N° 12-001-X au catalogue
ISSN 1712-5685

Techniques d'enquête

Comment décomposer la variance due à la non-réponse : une méthode fondée sur l'erreur d'enquête totale

par Keven Bosa, Serge Godbout, Fraser Mills et Frédéric Picard

Date de diffusion : le 20 décembre 2018

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.

Signes conventionnels dans les tableaux

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

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

Publication autorisée par le ministre responsable de Statistique Canada

© Sa Majesté la Reine du chef du Canada, représentée par le ministre de l'Industrie 2018

Tous droits réservés. L'utilisation de la présente publication est assujettie aux modalités de l'entente de licence ouverte de Statistique Canada.

Une version HTML est aussi disponible.

This publication is also available in English.

Techniques d'enquête, décembre 2018
Vol. 44, nº 2, pp. 319-337
Statistique Canada, nº 12-001-X au catalogue

Comment décomposer la variance due à la non-réponse : une méthode fondée sur l'erreur d'enquête totale

Keven Bosa, Serge Godbout, Fraser Mills et Frédéric Picard[1]

Résumé

Lorsqu'une méthode d'imputation linéaire est utilisée pour corriger la non-réponse, et sous certaines hypothèses, on peut attribuer au niveau des unités non-répondantes la variance totale. L'imputation linéaire n'est pas aussi restrictive qu'il n'y paraît car les méthodes les plus populaires comme l'imputation par ratio; donneur; moyenne et valeur auxiliaire sont toutes des méthodes d'imputation linéaires. Le cadre théorique ainsi que l'expression donnant la décomposition de la variance due à la non-réponse au niveau de l'unité seront présentés. Des résultats par simulation seront aussi présentés. Cette décomposition peut être utilisée pour prioriser le suivi de non-réponse, prioriser les corrections manuelles ou simplement orienter l'analyse des données.

Mots-clés : Variance totale; plan adaptatif; imputation.

1 Introduction

Biemer (2010) décrit l'erreur totale d'enquête comme « l'accumulation de toutes les erreurs susceptibles de se produire dans la conception, la collecte, le traitement et l'analyse des données d'enquête ». Il classe les composantes de l'erreur d'enquête en erreurs d'échantillonnage et erreurs non attribuables à l'échantillonnage, comme les erreurs dues à la non-réponse, la couverture, la mesure et le traitement des données. Ces erreurs sont susceptibles d'avoir une incidence sur la variance, le biais ou les deux. Le paradigme de l'erreur totale d'enquête vise à maximiser la qualité de l'enquête en réduisant le plus possible l'erreur totale d'enquête dans le cadre de contraintes de ressources prédéterminées comme le budget, les ressources humaines ou le temps.

À Statistique Canada, l'Architecture opérationnelle du Bureau a lancé le Programme intégré de la statistique des entreprises (PISE) comme plateforme normalisée de plus de 140 enquêtes économiques dans un objectif d'efficacité et d'amélioration de la qualité et de l'adaptabilité. Ainsi, l'un des piliers du programme consiste à réduire les coûts de collecte tout en gérant les erreurs dues à la non-réponse. C'est pourquoi le programme repose sur un plan adaptatif permettant de traiter différemment les différentes unités. Pour en savoir plus sur le PISE, voir Statistique Canada (2015). Groves et Heeringa (2006) montrent comment on peut utiliser les paradonnées pour augmenter le taux de réponse. Schouten, Calinescu et Luiten (2013) ont présenté un cadre général de plan d'enquête adaptatif et expliqué comment l'indicateur R pourrait être utilisé dans ce contexte.

On a élaboré un nouveau modèle de processus d'enquête, dit modèle des « estimations en continu », pour traiter le pilier du PISE mentionné ci-dessus. Le modèle des estimations en continu se fonde sur des cycles de traitement et d'estimation itératifs sur toute la période de collecte. Il s'agit essentiellement de calculer les estimations clés avec les indicateurs de qualité connexes à plusieurs moments précis de la période de

1. Keven Bosa, Serge Godbout, Fraser Mills et Frédéric Picard, Statistique Canada, 100, promenade du pré Tunney, Ottawa (Ontario), K1A 0T6. Courriel : serge.godbout@canada.ca.

