans la pratique, ils fassent habituellement l'objet d'une estimation.

Même si le modèle $f_r(\cdot)$ demeure inconnu, il est possible de calculer les espérances au niveau de la grappe pour la population U^B à partir d'un modèle défini au niveau de l'unité *élémentaire*, ce qu'indique $f_{re}(\cdot)$. Le modèle au niveau de l'unité élémentaire peut se formuler comme $y_{ik,r} = \tilde{y}_{ik,r} + u_{ik,r} = f_{re}(\mathbf{x}_{ik}; \boldsymbol{\varphi}_r) + u_{ik,r}$; $E_{M_{re}}\left(u_{ik,r}\right) = 0$; $E_{M_{re}}\left(u_{ik,r}^2\right) = \sigma_r^2$; $E_{M_{re}}(u_{ik,r}, u_{ik',r}) = \sigma_r^2 \;\rho_r \;\; \forall k \neq k'$; $E_{M_{re}}\left(u_{ik,r}, u_{i'k',r}\right) = 0 \;\; \forall i \neq i'$, où ρ_r est la corrélation intragrappe.

Les espérances des modèles au niveau de la grappe du côté droit de (4.1) peuvent aisément se calculer comme :

$$\tilde{y}_{i,r} = \sum_{k=1}^{M_i^B} \tilde{y}_{ik,r}; \;\; \sigma_{i,r}^2 = M_i^B \sigma_r^2 \left[1 + \left(M_i^B - 1\right)\rho_r\right]; \;\; E_{M_r}\left(u_{i,r}, u_{i',r}\right) = 0 \;\; \text{pour} \;\; i \neq i'.$$

À noter que les modèles *de travail* en (4.1) sont propres aux variables. Nous les présentons comme moyen utile d'établissement du plan de sondage, mais ils ne sont pas nécessairement la représentation fidèle des modèles réels qui génèrent les données.

Selon (4.1), les prédictions de modélisation et les variances des variables z sont données par

$$E_{M_r}\left(z_{j,r}\right) = \tilde{z}_{j,r} = \sum_{i=1}^{N^B} \tilde{L}_{j,i}^B \, \tilde{y}_{i,r} \quad \text{et} \quad V_{M_r}\left(z_{j,r}\right) = \sigma_{j,zr}^2 = \sum_{i=1}^{N^B} \left(\tilde{L}_{j,i}^B\right)^2 \sigma_{i,r}^2. \tag{4.2}$$

Dans le problème d'optimisation en (3.1), les termes de variance $V\left(\hat{Y}_v^A \left| \mathbf{m}^A \right.\right)$ et $V\left(\hat{Y}_r^B \left| \mathbf{m}^A \right.\right)$ sont donc remplacés par les variances anticipées. Si on désigne par $E(\cdot)$ l'espérance dans le plan de sondage, la variance anticipée (VA) de $\hat{Y}_v^A$ peut ainsi se reformuler :

$$\mathrm{VA}\left(\hat{Y}_v^A\right) = E_{M_v} E\left(\hat{Y}_v^A - Y_v^A\right)^2 = E_{M_v} V\left(\hat{Y}_v^A - Y_v^A\right) + V_{M_v} E\left(\hat{Y}_v^A - Y_v^A\right).$$

Nous avons

$$E\left(\hat{Y}_v^A - Y_v^A\right) = 0,$$

et

$$V\left(\hat{Y}_v^A - Y_v^A\right) = V\left(\hat{Y}_v^A \left| \mathbf{m}^A \right.\right) \cong \sum_{j \in U^A} \left(\frac{1}{\pi_j^A} - 1\right) \eta_{j,v}^2.$$

Le même résultat peut s'obtenir pour l'estimation $\hat{Y}_r^B$. Nous avons ainsi les expressions suivantes :

$$\mathrm{VA}\left(\hat{Y}_v^A\right) = E_{M_v} V\left(\hat{Y}_v^A \left| \mathbf{m}^A \right.\right) \cong \sum_{j \in U^A} \left(\frac{1}{\pi_j^A} - 1\right) E_{M_v}\left(\eta_{j,v}^2\right) \tag{4.3}$$

$$\mathrm{VA}\left(\hat{Y}_r^B\right) = E_{M_r} V\left(\hat{Y}_r^B \left| \mathbf{m}^A \right.\right) \cong \sum_{j \in U^A} \left(\frac{1}{\pi_j^A} - 1\right) E_{M_r}\left(\eta_{j,r}^2\right) \tag{4.4}$$

où $E_{M_v}\left(\eta_{j,v}^2\right)$ et $E_{M_r}\left(\eta_{j,r}^2\right)$ sont données par les expressions (A.2) et (B.2) dans les annexes A et B.

Le problème en (3.1) de recherche du vecteur $\boldsymbol{\pi}^A$ optimal se reformule de la manière suivante :

$$\begin{cases} \min \sum_{j \in U^A} c_j \ \pi_j^A \\ E_{M_v} V\left(\hat{Y}_v^A \left| \mathbf{m}^A \right.\right) \le V_v^* \quad \forall v = 1, \ldots, V \\ E_{M_r} V\left(\hat{Y}_r^B \left| \mathbf{m}^A \right.\right) \le V_r^* \quad \forall r = 1, \ldots, R \\ 0 < \pi_j^A \le 1 \quad \forall j = 1, \ldots, M^A. \end{cases} \tag{4.5}$$

Remarque 4.1. Les variances anticipées en (4.5) ont une formulation lourde. Une expression simplifiée et conservatrice de $E_{M_v} V\left(\hat{Y}_v^A \left| \mathbf{m}^A \right.\right)$ figure à la remarque 4.1 de Falorsi et Righi (2015). On peut obtenir des approximations conservatrices encore plus simplifiées tant pour $E_{M_v} V\left(\hat{Y}_v^A \left| \mathbf{m}^A \right.\right)$ que pour $E_{M_r} V\left(\hat{Y}_r^B \left| \mathbf{m}^A \right.\right)$ en approchant la variance de plan de sondage par la variance d'échantillonnage de Poisson. Nous avons alors

$$E_{M_v} V\left(\hat{Y}_v^A \left| \mathbf{m}^A \right.\right) \le \sum_{j \in U^A} \left(\frac{1}{\pi_j^A} - 1\right) E_{M_v}\left(y_{j,v}^2\right),\ E_{M_r} V\left(\hat{Y}_r^B \left| \mathbf{m}^A \right.\right) \le \sum_{j \in U^A} \left(\frac{1}{\pi_j^A} - 1\right) E_{M_r}\left(z_{j,r}^2\right),$$

remplaçant respectivement $\eta_{j,v}$ et $\eta_{j,r}$ par $y_{j,v}$ et $z_{j,r}$, où $E_{M_v}\left(y_{j,v}^2\right) = \tilde{y}_{j,v}^2 + \sigma_{j,v}^2$ et $E_{M_r}\left(z_{j,r}^2\right) = \tilde{z}_{j,r}^2 + \sigma_{j,zr}^2$ (voir l'annexe B). Les approximations conservatrices constituent un choix sûr dans ce cadre,

puisqu'elles écartent le risque de définition d'une taille d'échantillon insuffisante pour les précisions prévues.

Remarque 4.2. Lavallée et Labelle-Blanchet (2013) traitent du problème de l'échantillonnage indirect appliqué à des populations asymétriques en proposant huit autres méthodes de modification des liens $l_{j,ik}$, de manière à réduire la variance des estimations avec de telles populations, tout en gardant l'estimation sans biais. Par les méthodes 2 et 3 que proposent ces auteurs, nous pouvons exécuter l'algorithme en remplaçant simplement les liens $l_{j,ik}$ par les liens pondérés $\theta_{j,ik}$, dans $E_{M_r} V\left(\hat{Y}_r^B \,\middle|\, \mathbf{m}^A\right)$.

Contexte 2. Nous ne connaissons pas les liens $l_{j,ik}$ avec certitude, mais nous disposons des probabilités d'existence de liens sous la forme $\Pr\left(l_{j,ik} = 1\right) = \lambda_{j,ik}$.

Pour intégrer l'incertitude des liens à l'optimisation, nous posons que ceux-ci suivent un modèle de Bernoulli M_l, $l_{j,ik} \sim \mathrm{B}\left(\lambda_{j,ik}\right)$, où $E_{M_l}\left(l_{j,ik}\right) = \lambda_{j,ik}$ et $V_{M_l}\left(l_{j,ik}\right) = \lambda_{j,ik}\left(1-\lambda_{j,ik}\right)$. Nous supposons connus les paramètres $\lambda_{j,ik}$, bien que, dans la pratique, ils soient habituellement estimés par des méthodes probabilistes de couplage d'enregistrements (Lavallée et Caron, 2001). Dans l'exemple de l'agriculture, une telle situation peut se présenter quand, par exemple, la population d'exploitations est liée à la population de ménages ruraux par une méthode probabiliste de couplage, aucun identificateur commun n'étant présent. Dans ce cadre, la variance attendue doit tenir compte des deux modèles M_l et M_r. Comme

$$E_{M_l} E_{M_r} E\left(\hat{Y}_r^B - Y_r^B\right)^2 = E_{M_l} E_{M_r} V\left(\hat{Y}_r^B - Y_r^B\right) + E_{M_l} V_{M_r} E\left(\hat{Y}_r^B - Y_r^B\right) + V_{M_l} E_{M_r} E\left(\hat{Y}_r^B - Y_r^B\right)$$

et $E\left(\hat{Y}_r^B - Y_r^B\right) = 0$, le problème en (4.5) peut se reformuler :

$$\begin{cases} \min \sum_{j \in U^A} E_{M_l}\left(c_j\right)\ \pi_j^A \\ E_{M_v} V\left(\hat{Y}_v^A \,\middle|\, \mathbf{m}^A\right) \le V_v^* \quad \forall v = 1, \ldots, V \\ E_{M_l} E_{M_r} V\left(\hat{Y}_r^B \,\middle|\, \mathbf{m}^A\right) \le V_r^* \quad \forall r = 1, \ldots, R \\ 0 < \pi_j^A \le 1 \quad \forall j = 1, \ldots, M^A \end{cases} \tag{4.6}$$

où

$$E_{M_l} E_{M_r} V\left(\hat{Y}_r^B \,\middle|\, \mathbf{m}^A\right) \cong \sum_{j \in U^A} \left(\frac{1}{\pi_j^A} - 1\right) E_{M_l} E_{M_r}\left(\eta_{j,r}^2\right), \tag{4.7}$$

$$E_{M_l}\left(c_j\right) = f_c\left(\Lambda_j^A; C^B\right),$$

avec $\Lambda_j^A = \sum_{i=1}^{N^B} \Lambda_{j,i}^B$ et $\Lambda_{j,i}^B = \sum_{k=1}^{M_i^B} \lambda_{j,ik}$.

Les principaux résultats du calcul de l'expression de $E_{M_l} E_{M_r} V\left(\hat{Y}_r^B \,\middle|\, \mathbf{m}^A\right)$ sont donnés à l'annexe C. Ils s'obtiennent par une approximation en série de Taylor et un postulat d'indépendance du processus de génération des liens $l_{j,ik}$ par rapport au processus de création des variables d'intérêt $y_{i,r}$. Avec ces approximations, les valeurs prédites $\tilde{z}_{j,r}$ s'obtiennent comme

$$\tilde{z}_{j,r} \cong \sum_{i=1}^{N^B} \tilde{\Lambda}^B_{j,i} \tilde{y}_{i,r} \tag{4.8}$$

où

$$\tilde{\Lambda}^B_{j,i} = \frac{\Lambda^B_{j,i}}{\Lambda^B_i}$$

avec

$$\Lambda^B_i = \sum_{j=1}^{M^A} \Lambda^B_{j,i}. \tag{4.9}$$

L'incertitude quant au coût total d'enquête, qui dépend à la fois de l'échantillon prélevé et de l'incertitude du modèle de coût, nous oblige à considérer l'espérance de coût $E_{M_l}\left(c_j\right)$ dans le problème d'optimisation. Steel et Clark (2014) démontrent en quoi l'incertitude de l'espérance de coût peut influer sur l'exactitude du plan de sondage.

Contexte 3. Une intégration des données est impossible, parce que le processus de couplage d'enregistrements ne donne pas de bons liens ou simplement parce qu'il n'y a pas de base de sondage pour la population U^B.

C'est le contexte le plus fréquent dans les pays en développement. Il peut aussi être caractéristique de contextes particuliers d'enquête dans les pays développés s'il s'agit, par exemple, de populations difficiles à joindre. Pour en revenir à l'exemple de l'agriculture, cela signifierait que nous pourrions disposer d'une liste d'exploitations, mais non d'une liste de ménages ruraux. Dans ce cas, le problème d'un échantillonnage intégré optimal peut se résoudre par l'utilisation de toute l'information disponible, même de piètre qualité. Dans ce qui suit, nous illustrons trois modes de solution du problème d'optimisation entre l'exigence d'un minimum d'information et la nécessité d'exploiter une information plus riche pouvant se révéler coûteuse à obtenir.

Option 3.1. *Élaborer les prédictions des variables z et abaisser les seuils de variance V_r^* par un facteur d'échelle.* Posons qu'il est possible, à partir de la base de sondage U^A, de connaître les valeurs d'une variable de taille γ liée à l'ensemble des liens L_j^A des unités j. À titre d'exemple, si la population U^A comprend les exploitations et la population U^B les ménages, le nombre de travailleurs dans les exploitations (variable γ_j) peut offrir une bonne approximation du nombre total de liens, L_j^A, de ces exploitations. Posons en outre que les totaux ou les totaux estimés, $\tilde{Y}^B_{r(q)}$, sont disponibles à un certain niveau de domaine, $U^B_{(q)}$ $(q = 1, \ldots, Q)$, défini sur le plan géographique, avec $U^B = \bigcup_{q=1}^{Q} U^B_{(q)}$ et $U^B_{(q)} \cap U^B_{(q')} = \varnothing$ pour $q \neq q'$. Les variables z prédites peuvent alors se définir comme :

$$\tilde{z}_{j,r} = \frac{\gamma_j}{\sum_{l \in U^A_{(q)}} \gamma_l} \tilde{Y}^B_{r(q)} \quad \text{pour} \quad j \in U^A_{(q)}, \tag{4.10}$$

où $U^A_{(q)}$ désigne le domaine géographique q pour la population U^A. Dans la pratique, la méthode du quotient en (4.10) présuppose que l'unité j peut se voir attribuer une partie du total $\tilde{Y}^B_{r(q)}$ proportionnellement à la taille de l'unité même. D'autres exemples d'élaboration des prédictions des valeurs z sont cités à la section 5.3.2 de *Guidelines for the Integrated Survey Framework* (FAO, 2015).

Une fois élaborées les prédictions, $\tilde{z}_{j,r}$, il peut être raisonnable de supposer que la relation suivante se vérifie :

$$E_{M_{zr}}\left(z^2_{j,r}\right) = \tilde{z}^2_{j,r} + \sigma^2_{j,zr} \cong k_r \tilde{z}^2_{j,r}, \tag{4.11}$$

où $k_r > 1$. Selon (4.11), il est simple de démontrer que

$$E_{M_{zr}} V\left(\hat{Y}^B_r \,\middle|\, \mathbf{m}^A\right) \cong k_r V\left(\hat{\tilde{Y}}^B_r \,\middle|\, \mathbf{m}^A\right),$$

où $\hat{\tilde{Y}}^B_r = \sum_{j\in s^A} w^A_j \tilde{z}_{j,r}$. Nous pouvons calculer la variance d'échantillonnage $V\left(\hat{\tilde{Y}}^B_r \,\middle|\, \mathbf{m}^A\right)$ à l'aide des expressions en (2.2), (2.3), (2.4) et (2.5) par substitution de la variable $y_{j,v}$ (prédiction $\tilde{z}_{j,r}$). Nous pouvons alors reformuler le problème d'optimisation en recherchant le vecteur $\boldsymbol{\pi}^A$ optimal :

$$\begin{cases} \min \sum_{j\in U^A} E_{M_\Lambda}\left(c_j\right)\pi^A_j \\ E_{M_v} V\left(\hat{Y}^A_v \,\middle|\, \mathbf{m}^A\right) \le V^*_v \quad \forall v = 1, \ldots, V \\ V\left(\hat{\tilde{Y}}^B_r \,\middle|\, \mathbf{m}^A\right) \le V^*_r / k_r \quad \forall r = 1, \ldots, R \\ 0 < \pi^A_j \le 1 \quad \forall j = 1, \ldots, M^A. \end{cases} \tag{4.12}$$

L'auteur du plan de sondage peut trouver la solution en exécutant le problème en (4.12) avec d'autres choix raisonnables de la valeur k_r ($k_r = 2$, 3 ou 4, par exemple) et en étudiant la sensibilité des différentes solutions. À noter que $k_r \cong 1 + \left[\mathrm{CV}\left(z_{j,r}\right)\right]^2$, où $\left[\mathrm{CV}\left(z_{jr}\right)\right]^2 = \sigma^2_{j,zr} / \tilde{z}^2_{j,r}$. Ainsi, (4.11) se vérifie si les valeurs $\left[\mathrm{CV}\left(z_{j,r}\right)\right]^2$ sont approximativement constantes.

Option 3.2. *Scénario extrême du contexte 2 avec uniformité des liens dans certains domaines.* Si le nombre ou le nombre estimé de grappes et d'unités élémentaires $N^B_{(q)}$ et $M^B_{(q)}$ des domaines $U^B_{(q)}$ $(q = 1, \ldots, Q)$ est disponible, nous pouvons à bon droit supposer que, en cas d'absence d'information sur les liens $l_{j,ik}$, ceux-ci sont homogènes sur les domaines, c'est-à-dire que $l_{j,ik} \sim B\left(\lambda_{j,ik}\right)$, où $\lambda_{j,ik} = \gamma_j / M^B_{(q)}$.

Posons en outre que, dans ce contexte, les prédictions $\tilde{y}_{i,r}$ et les variances d'échantillonnage $\sigma^2_{i,r}$ peuvent être tenues pour homogènes à l'intérieur des domaines $U^B_{(q)}$, c'est-à-dire que $\tilde{y}_{i,r} = \tilde{y}_{r(q)}$ et $\sigma^2_{i,r} = \sigma^2_{r(q)}$ pour $i \in U^B_{(q)}$. Nous pouvons alors résoudre le problème d'optimisation comme cas extrême du contexte 2 avec uniformité des liens par domaine.

Remarque 4.3. Il convient de noter que, avec l'option (3.2), les prédictions $\tilde{z}_{j,r}$ équivalent à celles qui sont exprimées en (4.10). Il est en effet raisonnable de considérer que, en cas d'absence d'information, la taille en unités élémentaires de la grappe U^B_i peut être tenue pour égale à sa moyenne définie au niveau du domaine : $M^B_i \cong \bar{M}^B_{(q)} = M^B_{(q)} / N^B_{(q)}$ pour $U^B_i \in U^B_{(q)}$. Dans ce cas, les approximations suivantes se vérifient :

$$\Lambda_{j,i}^{B} = \sum_{k \in U_i^B} \lambda_{j,ik} \cong \bar{M}_{(q)}^{B} \frac{\gamma_j}{M_{(q)}^{B}} = \frac{\gamma_j}{N_{(q)}^{B}}; \quad \Lambda_i^B = \sum_{j \in U_{(q)}^{A}} \Lambda_{j,i}^{A} \cong \frac{1}{N_{(q)}^{B}} \sum_{j \in U_{(q)}^{A}} \gamma_j .$$

Si on pose $\tilde{Y}_{r(q)}^{B} = \tilde{y}_{r(q)}^{B} N_{(q)}^{B}$ et postule l'indépendance du processus qui génère les liens $l_{j,ik}$ par rapport au processus de création des variables d'intérêt $y_{i,r}$, nous pouvons obtenir

$$\tilde{z}_{j,r} \cong \sum_{i \in U_{(q)}^{B}} \frac{\Lambda_{j,i}^{B}}{\Lambda_i^B} \tilde{y}_{r(q)} = \sum_{i \in U_{(q)}^{B}} \frac{\gamma_j \big/ N_{(q)}^{B}}{\sum_{j \in U_{(q)}^{A}} \gamma_j \big/ N_{(q)}^{B}} \tilde{y}_{r(q)} = \frac{\gamma_j}{\sum_{j \in U_{(q)}^{A}} \gamma_l} \tilde{y}_{r(q)}^{B} N_{(q)}^{B} \quad \text{pour} \quad j \in U_{(q)}^{A}.$$