collecte. Au début, toutes les unités sont attribuées au traitement des enquêtes par autodéclaration, ce qui signifie qu'on demande aux répondants de remplir un questionnaire en ligne. Les efforts de collecte, comme le suivi de la non-réponse par interview téléphonique assistée par ordinateur, sont ensuite effectués sur les unités qui contribuent le plus aux estimations lorsque la qualité est jugée faible d'après les résultats préliminaires des estimations en continu. Cette méthode peut être considérée comme un plan adaptatif puisque les traitements sur les unités dépendent de la qualité des estimations produites pendant la période de collecte. La plupart des travaux relatifs à l'élaboration des plans d'enquête adaptatifs du PISE ont été réalisés après 2010. Godbout, Beaucage et Turmelle (2011), Turmelle, Godbout et Bosa (2012), Mills, Godbout, Bosa et Turmelle (2013), ainsi que Bosa et Godbout (2014) se sont appuyés sur cette idée dans le contexte du plan adaptatif du PISE afin de minimiser le nombre de suivis nécessaires pour atteindre la qualité visée en matière de coefficient de variation.

L'article revient sur le travail effectué jusqu'à maintenant pour le PISE et présente une méthode de décomposition de la variance due à la non-réponse dans un score au niveau de l'item pour une variable d'intérêt donnée dans un domaine. Au moyen de ce score d'item, on cherche à estimer la contribution à la variance assignée à une seule unité. Les unités ayant un score élevé contribueront le plus à réduire la variance et le coefficient de variation qui est souvent utilisé comme indicateur de qualité dans les enquêtes. Cependant, toute enquête comprend généralement plusieurs variables et domaines importants. La méthode proposée calcule d'abord, pour une unité donnée, les scores au niveau des items pour les variables et les domaines importants. Ensuite, on peut combiner les scores des items en un seul score d'unité afin de classer les unités. Par exemple, le score d'unité peut être une somme pondérée ou le maximum de ses scores d'item. L'utilisation la plus attrayante du score d'unité qui en résulte consiste à prioriser les unités, celles qui obtiennent les scores les plus élevés, pour les opérations de collecte les plus coûteuses comme le suivi téléphonique, l'interview téléphonique assistée par ordinateur ou l'interview sur place assistée par ordinateur. L'article suppose que la non-réponse totale et la non-réponse partielle sont traitées dans le plan adaptatif, mais que les traitements sont susceptibles de différer selon le type de non-réponse. Par exemple, des suivis téléphoniques pourraient être effectués en cas de non-réponse totale, tandis que les questionnaires comportant une non-réponse partielle pourraient être examinés par des analystes. Ce type de plan adaptatif produit de fortes interactions entre les activités de collecte, les données observées et la qualité mesurée. Bosa et Godbout (2014) montrent la façon dont cette technique a été mise en œuvre dans le PISE dans le cadre du modèle des estimations en continu.

L'article mettra l'accent sur la dérivation du score des items. Pour ce faire, on étudiera le cas particulier d'une seule variable d'intérêt dans un domaine. De plus, on utilisera une seule méthode d'imputation pour imputer la variable d'intérêt en cas de non-réponse afin de simplifier la notation et de rendre les résultats plus lisibles.

Le cadre d'inférence est décrit dans la section 2. Dans la section 3, la décomposition de la variance au niveau de l'unité est exprimée. En d'autres termes, la contribution de chaque unité non répondantes à la variance est calculée. On a réalisé une étude par simulation pour évaluer le score proposé. L'étude est décrite dans la section 4. Enfin, des réflexions et des conclusions sont présentées dans la section 5.

2 Cadre d'inférence

Supposons qu'un échantillon s de taille n est tiré d'une population U de taille N. Soit le total de la population définit par

$$t_d = \sum_{k \in U} d_k y_k \tag{2.1}$$

pour une variable, y, et un indicateur de domaine, d_k, qui prend la valeur $d_k = 1$ si l'unité k appartient au domaine d, et $d_k = 0$ sinon. En cas de réponse complète, t_d est estimé par $\hat{t}_d^0 = \sum_{k \in s} d_k w_k y_k$ où w_k pourrait être le poids d'échantillonnage ou un poids calé si un calage est effectué. Étant donné que les enquêtes sont généralement sujettes à la non-réponse, pour les unités comme pour les items, une unité d'échantillonnage est classée comme unité répondante ou unité non répondante en ce qui concerne la variable y à tout moment de la collecte des données. Le sous-ensemble s_r contient des unités répondant à la variable y, tandis que s_m contient les unités non répondantes pour cette variable. Notez que s_r et s_m, respectivement de taille n_r et n_m, forment une partition de l'échantillon s, $P_s = \{s_r, s_m\}$, avec $s_r \cup s_m = s$ et $s_r \cap s_m = \varnothing$.