Option 3.3. *Modéliser les valeurs* $z_{j,r}$. Une autre possibilité est de modéliser directement les valeurs z et le nombre total de liens L_j^A avec des modèles de la nature suivante :

$$\begin{cases} z_{j,r} = \tilde{z}_{j,r} + u_{j,zr} = f_{zr}\left(\mathbf{x}_j; \boldsymbol{\varphi}_r\right) + u_{j,zr} \\ E_{M_{zr}}\left(u_{j,zr}\right) = 0, \; E_{M_{zr}}\left(u_{j,zr}^2\right) = \sigma_{j,zr}^2, \; \forall j \\ E_{M_{zr}}\left(u_{j,zr}, \; u_{j',zr}\right) = 0, \; \forall j \neq j' \end{cases} , \quad \begin{cases} L_j^A = \Lambda_j^A + u_{j,\Lambda} = f_{\Lambda}\left(\boldsymbol{\theta}_j; \boldsymbol{\varphi}_{\Lambda}\right) + u_{j,\Lambda} \\ E_{M_{\Lambda}}\left(u_{j,\Lambda}\right) = 0, \; E_{M_{\Lambda}}\left(u_{j,\Lambda}^2\right) = \sigma_{j,\Lambda}^2, \; \forall j \\ E_{M_r}\left(u_{j,\Lambda}, \; u_{j',\Lambda}\right) = 0, \; \forall j \neq j' \end{cases} \quad (4.13)$$

où $\mathbf{x}_j$ et $\boldsymbol{\theta}_j$ sont des vecteurs de variables auxiliaires. Les prédictions Λ_j^A doivent être positives. Un modèle utile à cet égard est le modèle loglinéaire (Xu et Lavallée, 2009) : $\log\left(\Lambda_j^A\right) = \boldsymbol{\theta}_j' \boldsymbol{\varphi}_{\Lambda}$. Le modèle du côté droit de (4.13) permet de prédire le nombre total de liens Λ_j^A de l'unité j, ce qui définira l'espérance de coût d'enquête qui s'y attache. Nous pourrions exécuter le problème d'optimisation en nous reportant aux variances des prédictions des modèles (4.13).

Remarque 4.4. L'option 3.1 exige le minimum d'information pour l'élaboration des prédictions $\tilde{z}_{j,r}$. Il nous faut définir des valeurs vraisemblables pour les constantes k_r. L'option 3.2 fait intervenir la même information que l'option 3.1 pour cette élaboration (voir la remarque 4.3), mais nous aurons besoin d'une estimation des paramètres $\sigma_{r(q)}^2$. De telles estimations peuvent s'obtenir par des enquêtes pilotes ou antérieures portant directement sur la population U^B. L'option 3.3 est la plus complexe et la plus coûteuse, puisqu'il faudra des enquêtes pilotes indirectes sur la population U^A pour que puissent être élaborées des prédictions vraisemblables des paramètres $\tilde{z}_{j,r}$, Λ_j^A, $\sigma_{j,zr}^2$ et $\sigma_{j,\Lambda}^2$.

Remarque 4.5. Une bonne stratégie qui, par sa robustesse, devrait résister aux défaillances du modèle consiste à choisir un échantillon équilibré relativement aux variables auxiliaires $\mathbf{x}_j$. Dans ce cas, les variables auxiliaires $\mathbf{d}_j$ des équations d'équilibre sont remplacées par les variables augmentées $\mathbf{d}_j^* = \left(\mathbf{d}_j', \mathbf{x}_j' \big/ \pi_j^A\right)'$. Pour le calcul des variances, nous remplaçons les résidus $\eta_{j,v}$ par les résidus modifiés $\eta_{j,v}^* = y_{j,v} - \pi_j^A \left(\mathbf{d}_j^*\right)' \boldsymbol{\beta}_v^*$, où $\boldsymbol{\beta}_v^* = \left(\boldsymbol{\Delta}^*\right)^{-1} \sum_{j \in U^A} \pi_j^A \left(\frac{1}{\pi_j^A} - 1\right) \mathbf{d}_j^* y_{j,v}$ et $\boldsymbol{\Delta}^* = \sum_{j \in U^A} \mathbf{d}_j^* \left(\mathbf{d}_j^*\right)' \pi_j^A \left(1 - \pi_j^A\right)$. Nous employons les mêmes expressions pour les résidus modifiés $\eta_{j,r}^*$.

Remarque 4.6. Une répartition proportionnelle à la taille de la population peut représenter une stratégie raisonnable pour les plans d'échantillonnage stratifié où la taille totale d'échantillon m^A est fixe. Dans ce cas, nous pouvons définir la taille d'échantillon de strate, m_h^A, comme $m_h^A = m^A \left(\sum_{j \in U_h^A} x_j \big/ \sum_{j \in U^A} x_j\right)$, où x_j est la mesure de la taille.

5 Résultats empiriques

Nous obtenons les résultats ici présentés à l'aide de données réelles des districts 7, 8 et 9 de la province de Gaza au Mozambique. Cette source d'information résume les résultats empiriques d'une étude d'évaluation décrite dans FAO (2014). Nous avons pris d'autres résultats empiriques pour les stratégies proposées (FAO, 2015) avec la base de données de ménages agricoles du recensement général de l'agriculture du Burkina Faso pour confirmer les résultats généraux livrés ci-après.

Dans cette analyse utilisant les données du Mozambique, la population U^A se compose d'exploitations. La base de données d'expérimentation, qui est formée de variables environnementales et économiques, reçoit l'information du recensement de 2007 des exploitations grandes et moyennes et d'une enquête par sondage sur les petites exploitations (pour la même année). Le nombre total d'enregistrements est d'environ 36 890, dont 890 pour les exploitations grandes et moyennes.

La seconde population, U^B, est celle du recensement des ménages de 2007. Les enregistrements portent sur les particuliers s'adonnant à l'agriculture, la pêche ou l'activité forestière. La base compte quelque 54 000 enregistrements et comprend un certain nombre de variables sociodémographiques, environnementales et économiques. Nous avons fusionné les sources d'information des deux populations, créant ainsi une base de sondage principale (BSP) renfermant des liens artificiels entre particuliers et exploitations. Nous avons retenu les variables suivantes dans la procédure de fusionnement : type de travail et district de résidence pour les particuliers; secteur, district et nombre de travailleurs par type de travail pour les exploitations. Avant de fusionner, nous avons nettoyé U^B, retranchant les enregistrements où ne figurait pas la variable du type de travail (au nombre de 9 000 environ). Par la suite, nous avons établi un lien univoque exploitation-particulier pour les quelque 36 000 enregistrements de U^B relatifs à des exploitations n'employant pas de travailleurs. Nous avons rattaché le reste des particuliers aux 890 exploitations en appliquant les règles hiérarchiques suivantes :

- nous avons lié chaque exploitation à un nombre de particuliers correspondant au nombre de travailleurs selon le type de travail;
- nous avons lié les particuliers et les exploitations d'un même district;
- nous avons lié les particuliers aux exploitations privées ou publiques/gouvernementales là où le type de travail correspondait à la nature du secteur agricole.

Nous avons produit les liens au hasard en suivant les catégories définies par les règles hiérarchiques. Dans cet exercice, nous n'avons pas cherché à prédire les liens qui existent effectivement dans les deux populations, mais plutôt à créer un ensemble de données réalistes à des fins d'évaluation.

Les ensembles de données sur les deux populations comportent plusieurs variables, mais nous en avons retenu deux seulement dans notre étude. Pour U^A, nous prenons en compte le nombre d'animaux et, pour U^B, le nombre d'arbres. Le but est de mieux mettre en lumière l'incidence (sur le plan tant de l'exactitude que de la taille d'échantillon) que les différents contextes décrits à la section 4 peuvent avoir sur l'échantillon de la population U^B. Nous présentons au tableau 5.1 les statistiques sommaires de ces variables.

Tableau 5.1
Variables de la simulation avec des données du Mozambique

Population*	Nombre d'enregistrements*	Variable	Valeur moyenne	CV en %**
U^A : Exploitations	36 890***	Nombre d'animaux	11,1	681,6
U^B : Ménages	45 000	Nombre d'arbres	4,5	107,5

* Districts 7, 8 et 9 de la province de Gaza au Mozambique.
** CV en % = (écart-type de population/moyenne) × 100.
*** Sur ce nombre, 890 sont des exploitations grandes et moyennes.

Pour les deux populations, nous avons pris comme domaines d'intérêt les districts (3 domaines) et la province (1 domaine). Nous regardons donc au total huit totaux cibles d'intérêt (2 variables × 4 domaines).

5.1 Plans de sondage optimaux pour les différents contextes

Il sera question ici de quatre contextes :

Contexte 0. Absence de contrôle de l'échantillon de la population U^B. Dans cet échantillon planifié, il n'y a de contrôle que pour l'exactitude des estimations des variables des exploitations. Une fois l'échantillon de U^A prélevé, les unités de U^B liées aux unités de cet échantillon sont sélectionnées par le mécanisme d'échantillonnage indirect. Les CV en % prévus des estimations tirées de l'échantillon indirect des ménages sont alors calculés comme $\%\text{CV} = \left(\sqrt{\text{VA}\left(\hat{Y}\right)}\Big/Y\right)\times 100$.

Contexte 1. Il existe des bases de sondage pour les deux populations. Tous les liens sont connus et un plan d'échantillonnage intégré est utilisé dans la recherche d'une solution optimale compte tenu des deux. Il y a donc échantillonnage à plusieurs variables avec contrôle de l'exactitude des estimations tant de l'échantillon direct d'exploitations que de l'échantillon indirect de particuliers.

Contexte 2. Des bases de sondage existent pour les deux populations, mais les liens sont des probabilités estimées et un plan d'échantillonnage intégré est utilisé.

Contexte 3. Une seule base de sondage existe, celle de la population U^A. Nous étudions un plan d'échantillonnage intégré avec les options 3.1 et 3.2 représentant les solutions les plus praticables dans un contexte réel.

Les contextes 1, 2 et 3 sont ceux que nous avons définis à la section 4. Nous introduisons le contexte 0 qui constitue un moyen utile d'évaluation de la stratégie d'intégration.

Nous posons l'existence d'un mécanisme d'échantillonnage stratifié pour la première population U^A, où les strates U_h^A sont définies comme districts (7, 8 et 9) par catégorie de taille (1, 2, 3-4, 5-9, 10-19, 20-49, 50-99, 100+) selon le nombre de travailleurs agricoles, ce qui donne 21 strates. Pour ce qui est des modèles en (4.1), nous considérons des modèles de strate en moyenne avec $\tilde{y}_{j,v} = \tilde{y}_{v,(h)}$ et $\sigma^2_{j,v} = \sigma^2_{v,(h)}$ pour $j \in U_h^A$. Ces spécifications mènent à un plan type EASSSR pour les exploitations où les strates

coïncident avec les domaines planifiés (voir Falorsi et Righi (2015), remarque 4.2). Pour l'évaluation, nous appliquons la formule exacte de la variance d'un plan EASSSR au lieu d'approcher cette variance comme à la section 2, mais les deux expressions sont en gros équivalentes. Pour U^B, nous considérons aussi un modèle en moyenne défini au niveau du district d avec $\tilde{y}_{i,r} = \tilde{y}_{r,(d)}$, et $\sigma^2_{i,r} = \sigma^2_{r,(d)}$ pour $i \in U^B_d$.

Dans les études d'évaluation, on utilise le logiciel qui, conçu dans le langage *R*, met en œuvre l'échantillonnage optimal des plans types EASSSR, tout comme des plans de sondage plus généraux (plans équilibrés et plans à stratification incomplète). Ce logiciel est disponible à l'adresse http://www.istat.it/en/tools/methods-and-it-tools/design-tools/multiwaysampleallocation. Une fois installé, il s'accompagne d'un guide complet de l'utilisateur en anglais. Un autre logiciel qui existe seulement pour les plans EASSSR est le « MAUSS-R » disponible à l'adresse http://www.istat.it/it/strumenti/metodi-e-software/software/mauss-rdownload.

Pour chaque contexte, nous exprimons les contraintes de variance en coefficients de variation en pourcentage. Les analyses de la présente section se rapportent aux contextes, et nous employons une version simplifiée des fonctions de coût. Le coût c_j d'observation de l'unité j dans la population U^A avec les unités liées de la population U^B est fixe et égal à l'unité. Nous présentons à la section 5.2 des analyses plus détaillées sur les coûts.

On pourra se reporter à d'autres spécifications illustrées au tableau 5.2 pour chaque contexte.

Contexte 0. Les contraintes de variance sont fixées (seulement pour les estimations des exploitations en nombre d'animaux) à 6,5 % au niveau de la province et à 10 % au niveau des districts, d'où un échantillon de 2 122 exploitations.

Contexte 1. Nous avons fixé les contraintes des estimations des exploitations et des ménages pour constituer un échantillon d'environ 2 100 exploitations. Nous avons arrêté par là les contraintes de variance des estimations des exploitations avec 10 % pour les animaux au niveau de la province et 15 % au niveau des districts. Pour les estimations des ménages, ces mêmes contraintes sont de 2,5 % au niveau de la province et de 5 % au niveau des districts. À noter que ce choix de contraintes rend possible une comparaison des deux contextes avec en gros la même taille d'échantillon, bien que, dans le cas de la population U^A, les contraintes de variance des estimations soient plus grandes pour le contexte 1 que pour le contexte 0.

Contexte 2. Les contraintes de CV correspondent à celles du contexte 1 pour les estimations des ménages et des exploitations. Nous planifions l'observation intégrée à l'étape du plan de sondage en tenant compte de l'incertitude des liens, ce que nous faisons en regardant un modèle simplifié où nous posons que, pour chaque travailleur d'une exploitation, il n'existe qu'un lien *fort* (de valeur ψ) avec un particulier de la population de ménages et α liens faibles (de valeur τ) avec d'autres particuliers du même district, et où ψ et τ sont des probabilités avec $\psi \gg \tau$. Soit $l_{j\omega,ik}$ le lien entre le travailleur ω de l'exploitation j et le particulier k du ménage i. Posons que ces liens suivent un modèle de Bernoulli M_l, où

$$E_{M_l}\left(l_{j\omega,ik}\right) = \lambda_{j\omega,ik} = \begin{cases} \psi \text{ pour un seul travailleur } j\omega \in U^A \text{ et un particulier } ik \in U^B \\ \tau \text{ pour un seul travailleur } j\omega \in U^A \text{ et } \alpha \text{ particuliers } ik \in U^B \end{cases}, \quad (5.1)$$

où $\tau = \frac{1-\psi}{\alpha}$.

Dans cette simulation, nous avons envisagé différentes combinaisons des valeurs pour les probabilités de lien fort ψ ou de lien faible τ, ainsi que pour le nombre α de particuliers en cas de lien faible. Le tableau 5.3 illustre ces combinaisons.

Contexte 3. Les contraintes de CV pour les estimations des ménages et des exploitations sont celles du contexte 1. Au tableau 5.3, nous avons calculé la répartition avec l'option 3.2 comme elle est proposée pour le contexte 3. Nous présentons les résultats de l'option 3.1 à la fin de cette section.

Il convient enfin de noter que, pour les trois contextes, nous formulons le problème d'optimisation sous la forme π_j^A. Dans le cas d'un plan de sondage EASSSR, on pourrait y voir un problème de répartition en échantillonnage stratifié.

Tableau 5.2
Contraintes de variance dans les différents contextes

Contextes	Contraintes de variance*			
	U^A : variable des animaux		U^B : variable des arbres	
	Province	District	Province	District
Contexte 0	6,5 %	10 %	Contraintes nulles	Contraintes nulles
Contexte 1	10 %	15 %	2,5 %	5 %
Contexte 2	10 %	15 %	2,5 %	5 %
Contexte 3	10 %	15 %	2,5 %	5 %

* Contraintes exprimées en CV en %.

Tableau 5.3
Principaux résultats de l'évaluation

Contextes		Taille d'échantillon	Coefficient de variation réalisé (en %)						
			U^A : variable des animaux				U^B : variable des arbres		
			Province	District			Province	District	
				7	8	9		7	8	9
Contexte 0		2 122	6,5	10,0	10,0	10,0	1,5	6,8	12,7	1,4
Contexte 1		2 106	8,8	7,5	4,1	15,0	1,8	5,0	5,0	2,0
Contexte 2	$\psi = 0{,}90;\ \tau = 0{,}10;\ \alpha = 1$	2 146	8,8	7,2	4,1	15,0	2,2	5,0	5,0	2,4
	$\psi = 0{,}50;\ \tau = 0{,}10;\ \alpha = 5$	2 573	7,5	6,5	4,0	12,7	2,5	5,0	5,0	2,8
	$\psi = 0{,}30;\ \tau = 0{,}08;\ \alpha = 9$	2 767	7,0	6,4	4,0	11,9	2,5	5,0	5,0	2,8
	$\psi = 0{,}10;\ \tau = 0{,}09;\ \alpha = 9$	2 826	6,9	6,2	4,0	11,6	2,5	5,0	5,0	2,8
Contexte 3	Option 3.2	2 936	6,6	6,2	3,9	11,2	2,5	5,0	5,0	2,8

Les constatations suivantes ressortent de l'examen des principaux résultats de l'évaluation au tableau 5.3 :

Contexte 0 et contexte 1. Dans ces deux contextes, la taille de l'échantillon des exploitations est d'environ 2 100 fermes.

- Dans le contexte 0, le CV en % espéré des estimations des exploitations au niveau du district correspond exactement au niveau de contrainte de 10 % défini pour ce contexte.
- Dans le contexte 1, nous pouvons voir que tous les CV en % des estimations des exploitations au niveau du district respectent la contrainte de 15 % (définie pour ce contexte), mais ils deviennent de bien moins de 10 % pour les districts 7 et 8, indice que ces districts sont quelque peu suréchantillonnés par rapport à la précision visée, et ce, parce que, dans la deuxième répartition, il faut qu'une partie de l'échantillon d'exploitations réalise l'échantillon indirect requis de ménages (ce problème d'inefficacité est étudié plus en détail dans FAO (2014)).
- Si nous considérons maintenant la précision des estimations de la population U^B, nous constatons que les tailles espérées d'échantillon des ménages sont d'environ 5 300 enregistrements dans les deux contextes. Dans le contexte 0, les CV en % sont bien supérieurs au niveau visé de 5 % et dépassent même les 12 % dans le district 8. Avec la répartition de l'échantillon dans le contexte 1, la précision visée des estimations de la population U^B est toujours respectée, tout comme celle des estimations de la population U^A, bien que les contraintes de ces estimations reçoivent une définition plus large (valeurs supérieures) que celle des estimations du contexte 0.
- Ainsi, l'échantillonnage intégré dans le contexte 1 permet de contrôler la précision des estimations des deux populations d'intérêt, mais au prix d'une certaine perte de précision pour les estimations de la population U^A.

Contexte 1 et contexte 2. Si nous comparons les contextes 1 et 2, l'analyse porte sur les tailles globales d'échantillon, puisque les CV en % sont inférieurs aux niveaux de contrainte de ces deux contextes.

- En cas de présence de liens forts pour le contexte 2 ($\psi = 0{,}90$; $\tau = 0{,}10$; $\alpha = 1$), les tailles d'échantillon augmentent un peu seulement (40 exploitations), et les CV restent en deçà du niveau visé de précision, bien qu'étant légèrement en hausse pour les estimations des ménages.
- À mesure que les liens sont faibles, les tailles d'échantillon sont en hausse significative, ce que l'on doit à la réalisation des CV en % prévus pour les estimations des ménages.
- Dans le contexte 2 à l'inverse, les CV prévus des estimations des exploitations sont inférieurs aux niveaux visés, d'où l'impression que les exploitations sont quelque peu suréchantillonnées par rapport aux niveaux visés de précision.