La méthode proposée dans l'article suppose que l'imputation est utilisée en cas de non-réponse, ce qui est couramment le cas dans les enquêtes auprès des entreprises. De plus, on peut envisager cette technique pour la non-réponse partielle ou totale tant que l'imputation est utilisée. Toutefois, comme une seule variable d'intérêt y est prise en compte ici à des fins de simplicité, aucune distinction n'est faite suivant que la variable y est imputée en raison de la non-réponse totale ou partielle. De plus, les ensembles s_r et s_m ne sont pas indexés par un numéro d'item pour des questions de simplicité sans perte de généralité. Cependant, l'action qui suit le calcul d'un score d'unité peut être différente suivant que l'unité soit répondante ou non-répondante.

2.1 Estimation par imputation

Le cadre nécessite des méthodes d'imputation linéaire. En d'autres termes, la valeur imputée, y_k^*, peut être exprimée comme une combinaison linéaire des valeurs déclarées par les autres unités. Cette combinaison linéaire est donnée par $y_k^* = \varphi_{0k} + \sum_{l \in s_r} \varphi_{lk} y_l$. Les quantités, φ_{0k} et φ_{lk} ne dépendent pas des valeurs de la variable d'intérêt, y, mais elles peuvent dépendre de s, s_r et des données auxiliaires des non-répondants disponibles dans la base de sondage, les registres ou ailleurs. Les méthodes d'imputation linéaire recouvrent la plupart des méthodes utilisées en pratique, comme l'imputation des valeurs auxiliaires (Beaumont, Haziza et Bocci, 2011) et l'imputation par régression linéaire, ainsi que l'imputation par donneur, souvent utilisée pour imputer les variables catégoriques.

Il est courant d'utiliser une imputation composite, qui consiste à appliquer plusieurs méthodes d'imputation de façon séquentielle à une même variable. Il est possible d'utiliser plus d'une méthode d'imputation linéaire pour imputer les unités non répondantes. La section 2 de Beaumont et Bissonnette (2011) définit en détail l'imputation composite. En bref, supposons que l'ensemble de non-répondants est divisé en deux groupes ou plus et qu'une méthode d'imputation différente est utilisée dans chaque groupe.

Par exemple, supposons que $\mathbf{x}_k$ soit le vecteur complet des variables auxiliaires pour l'unité k, et qu'on utilise l'imputation par régression pour imputer la variable d'intérêt. Cependant, si, dans certains cas, $\mathbf{x}_k$ était incomplet, une autre méthode d'imputation, fondée sur le sous-ensemble disponible de $\mathbf{x}_k$, serait utilisée. On peut généraliser la méthode présentée dans le présent article pour y inclure l'imputation composite tant que des méthodes d'imputation linéaire sont utilisées. Pour simplifier la notation, on présente le cas d'une seule méthode d'imputation linéaire.

L'estimateur du total du domaine après l'imputation est donné par

$$\hat{t}_d = \sum\nolimits_{l \in s_r} w_l d_l y_l + \sum\nolimits_{k \in s_m} w_k d_k y_k^* \tag{2.2}$$

où w_k est le poids d'échantillonnage ou un poids calé. L'estimateur présenté dans l'équation (2.2) peut être réécrite ainsi

$$\begin{aligned}
\hat{t}_d &= \sum\nolimits_{l \in s_r} w_l d_l y_l + \sum\nolimits_{k \in s_m} w_k d_k y_k^* \\
&= \sum\nolimits_{l \in s_r} w_l d_l y_l + \sum\nolimits_{k \in s_m} w_k d_k \left(\varphi_{0k} + \sum_{l \in s_r} \varphi_{lk} y_l \right) \\
&= \sum\nolimits_{l \in s_r} w_l d_l y_l + \sum\nolimits_{k \in s_m} w_k d_k \varphi_{0k} + \sum\nolimits_{l \in s_r} y_l \sum\nolimits_{k \in s_m} w_k d_k \varphi_{lk} \\
&= W_{0d} + \sum\nolimits_{l \in s_r} w_l d_l y_l + \sum\nolimits_{l \in s_r} y_l W_{dl} \\
&= W_{0d} + \sum\nolimits_{l \in s_r} y_l \left(w_l d_l + W_{dl} \right).
\end{aligned}$$