Contexte 3 et autres contextes. Le contexte 3 peut être considéré comme un cas extrême du contexte 2 étant donné l'option 3.2 au tableau 5.3. Même dans ce cas, l'analyse porte sur les tailles globales d'échantillon, car tous les CV en % se situent sous les niveaux de contrainte :

- La maximisation de l'incertitude des liens, représentée par l'option 3.2, majore la taille d'échantillon d'environ 30 %; nous passons ainsi de 2 106 à 2 936.

- Si nous examinons le contexte 2, nous notons que nos résultats sont semblables à ceux du contexte 3 quand le niveau, ψ, du lien fort est d'environ 10 %.
- Même dans ce cas, les exploitations sont quelque peu suréchantillonnées par rapport aux niveaux de précision visés.

Analyse plus détaillée du contexte 3. Voici des analyses quelque peu plus détaillées par lesquelles nous voulons clarifier certains aspects du problème de la répartition d'échantillon pour l'observation intégrée de deux populations liées. Nous examinons l'option 3.1 et la répartition proportionnelle proposée à la remarque 4.6 en raison de leur importance pratique. Pour cette répartition, nous avons considéré comme mesure de taille (voir la remarque 4.6) le nombre total de travailleurs. Les $\tilde{z}_{j,r}$ s'obtiennent par l'expression en (4.10). Dans ce contexte, nous avons à définir la valeur k_r. Pour dégager une valeur unique k_r nous avons exploité les données du contexte 1 et calculé d'abord pour chaque strate le coefficient de variation de $z_{j,r}$, $\mathrm{CV}\left(z_{h,r}\right)$. Nous avons ensuite calculé les valeurs spécifiques k_r au niveau de la strate comme $k_{hr} = 1 + \left[\mathrm{CV}\left(z_{h,r}\right)\right]^2$. Nous avons enfin dégagé la valeur k_r considérée dans cette évaluation comme une moyenne pondérée des valeurs k_{hr} : $k_r = \sum_h k_{h,r} w_h$. Nous avons calculé les poids w_h de deux manières, ce qui nous a donné deux valeurs, soit $k_r = 2{,}75$ et $k_r = 2{,}16$. Nous avons défini les w_h, d'une part, proportionnellement à la somme des poids L_j^A au niveau de la strate et, d'autre part, proportionnellement à la quantité $\sqrt{\mathrm{CV}\left(z_{h,r}\right)}\,\bar{Y}_{r,h}^B\, N_h^A$, où $\bar{Y}_{r,h}^B$ et N_h^A sont respectivement la valeur moyenne de la variable y_r et le nombre d'unités de la strate. Dans chaque cas, nous avons exécuté le problème en (4.12) avec les contraintes définies au tableau 5.2 pour le contexte 1, obtenant ainsi une taille globale d'échantillon, n^A, correspondant respectivement à 1 639 et 1 517. Le tableau 5.4 illustre les principaux résultats de cette expérience. Pour les deux cas, nous indiquons (i) les CV espérés en % comme solution du problème en (4.12) dans l'hypothèse que la relation (4.11) se vérifie, (ii) les *vrais* CV espérés en %, c'est-à-dire obtenus dans le contexte 1 en fonction des tailles d'échantillon de strate définies par la solution du problème en (4.12) et (iii) les vrais CV en % obtenus dans le contexte 1 avec la répartition proportionnelle proposée à la remarque 4.6.

Tableau 5.4
CV espérés en % et réalisés des estimations de domaine du nombre total d'arbres avec la répartition d'échantillon obtenue comme solution du problème (4.12) et avec la répartition proportionnelle

Domaines d'estimation	$k_r = 2{,}75;\ n^A = 1\,639$			$k_r = 2{,}16;\ n^A = 1\,517$		
	CV en % espérés, obtenus comme solution du problème (4.12) en supposant que (4.11) se vérifie	vrais CV en % espérés dans le contexte 1 avec la répartition définie par (4.12)	vrais CV en % espérés dans le contexte 1 avec la répartition proportionnelle	CV en % espérés, obtenus comme solution du problème (4.12) en supposant que (4.11) se vérifie	vrais CV en % espérés dans le contexte 1 avec la répartition définie par (4.12)	vrais CV en % espérés dans le contexte 1 avec la répartition proportionnelle
Province	2,11	1,94	1,76	2,11	2,04	1,83
District 7	4,95	6,80	6,10	4,95	8,20	6,34
District 8	4,99	6,45	13,23	4,99	6,45	13,79
District 9	2,36	2,0	1,81	2,36	2,0	1,88

Voici les principaux résultats de cette évaluation :

- La stratégie de l'option 3.1 paraît efficace, permettant de contrôler les erreurs d'échantillonnage; on se trouve à éviter toute situation où ces erreurs dépassent largement le niveau visé d'exactitude pour les différents domaines d'estimation.
- Avec un k_r unique, les vrais CV espérés en % (colonnes 3 et 7 du tableau 5.4) pour certains domaines d'estimation sont supérieurs aux valeurs repères définies et, pour certains autres, les estimations sont d'une exactitude bien plus grande que ce qui est visé.
- Le choix d'une valeur supérieure du paramètre k_r semble être prudent, si la répartition de l'échantillon vise principalement à éviter les erreurs d'échantillonnage dans certains domaines aux valeurs trop grandes.
- Même si elle paraît efficace pour l'exactitude d'une estimation globale au niveau de la province, la répartition proportionnelle (colonnes 4 et 8 du tableau 5.4) ne permet pas de contrôler les écarts extrêmes par rapport au niveau espéré d'exactitude dans certains domaines d'estimation (voir le district 8).

5.2 Évaluation de coût

Dans cette évaluation, nous considérons le contexte 1 où nous disposons de bases de sondage pour les deux populations et où une observation intégrée des deux est possible. Nous mettons l'accent sur deux stratégies d'observation, la première avec deux échantillons indépendants, l'un pour les exploitations et l'autre pour les particuliers. Il est alors impossible de procéder à une analyse véritablement intégrée. Dans la seconde stratégie d'observation, nous appliquons un plan d'échantillonnage intégré avec un échantillon direct des exploitations et un échantillon indirect des ménages des travailleurs des exploitations échantillonnées.

Nous avons adopté les contraintes de variance établies pour le contexte 1 (voir le tableau 5.5).

Tableau 5.5
Contraintes de variance dans l'évaluation de coût

Contraintes de variance *			
U^A : variable des animaux		U^B : variable des arbres	
Province	**District**	**Province**	**District**
10 %	15 %	2,5 %	5 %

* Contraintes exprimées en CV en %.

Pour l'échantillon direct, nous avons adopté un plan EASSSR où la population U^A a été stratifiée par croisement des districts et des catégories de taille des exploitations et où la population U^B a été stratifiée par district. Le coût d'interview des exploitations est variable $(C^A = 1, 2, 5 \text{ et } 10)$, ce qui nous amène à effectuer quatre évaluations. Le coût C^B d'interview d'un particulier est fixé à l'unité.

Pour les plans d'échantillonnage indirect, nous définissons le coût global d'interview de l'exploitation et de ses travailleurs ensemble par deux spécifications distinctes de l'équation en (3.2) :

$$c_j = C^A + L_j^A C^B, \tag{5.2}$$

$$c_j = C^A + \sqrt{L_j^A}\, C^B. \tag{5.3}$$

La fonction de coût en (5.3) augmente moins que la fonction de coût en (5.2) lorsque L_j^A s'accroît.

Nous avons fait une répartition optimale sous contrainte de précision pour les deux plans de sondage indépendants. Les différentes valeurs C^A (1, 2, 5 et 10) n'influent pas sur la taille de l'échantillon des exploitations, le coût augmentant en proportion. Pour les contraintes de variance au tableau 5.5 avec la stratégie d'indépendance, les tailles des échantillons des exploitations et des particuliers sont de 1 010 et 3 388. Le coût total est alors de 4 398 si $C^A = 1$. Dans la stratégie d'intégration, le coût influe effectivement sur la répartition, essentiellement parce que si le coût d'interview de l'exploitation est en hausse, le nombre d'exploitations échantillonnées baisse et la répartition majore les tailles d'échantillon des strates des exploitations les plus grandes.

Le tableau 5.6 qui suit présente les tailles d'échantillon des exploitations et les tailles prévues d'échantillon des particuliers lorsque le modèle de coût en (5.2) sert à calculer le coût des interviews des particuliers en échantillonnage intégré. Nous pouvons voir que l'échantillon des exploitations est plus du double de la taille d'échantillon, à ne considérer que les seules fermes (1 101). Si la taille augmente, c'est à cause des contraintes de précision des estimations des ménages.

Tableau 5.6
Tailles d'échantillon pour la répartition intégrée lorsque le coût global d'interview d'un particulier est donné par (5.2)

Coût par interview d'exploitation (C^A)	1	2	5	10
Exploitations	2 388	2 289	2 190	2 137
Particuliers	4 504	4 491	4 862	4 905

Le tableau 5.7 qui suit indique la répartition quand l'équation (5.3) est appliquée au coût de l'interview d'un particulier en échantillonnage intégré.

Tableau 5.7
Tailles d'échantillon en échantillonnage intégré lorsque le coût global d'interview d'un particulier est donné par (5.3)

Coût par interview d'exploitation (C^A)	1	2	5	10
Exploitations	2 135	2 121	2 111	2 108
Particuliers	4 834	4 874	5 283	5 360

Les tableaux (5.6) et (5.7) indiquent que la taille d'échantillon intégré des exploitations est en gros du double de celle de l'échantillon indépendant correspondant. Ainsi, la variance prévue des estimations sera bien moindre que les contraintes de variance visées, d'où l'impression que l'échantillonnage intégré est surtout tributaire des contraintes de variance liées aux paramètres des particuliers à estimer.

Les figures 5.1 et 5.2 indiquent les coûts respectifs de l'échantillonnage indépendant et de l'échantillonnage intégré. La stratégie d'observation intégrée coûte généralement plus cher sauf là où le coût par interview d'exploitation correspond à l'unité et où la fonction de coût est donnée par (5.3). Dans cette évaluation, la nature intégrée de l'échantillon n'a pas à intervenir, car nous n'examinons aucun croisement des variables respectives de la population U^A et de la population U^B; dans ce cas, l'échantillonnage indépendant sera plus efficace en matière de précision. Une autre fonction de coût pourrait cependant amener en partie un nouvel équilibre entre les deux stratégies d'observation pour ce qui est du coût.

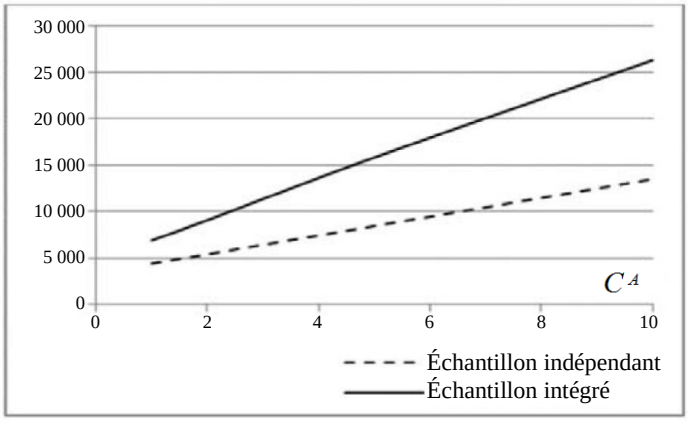

Figure 5.1 Coût global selon l'échantillonnage intégré et deux échantillonnages indépendants avec (5.2).

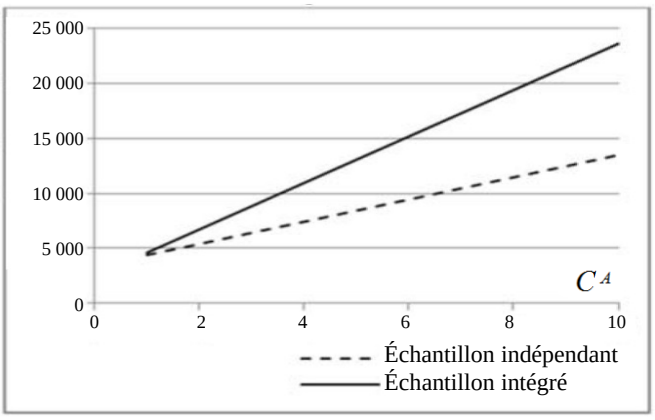

Figure 5.2 Coût global selon l'échantillonnage intégré et deux échantillonnages indépendants avec (5.3).

6 Conclusions

Nous avons étudié le problème de la définition de plans de sondage optimaux pour les stratégies d'enquête visant à l'observation intégrée de populations statistiques liées entre elles. Cette question présente un intérêt particulier dans le secteur agricole où l'observation intégrée permet de mesurer des phénomènes d'ensemble qui concernent, par exemple, des populations statistiques différentes d'exploitations et de ménages. Il y a observation intégrée lorsqu'on échantillonne directement la première population et observe indirectement la seconde en exploitant les liens qui existent entre leurs unités respectives. Nous avons étudié ce problème en nous attachant à trois contextes d'information sur les liens. Dans deux de ces contextes, l'information est très riche et, dans le troisième, très pauvre. Nous traitons l'incertitude dont peuvent être entachées les variables des deux populations, les liens et les variables z (issues du mécanisme d'échantillonnage indirect) en introduisant des modèles appropriés de superpopulation dont les valeurs espérées (du premier et du second ordre) sont considérées comme connues au lancement de l'algorithme de l'échantillonnage optimal. Nous avons soumis à des études empiriques les données réelles venant d'un pays en développement, le Mozambique.

Voici en résumé les grandes conclusions tirées :

Observation intégrée et observation indépendante. L'observation intégrée est essentielle à une mesure méthodique de phénomènes d'ensemble influant sur des populations différentes. Son grand avantage est de permettre le croisement des variables de la population U^A et de la population U^B. Elle se révèle en outre nécessaire lorsqu'il n'existe pas de base de sondage pour la population U^B et qu'on a besoin d'un mécanisme d'échantillonnage indirect. Tel est le cas examiné dans le contexte 3. Il reste que, dans les contextes 1 et 2, l'échantillonnage indépendant sera plus efficace si on se contente d'examiner des agrégats indépendamment l'un de l'autre pour les deux populations.

Questions de coût. Il est possible de réduire la perte d'efficacité de l'observation intégrée si, comme nous le posons avec la fonction de coût en (5.3), le coût moyen d'observation de l'unité élémentaire de U^B décroît à mesure qu'augmente la taille des grappes indirectement observées. Dans ce cas, l'échantillonnage intégré et les deux échantillonnages indépendants pourraient être d'un rendement plus proche ou analogue comme dans l'étude d'évaluation. Il est néanmoins complexe d'établir quelle relation entre C^A et C^B conduit à deux stratégies au coût semblable, car les répartitions en question dépendent non seulement du coût de l'interview, mais aussi de la variabilité des paramètres cibles des deux populations et du jeu de contraintes de variance.

Contrôle des erreurs à l'étape du plan de sondage. L'échantillonnage intégré permet de contrôler les CV des estimations des populations intégrées, sinon les CV de la population indirectement observée pourraient être très élevés.

Incidence de l'incertitude sur les tailles d'échantillon. Une hausse des variances du modèle (sur les variables ou les liens) se traduit par une augmentation significative des tailles d'échantillon, d'où la nécessité de disposer de bons modèles de prédiction des variables ou des liens inconnus.

Annexe A

Pour obtenir l'espérance du modèle $E_{M_v}\left(\eta_{j,v}^2\right)$, soit $\boldsymbol{\eta}_v = \{\eta_{j,v}\}$ le vecteur de taille M^A des résidus, où

$$\boldsymbol{\eta}_v = \mathbf{Y}_v - \boldsymbol{\Pi}\mathbf{D}\boldsymbol{\Delta}^{-1}\mathbf{D}'(\mathbf{I}-\boldsymbol{\Pi})\,\mathbf{Y}_v, \tag{A.1}$$

où $\mathbf{Y}_v = \{y_{j,v}\}$ désigne le vecteur de taille M^A avec les valeurs de la v^{e} variable d'intérêt et où $\boldsymbol{\Pi} = \text{diag}\{\pi_j^A\}$ indique la matrice diagonale avec les M^A probabilités d'inclusion. D'après le modèle en (4.1), le vecteur $\mathbf{Y}_v$ peut s'exprimer comme $\mathbf{Y}_v = \tilde{\mathbf{Y}}_v + \mathbf{u}_v$, où $\tilde{\mathbf{Y}}_v = \{\tilde{y}_{i,v}\}$ et $\mathbf{u}_v = \{u_{i,v}\}$ désigne les vecteurs de taille M^A des prédictions et des résidus du modèle. Si nous adoptons la notation matricielle qui précède, nous pouvons exprimer les résidus spécifiques $\eta_{j,v}$ comme $\eta_{j,v} = \left(\tilde{y}_{j,v} + u_{j,v}\right) - \pi_j\mathbf{d}_j'\boldsymbol{\Delta}^{-1}\mathbf{D}'(\mathbf{I}-\boldsymbol{\Pi})\left(\tilde{\mathbf{Y}}_v + \mathbf{u}_v\right)$. Ainsi, les valeurs prévues du modèle pour les termes au carré sont données par :

$$\begin{aligned} E_{M_v}\left(\eta_{j,v}^2\right) = \tilde{y}_{j,v}^2 + \sigma_{j,v}^2 \quad &- 2\pi_j\mathbf{d}_j'\boldsymbol{\Delta}^{-1}\mathbf{D}'(\mathbf{I}-\boldsymbol{\Pi})\,\tilde{\mathbf{Y}}_v - 2\pi_j\mathbf{d}_j'\boldsymbol{\Delta}^{-1}\mathbf{D}'(\mathbf{I}-\boldsymbol{\Pi})\,\mathbf{0}_{j\sigma_v} \\ &+ \pi_j^2\tilde{\mathbf{Y}}_v'(\mathbf{I}-\boldsymbol{\Pi})\,\mathbf{D}\boldsymbol{\Delta}^{-1}\mathbf{d}_j\mathbf{d}_j'\boldsymbol{\Delta}^{-1}\mathbf{D}'(\mathbf{I}-\boldsymbol{\Pi})\,\tilde{\mathbf{Y}}_v \\ &+ \pi_j^2\boldsymbol{\sigma}_v'(\mathbf{I}-\boldsymbol{\Pi})\,\mathbf{D}\boldsymbol{\Delta}^{-1}\mathbf{d}_j\mathbf{d}_j'\boldsymbol{\Delta}^{-1}\mathbf{D}'(\mathbf{I}-\boldsymbol{\Pi})\,\boldsymbol{\sigma}_v \end{aligned} \tag{A.2}$$

où $\boldsymbol{\sigma}_v = \{\sigma_{j,v}\}$ est le vecteur colonne de taille M^A des erreurs-types du modèle pour les variables V et où $\mathbf{0}_{j\sigma_v} = \left(0, \ldots, \sigma_{j,v}^2, \ldots, 0\right)'$ est un vecteur où le j^{e} élément correspond à $\sigma_{j,v}^2$, tous les autres éléments étant des zéros. Avec la notation matricielle qui précède et en suivant Falorsi et Righi (2015), nous pouvons approcher la variance anticipée par l'expression suivante :

$$E_{M_v}\left[V\left(\hat{Y}_v^A \,\middle|\, \mathbf{m}^A\right)\right] = \left[M^A/(M^A - H)\right]\left[\tilde{\mathbf{Y}}_v'\boldsymbol{\Pi}^{-1}\tilde{\mathbf{Y}}_v + \boldsymbol{\sigma}_v'\boldsymbol{\Pi}^{-1}\boldsymbol{\sigma}_v - \tilde{\mathbf{Y}}_v'\tilde{\mathbf{Y}}_v + \boldsymbol{\sigma}_v'\boldsymbol{\sigma}_v + \text{AVA}_{3,v}\right].$$

Soit $\mathbf{a}_v = \mathbf{D}\boldsymbol{\Delta}^{-1}\mathbf{D}'(\mathbf{I}-\boldsymbol{\Pi})\tilde{\mathbf{Y}}_v$, $\mathbf{b}_v = \boldsymbol{\sigma}_v'\mathbf{D}\boldsymbol{\Delta}^{-1}\mathbf{D}'(\mathbf{I}-\boldsymbol{\Pi})\,\boldsymbol{\sigma}_v$ et $\mathbf{c}_v = \boldsymbol{\sigma}_v'\text{diag}\left[\mathbf{D}\boldsymbol{\Delta}^{-1}\mathbf{D}'(\mathbf{I}-\boldsymbol{\Pi})'(\mathbf{I}-\boldsymbol{\Pi})\,\mathbf{D}\boldsymbol{\Delta}^{-1}\mathbf{D}'\right]\boldsymbol{\sigma}_v$. Nous avons alors $\text{AVA}_{3,v} = \mathbf{a}_v'(\mathbf{I}-\boldsymbol{\Pi})\left(2\tilde{\mathbf{Y}}_v - \boldsymbol{\Pi}\mathbf{a}_v\right) + \mathbf{1}'(\mathbf{I}-\boldsymbol{\Pi})\left(2\mathbf{b}_v - \boldsymbol{\Pi}\mathbf{c}_v\right)$, où les scalaires définis en (A.1.4), (A.1.7) et (A.1.8) dans Falorsi et Righi (2015) sont respectivement les éléments des vecteurs $\mathbf{a}_v$, $\mathbf{b}_v$ et $\mathbf{c}_v$.