Les quantités W_{dl} et W_{0d} désignent les facteurs de pondération compensatoire (ou poids d'ajustement) définis comme suit :

$$\begin{aligned}
W_{dl} &= \sum_{k \in s_m} w_k d_k \varphi_{lk} \\
W_{0d} &= \sum_{k \in s_m} w_k d_k \varphi_{0k} .
\end{aligned}$$

Elles représentent l'effet de la non-réponse dans le domaine, d, porté par l'unité du répondant, $l \in s_r$, avec une valeur déclarée, y_l.

2.2 Estimation de la variance

Soit un modèle d'imputation, η, décrivant la relation entre la variable y et le vecteur des variables auxiliaires observées $\mathbf{x}^{\text{obs}}$. Soit $E_\eta(.)$, $\text{Var}_\eta(.)$ et $\text{cov}_\eta(.)$ qui désignent respectivement l'espérance, la variance et la covariance par rapport au modèle d'imputation η. Le modèle d'imputation est :

$$\begin{aligned}
E_\eta \left(y_k \mid \mathbf{X}^{\text{obs}} \right) &= \mu_k \\
V_\eta \left(y_k \mid \mathbf{X}^{\text{obs}} \right) &= \sigma_k^2 \\
\text{cov}_\eta \left(y_k, y_{k'} \mid \mathbf{X}^{\text{obs}} \right) &= 0
\end{aligned}$$

où $k, k' \in U$ et $k \neq k'$. La matrice $\mathbf{X}^{\text{obs}}$ contient tous les vecteurs observés $\mathbf{x}^{\text{obs}}$. Les quantités μ_k et σ_k^2 peuvent être estimées par $\hat{\mu}_k$ et $\hat{\sigma}_k^2$ respectivement. Nous supposons que ces estimateurs sont sans biais

par rapport au modèle d'imputation η. Ces estimateurs seront utiles plus tard pour l'estimation des composantes de la variance totale et les décompositions des composantes au niveau des unités.

On peut exprimer l'erreur totale de l'estimateur (2.2) comme suit :

$$\hat{t}_d - t_d = \left(\hat{t}_d^0 - t_d\right) + \left(\hat{t}_d - \hat{t}_d^0\right), \tag{2.3}$$

où $\hat{t}_d^0$ est l'estimateur en présence de réponse complète donnée par (2.1). Habituellement, on appelle le premier terme de la partie droite de l'égalité (2.3) erreur d'échantillonnage et le deuxième terme erreur de non-réponse. Comme le proposent Särndal (1992) et Beaumont et Bissonnette (2011), l'erreur quadratique moyenne de $\hat{t}_d$ en utilisant (2.3) peut être décomposée en trois composantes et est obtenue au moyen de

$$\begin{aligned} E_{\eta pq}\left(\hat{t}_d - t_d\right)^2 &= E_\eta V_p\left(\hat{t}_d\right) + E_{pq}E_\eta\left[\left(\hat{t}_d - \hat{t}_d^0\right)^2 \mid s,\, s_r\right] \\ &\quad + 2E_{pq}E_\eta\left[\left(\hat{t}_d - \hat{t}_d^0\right)\left(\hat{t}_d^0 - t_d\right) \mid s,\, s_r\right], \end{aligned} \tag{2.4}$$

avec le modèle d'imputation, η, le plan d'échantillonnage, p, et le mécanisme de réponse, q. $E_{\eta pq}\left(\hat{t}_d - t_d\right)^2$ équivaut approximativement à la variance $V_{\eta pq}\left(\hat{t}_d - t_d\right)$ en supposant que le biais global est négligeable. Ainsi, l'équation (2.4) équivaut à $V_{\eta pq}\left(\hat{t}_d - t_d\right) \equiv V_{\text{TOT}}\left(\hat{t}_d\right) = V_{\text{SAM}}\left(\hat{t}_d\right) + V_{\text{NR}}\left(\hat{t}_d\right) + V_{\text{MIX}}\left(\hat{t}_d\right)$, où :