Annexe B

Si nous adoptons la notation matricielle, nous pouvons exprimer les résidus $\eta_{j,r}$ comme

$$\eta_{j,r} = \mathbf{l}_j'\left(\tilde{\mathbf{Y}}_r + \mathbf{u}_r\right) - \pi_j\boldsymbol{\delta}_j'\boldsymbol{\Delta}^{-1}\mathbf{D}'(\mathbf{I}-\boldsymbol{\Pi})\,\mathbf{L}\left(\tilde{\mathbf{Y}}_r + \mathbf{u}_r\right), \tag{B.1}$$

où $\mathbf{L} = \left\{\tilde{L}_{j,i}^B\right\}$ est la matrice de taille $M^A \times N^B$ des liens centrés réduits, où $\tilde{\mathbf{Y}}_r = \{y_{i,r}\}$ et $\mathbf{u}_r = \{u_{i,r}\}$ désignent respectivement les vecteurs de taille N^B avec les valeurs des prédictions et des résidus de la r^{e} variable d'intérêt et où $\mathbf{l}_j'$ est la j^{e} ligne de la matrice $\mathbf{L}$. Ainsi, les valeurs espérées du modèle pour les termes au carré sont données par :

$$
\begin{aligned}
E_{M_r}\left(\eta_{j,r}^2\right) &= \tilde{\mathbf{Y}}_r'\mathbf{l}_j\mathbf{l}_j'\tilde{\mathbf{Y}}_r + \boldsymbol{\sigma}_r'\mathbf{l}_j\mathbf{l}_j'\boldsymbol{\sigma}_r \\
&\quad - 2\pi_j\tilde{\mathbf{Y}}_r'\mathbf{d}_j\boldsymbol{\Delta}^{-1}\mathbf{D}'(\mathbf{I}-\boldsymbol{\Pi})\,\mathbf{L}\tilde{\mathbf{Y}}_r - 2\pi_j\boldsymbol{\sigma}_r'\mathbf{d}_j\boldsymbol{\Delta}^{-1}\mathbf{D}'(\mathbf{I}-\boldsymbol{\Pi})\,\mathbf{L}\boldsymbol{\sigma}_r \\
&\quad + \pi_j^2\tilde{\mathbf{Y}}_r'\mathbf{L}'(\mathbf{I}-\boldsymbol{\Pi})\,\mathbf{D}\boldsymbol{\Delta}^{-1}\mathbf{d}_j\mathbf{d}_j'\boldsymbol{\Delta}^{-1}\mathbf{D}'(\mathbf{I}-\boldsymbol{\Pi})\,\mathbf{L}\tilde{\mathbf{Y}}_r \\
&\quad + \pi_j^2\boldsymbol{\sigma}_r'\mathbf{L}'(\mathbf{I}-\boldsymbol{\Pi})\,\mathbf{D}\boldsymbol{\Delta}^{-1}\mathbf{d}_j\mathbf{d}_j'\boldsymbol{\Delta}^{-1}\mathbf{D}'(\mathbf{I}-\boldsymbol{\Pi})\,\mathbf{L}\boldsymbol{\sigma}_r \qquad \text{(B.2)}
\end{aligned}
$$

où $\boldsymbol{\sigma}_r = \{\sigma_{j,r}\}$ est le vecteur colonne N^B des erreurs-types du modèle pour les variables y_r. Suivant la même notation, nous avons :

$$
E_{M_v}\left[V\left(\hat{Y}_r^A \,\middle|\, \mathbf{m}^A\right)\right] = \left[M^A/(M^A-H)\right]\left[\tilde{\mathbf{Y}}_r'\mathbf{L}'\boldsymbol{\Pi}^{-1}\mathbf{L}\tilde{\mathbf{Y}}_r + \boldsymbol{\sigma}_r'\mathbf{L}'\boldsymbol{\Pi}^{-1}\mathbf{L}\boldsymbol{\sigma}_r - \tilde{\mathbf{Y}}_r'\mathbf{L}'\mathbf{L}\tilde{\mathbf{Y}}_r + \boldsymbol{\sigma}_r'\mathbf{L}'\mathbf{L}\boldsymbol{\sigma}_r + \mathrm{AVA}_{3,r}\right].
$$

Soit $\mathbf{a}_r = \mathbf{D}\boldsymbol{\Delta}^{-1}\mathbf{D}'(\mathbf{I}-\boldsymbol{\Pi})\,\mathbf{L}\tilde{\mathbf{Y}}_r$, $\mathbf{b}_r = \boldsymbol{\sigma}_r'\mathbf{D}\boldsymbol{\Delta}^{-1}\mathbf{D}'(\mathbf{I}-\boldsymbol{\Pi})\,\boldsymbol{\sigma}_r$, $\mathbf{c}_r = \boldsymbol{\sigma}_r'\mathbf{D}\boldsymbol{\Delta}^{-1}\mathbf{D}'(\mathbf{I}-\boldsymbol{\Pi})'(\mathbf{I}-\boldsymbol{\Pi})\,\mathbf{D}\boldsymbol{\Delta}^{-1}\mathbf{D}'\boldsymbol{\sigma}_r$. Nous avons alors $\mathrm{AVA}_{3,r} = \mathbf{a}_r'(\mathbf{I}-\boldsymbol{\Pi})\left(2\mathbf{L}\tilde{\mathbf{Y}}_r - \boldsymbol{\Pi}\mathbf{a}_r\right) + \mathbf{1}'(\mathbf{I}-\boldsymbol{\Pi})\left(2\mathbf{b}_r - \boldsymbol{\Pi}\mathbf{c}_r\right)$.

Enfin, nous obtenons ce qui suit :

$$
\begin{aligned}
E_{M_r}\left(z_{j,r}^2\right) &= \sum_{i=1}^{N^B}\left(\tilde{L}_{j,i}^B\right)^2\left(\tilde{y}_{i,r}^2 + \sigma_{i,r}^2\right) + \sum_{i=1}^{N^B}\tilde{L}_{j,i}^B\,\tilde{y}_{i,r}\sum_{i'\neq i}\tilde{L}_{j,i'}^B\,\tilde{y}_{i',r} \\
&= \sum_{i=1}^{N^B}\left(\tilde{L}_{j,i}^B\right)^2\left(\tilde{y}_{i,r}^2 + \sigma_{i,r}^2\right) + \sum_{i=1}^{N^B}\tilde{L}_{j,i}^B\,\tilde{y}_{i,r}\left(\tilde{z}_{j,r} - \tilde{L}_{j,i}^B\,\tilde{y}_{i,r}\right) \\
&= \sum_{i=1}^{N^B}\left(\tilde{L}_{j,i}^B\right)^2\left(\tilde{y}_{i,r}^2 + \sigma_{i,r}^2\right) + \tilde{z}_{j,r}^2 - \sum_{i=1}^{N^B}\left(\tilde{L}_{j,i}^B\right)^2\tilde{y}_{i,r}^2 \\
&= \tilde{z}_{j,r}^2 + \sigma_{j,zr}^2.
\end{aligned}
$$

Annexe C

À partir de (B.2), nous obtenons :

$$
\begin{aligned}
E_{M_l}E_{M_r}\left(\eta_{j,r}^2\right) &= \tilde{\mathbf{Y}}_r'E_{M_l}\left(\mathbf{l}_j\mathbf{l}_j'\right)\tilde{\mathbf{Y}}_r + \boldsymbol{\sigma}_r'E_{M_l}\left(\mathbf{l}_j\mathbf{l}_j'\right)\boldsymbol{\sigma}_r \\
&\quad - 2\pi_j\tilde{\mathbf{Y}}_r'E_{M_l}\left(\mathbf{l}_j\mathbf{d}_j'\boldsymbol{\Delta}^{-1}\mathbf{D}'(\mathbf{I}-\boldsymbol{\Pi})\,\mathbf{L}\right)\tilde{\mathbf{Y}}_r \\
&\quad - 2\pi_j\boldsymbol{\sigma}_r'E_{M_l}\left(\mathbf{l}_j\mathbf{d}_j'\boldsymbol{\Delta}^{-1}\mathbf{D}'(\mathbf{I}-\boldsymbol{\Pi})\,\mathbf{L}\right)\boldsymbol{\sigma}_r \\
&\quad + \pi_j^2\tilde{\mathbf{Y}}_r'E_{M_l}\left(\mathbf{L}'(\mathbf{I}-\boldsymbol{\Pi})\,\mathbf{D}\boldsymbol{\Delta}^{-1}\mathbf{d}_j\mathbf{d}_j'\boldsymbol{\Delta}^{-1}\mathbf{D}'(\mathbf{I}-\boldsymbol{\Pi})\,\mathbf{L}\right)\tilde{\mathbf{Y}}_r \\
&\quad + \pi_j^2\boldsymbol{\sigma}_r'E_{M_l}\left(\mathbf{L}'(\mathbf{I}-\boldsymbol{\Pi})\,\mathbf{D}\boldsymbol{\Delta}^{-1}\mathbf{d}_j\mathbf{d}_j'\boldsymbol{\Delta}^{-1}\mathbf{D}'(\mathbf{I}-\boldsymbol{\Pi})\,\mathbf{L}\right)\boldsymbol{\sigma}_r. \qquad \text{(C.1)}
\end{aligned}
$$

Nous pouvons aisément calculer la valeur prévue qui précède en fonction du résultat général qui suit. Soit $\mathbf{A} = \{a_{j,j'}\}$ une matrice générique de taille $M^A \times M^A$. L'élément générique $g_{i,i'}$ à la position i, i' de la matrice carrée de taille $N^B \times N^B$ $\mathbf{L}'\mathbf{A}\mathbf{L} = \{g_{i,i'}\}$ est donné par $g_{i,i'} = \sum_{j=1}^{M^A}\sum_{j'=1}^{M^A}\tilde{L}_{j,i}^B\,\tilde{L}_{j',i'}^B\,a_{j,j'}$.

Nous avons $E_{M_l}\left(g_{i,i'}\right) = \sum_{j=1}^{M^A}\sum_{j'=1}^{M^A}\tilde{\Lambda}_{j,i}^B\,\tilde{\Lambda}_{j',i'}^B\,a_{j,j'} + \sum_{j=1}^{M^A}\sum_{j'=1}^{M^A}\mathrm{Cov}_{M_l}(\tilde{L}_{j,i}^B, \tilde{L}_{j',i'}^B)\;a_{j,j'}$.

L'approximation du premier ordre en série de Taylor de $\tilde{\Lambda}_{j,i}$ est donnée par $\tilde{L}_{j,i}^B \cong \frac{1}{\Lambda_i}\left(L_{j,i}^B - \tilde{\Lambda}_{j,i}^B\,L_i^B\right)$. Nous avons donc

$$\mathrm{Cov}_{M_l}(\tilde{L}^B_{j,i},\,\tilde{L}^B_{j',i'}) \cong \frac{1}{\Lambda^B_i}\frac{1}{\Lambda^B_{i'}}\mathrm{Cov}_{M_l}\left[\left(L^B_{j,i}-\tilde{\Lambda}^B_{j,i}L^B_i\right),\left(L^B_{j',i'}-\tilde{\Lambda}^B_{j',i'}L^B_{i'}\right)\right]$$

$$= \begin{cases} \left[V_{M_l}(L^B_{j,i})(1-2\tilde{\Lambda}^B_{j,i})+(\tilde{\Lambda}^B_{j,i})^2 V_{M_l}(L^B_i)\right]\Big/(\Lambda^B_i)^2 & \text{pour } j=j' \text{ et } i=i' \\ \left[\tilde{\Lambda}^B_{j,i}\tilde{\Lambda}^B_{j,i'}V_{M_l}(L^B_i)-\tilde{\Lambda}^B_{j,i}V_{M_l}(L^B_{j,i})-\tilde{\Lambda}^B_{j,i'}V_{M_l}(L^B_{j',i})\right]\Big/(\Lambda^B_i)^2 & \text{pour } j\neq j' \text{ et } i=i' \\ 0 & \text{pour } j=j' \text{ et } i\neq i' \\ 0 & \text{pour } j\neq j' \text{ et } i\neq i' \end{cases} \quad \text{(C.2)}$$

où

$$V_{M_l}\left(L^B_{j,i}\right)=\sum_{k=1}^{M^b_l}\lambda_{j,ik}\left(1-\lambda_{j,ik}\right),\ V_{M_l}(L^B_i)=\sum_{j=1}^{M^A}V_{M_l}\left(L^B_{j,i}\right). \quad \text{(C.3)}$$

L'équation (C.2) est tirée du résultat suivant. Pour $i=i'$ et $j=j'$, nous obtenons

$$\begin{aligned}\mathrm{Cov}_{M_l}\left(\tilde{L}^B_{j,i},\,\tilde{L}^B_{j,i}\right) &= V_{M_l}\left(\tilde{L}^B_{j,i}\right) \\ &\cong \frac{1}{\Lambda^2_i}\left[V_{M_l}\left(L^B_{j,i}\right)+\left(\tilde{\Lambda}^B_{j,i}\right)^2 V_{M_l}(L^B_i)-2\tilde{\Lambda}^B_{j,i}\,\mathrm{Cov}_{M_l}\left(L^B_{j,i},\,L^B_i\right)\right] \\ &= \frac{1}{\Lambda^2_i}\left[V_{M_l}\left(L^B_{j,i}\right)\left(1-2\tilde{\Lambda}^B_{j,i}\right)+\left(\tilde{\Lambda}^B_{j,i}\right)^2 V_{M_l}(L^B_i)\right] \quad \text{(C.4)}\end{aligned}$$

où $\mathrm{Cov}_{M_l}\left(L^B_{j,i},\,L^B_i\right)=V_{M_l}\left(L^B_{j,i}\right)$. Pour $i=i'$ et $j\neq j'$, nous obtenons

$$\begin{aligned}\mathrm{Cov}_{M_l}\left(\tilde{L}^B_{j,i},\,\tilde{L}^B_{j',i}\right) &\cong \frac{1}{\Lambda^2_i}\mathrm{Cov}_{M_l}\left[\left(L^B_{j,i}-\tilde{\Lambda}^B_{j,i}\,L_i\right),\left(L^B_{j',i}-\tilde{\Lambda}^B_{j',i}L_i\right)\right] \\ &= \frac{1}{\Lambda^2_i}V_{M_l}\left[L^B_{j,i}L^B_{j',i}-\tilde{\Lambda}^B_{j',i}L^B_{j,i}L_i-\tilde{\Lambda}^B_{j,i}L^B_{j',i}L_i+\tilde{\Lambda}^B_{j,i}\,\tilde{\Lambda}^B_{j',i}\,L^2_i\right] \\ &= \frac{1}{\Lambda^2_i}\left[\tilde{\Lambda}^B_{j,i}\,\tilde{\Lambda}^B_{j',i}V_{M_l}(L^B_i)-\tilde{\Lambda}^B_{j,i}V_{M_l}\left(L^B_{j,i}\right)-\tilde{\Lambda}^B_{j',i}V_{M_l}\left(L^B_{j',i}\right)\right]. \quad \text{(C.5)}\end{aligned}$$

Soit $\mathbf{a}=\{a_j\}$ un vecteur générique de taille M^A. L'élément générique ${}^j g_{i,i'}$ à la position i, i' de la matrice carrée de taille $N^B\times N^B$ $\mathbf{l}_j\mathbf{a}'\mathbf{L}=\left\{{}^j g_{i,i'}\right\}$ est donné par ${}^j g_{i,i'}=\sum_{j'=1}^{M^A}\tilde{L}^B_{j,i}\,\tilde{L}^B_{j',i'}a_{j'}$ où $E_{M_l}\left({}^j g_{i,i'}\right)=\sum_{j'=1}^{M^A}\tilde{\Lambda}^B_{j,i}\,\tilde{\Lambda}^B_{j',i'}\,a_{j'}+\sum_{j'=1}^{M^A}\mathrm{Cov}_{M_l}\left(\tilde{L}^B_{j,i},\,\tilde{L}^B_{j',i'}\right)\,a_{j'}$.

Enfin, soit $\left\{{}^{jj} g_{i,i'}\right\}$ l'élément générique à la position i, i'^{e} de la matrice $\mathbf{l}_j\mathbf{l}'_j$. Sa valeur générique espérée est donnée par $E_{M_l}\left({}^{jj} g_{i,i'}\right)=\tilde{\Lambda}^B_{j,i}\,\tilde{\Lambda}^B_{j',i'}+\mathrm{Cov}_{M_l}\left(\tilde{L}^B_{j,i},\,\tilde{L}^B_{j',i'}\right)$.

Bibliographie

Bethel, J. (1989). Répartition de l'échantillon dans les enquêtes à plusieurs variables. *Techniques d'enquête*, 15, 1, 49-60. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/1989001/article/14578-fra.pdf.

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

Brewer, K.R.W., et Gregoire, T.G. (2009). Introduction to survey sampling. Dans *Handbook of Statistics – Sample Surveys: Design, Methods and Applications*, (Éds., D. Pfeffermann et C.R. Rao), Elsevier B.V. *29A*, 9-37.

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*, 38, 1, 25-32. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2012001/article/11682-fra.pdf.

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. (2005). Variance approximation under balanced sampling. *Journal of Statistical Planning and Inference*, 128, 569-591.

Falorsi, P.D., et Righi, P. (2015). 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. *Techniques d'enquête*, 41, 1, 225-247. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2015001/article/14149-fra.pdf.

FAO (2014). *Technical Report on the Integrated Survey Framework*, rapport technique série GO-02-2014. http://gsars.org/wp-content/uploads/2014/07/Technical_report_on-ISF-Final.pdf.

FAO (2015). *Guidelines on Integrated Survey Framework*. GUIDELINES & HANDBOOKS http://gsars.org/en/guidelines-for-the-integrated-survey-framework/. Consulté en août 2016.

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.

Lavallée, P. (2002). Le sondage indirect, ou la méthode du partage des poids. Éditions de l'Université de Bruxelles (Belgique) et Éditions Ellipses (France), 215 pages.

Lavallée, P. (2007). *Indirect Sampling*. New York: Springer.

Lavallée, P., et Caron, P. (2001). Estimation par la méthode généralisée du partage des poids : le cas du couplage d'enregistrements. *Techniques d'enquête*, 27, 2, 171-187. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2001002/article/6092-fra.pdf.

Lavallée, P., et Labelle-Blanchet, S. (2013). Le sondage indirect appliqué aux populations asymétriques. *Techniques d'enquête*, 39, 1, 207-241. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2013001/article/11829-fra.pdf.

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

Steel, D.G., et Clark, R.G. (2014). Gains possibles lors de l'utilisation de l'information sur les coûts au niveau de l'unité dans un cadre assisté par modèle. *Techniques d'enquête*, 40, 2, 257-269. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2014002/article/14110-fra.pdf.

Wallgren, A., et Wallgren, B. (2014). *Register-Based Statistics: Administrative Data for Statistical Purposes*. New York: John Wiley & Sons, Inc. Chichester, Royaume-Uni. ISBN: ISBN 978-1-119-94213-9.

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.

Xu, X., et Lavallée, P. (2009). Traitements de la non-réponse de lien dans l'échantillonnage indirect. *Techniques d'enquête*, 35, 2, 165-177. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2009002/article/11038-fra.pdf.