- $V_{\text{SAM}}\left(\hat{t}_d\right) \equiv E_\eta V_p\left(\hat{t}_d\right)$ est la variance d'échantillonnage;
- $V_{\text{NR}}\left(\hat{t}_d\right) \equiv E_{pq}E_\eta\left[\left(\hat{t}_d - \hat{t}_d^0\right)^2 \mid s,\, s_r\right]$ est la variance due à la non-réponse;
- $V_{\text{MIX}}\left(\hat{t}_d\right) \equiv 2E_{pq}E_\eta\left[\left(\hat{t}_d - \hat{t}_d^0\right)\left(\hat{t}_d^0 - t_d\right) \mid s,\, s_r\right]$ est la covariance entre les termes d'échantillonnage et d'erreur due à la non-réponse, aussi appelée composante de variance mixte.

Beaumont et Bissonnette (2011) proposent les estimateurs suivants pour $V_{\text{SAM}}\left(\hat{t}_d\right)$, $V_{\text{NR}}\left(\hat{t}_d\right)$ et $V_{\text{MIX}}\left(\hat{t}_d\right)$.

1. $\hat{V}_{\text{SAM}}\left(\hat{t}_d\right) = \hat{V}_{\text{ORD}}\left(\hat{t}_d\right) + \hat{V}_{\text{DIF}}\left(\hat{t}_d\right)$ où :
   - $\hat{V}_{\text{ORD}}\left(\hat{t}_d\right)$ est l'estimateur naïf de la variance d'échantillonnage qui utilise les valeurs imputées comme s'il s'agissait de valeurs déclarées.
   - $\hat{V}_{\text{DIF}}\left(\hat{t}_d\right) = \sum_{k \in s_m}\left(1 - \pi_k\right) w_k^2 d_k \hat{\sigma}_k^2$ est une correction à $\hat{V}_{\text{ORD}}\left(\hat{t}_d\right)$ afin de réduire le biais de $\hat{V}_{\text{ORD}}\left(\hat{t}_d\right)$, comme le proposent Beaumont et Bocci (2009), puisque la composante de variance $\hat{V}_{\text{ORD}}\left(\hat{t}_d\right)$ repose sur l'utilisation de valeurs imputées, généralement plus homogènes que les valeurs déclarées.
2. $\hat{V}_{\text{NR}}\left(\hat{t}_d\right) = \sum_{l \in s_r} W_{dl}^2 \hat{\sigma}_l^2 + \sum_{k \in s_m} w_k^2 d_k \hat{\sigma}_k^2$ est l'estimateur de la composante de non-réponse de la variance.
3. $\hat{V}_{\text{MIX}}\left(\hat{t}_d\right) = 2\sum_{l \in s_r} W_{dl}\left(w_l - 1\right) d_l \hat{\sigma}_l^2 - 2\sum_{k \in s_m} w_k\left(w_k - 1\right) d_k \hat{\sigma}_k^2$ est l'estimateur de la composante de variance mixte.

En présence de réponse complète, $s_m = \varnothing$, les facteurs de pondération compensatoires sont $W_{dl} = 0$, et les composantes de la variance, $\hat{V}_{\text{DIF}}\left(\hat{t}_d\right)$, $\hat{V}_{\text{NR}}\left(\hat{t}_d\right)$, et $\hat{V}_{\text{MIX}}\left(\hat{t}_d\right)$, sont également égaux à 0, ce qui donne une variance totale de $\hat{V}_{\text{TOT}}\left(\hat{t}_d\right) = \hat{V}_{\text{ORD}}\left(\hat{t}_d\right)$. Dans un recensement, $s = U$, les composantes de la variance, $\hat{V}_{\text{DIF}}\left(\hat{t}_d\right)$, $\hat{V}_{\text{ORD}}\left(\hat{t}_d\right)$, et $\hat{V}_{\text{MIX}}\left(\hat{t}_d\right)$, sont égales à 0, ce qui donne une variance totale de $\hat{V}_{\text{TOT}}\left(\hat{t}_d\right) = \hat{V}_{\text{NR}}\left(\hat{t}_d\right)$.

2.3 Biais de non-réponse

Dans tous les cas, la réduction du biais de non-réponse est souhaitable. On peut y parvenir au moyen d'un plan adaptatif ou d'une méthode appropriée de traitement des valeurs manquantes. Notre cadre suppose que le biais de non-réponse est éliminé par des méthodes d'imputation utilisant l'information auxiliaire pertinente. En pratique, il est probable que l'imputation réduise le biais de non-réponse, mais ne l'élimine pas. Nous pourrions alors nous interroger sur la possibilité d'utiliser des plans adaptatifs pour réduire davantage le biais. Dans le contexte de la pondération de la non-réponse, Beaumont, Bocci et Haziza (2014) soutiennent que l'information auxiliaire utilisée dans un plan adaptatif pour la réduction du biais attribuable à la non-réponse peut aussi servir à la pondération de la non-réponse pour réduire d'autant le biais. On peut aussi avancer leur argument dans le contexte de l'imputation. Il justifie que nous mettions l'accent sur la réduction de la variance plutôt que sur la réduction du biais. Nous savons qu'un biais pourrait demeurer après l'imputation, mais nous l'ignorerons parce qu'il ne sera peut-être pas possible de le réduire davantage au moyen d'un plan adaptatif sans information auxiliaire supplémentaire. Il est en revanche possible de réduire la variance au moyen d'un plan adaptatif.

3 Décomposition au niveau de l'unité de l'erreur provenant des composantes de la variance

Cette section décrit la méthode utilisée pour évaluer la contribution d'une unité non répondante donnée, $\lambda \in s_m$, à la variance totale estimée pour l'estimation du total pour une variable donnée.

La décomposition de l'erreur au niveau de l'unité, δ_λ, de la variance totale pour une unité donnée, λ, est définie comme la différence entre la variance totale estimée et la variance totale projetée, c'est-à-dire $\delta_\lambda\left(\hat{V}_{\text{TOT}}\left(\hat{t}_d\right)\right) \equiv \hat{V}_{\text{TOT}}\left(\hat{t}_d\right) - \hat{V}_{\text{TOT}}^{(\lambda)}\left(\hat{t}_d\right)$. L'exposant (λ) indique les quantités projetées lorsque l'unité λ est convertie en unité répondante. Donc, on peut considérer $\delta_\lambda\left(\hat{V}_{\text{TOT}}\left(\hat{t}_d\right)\right)$ comme le gain attendu, pour ce qui est de la variance totale, de la conversion d'une unité non répondante λ en unité répondante.

Pour obtenir $\delta_\lambda\left(\hat{V}_{\text{TOT}}\left(\hat{t}_d\right)\right)$, λ est déplacée de s_m à s_r, ce qui génère la nouvelle partition $P_s^{(\lambda)}$ de l'échantillon de P_s où $P_s^{(\lambda)} = \left\{s_r^{(\lambda)}, s_m^{(\lambda)}\right\}$, $s_r^{(\lambda)} = s_r \cup \{\lambda\}$ et $s_m^{(\lambda)} = s_m \setminus \{\lambda\}$, comme l'illustre la figure 3.1.

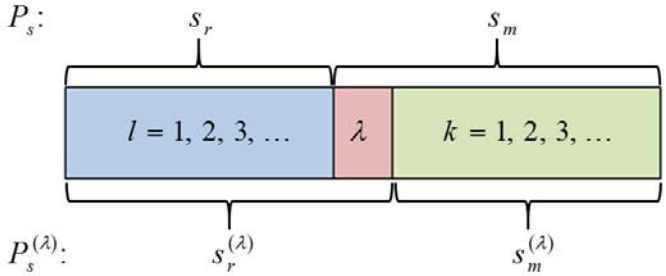

Figure 3.1 Partitions d'échantillon.

Certaines hypothèses sont nécessaires pour décomposer les composantes de la variance. On sait que ces hypothèses ne sont pas nécessairement exactes. Elles peuvent cependant servir à produire de bons résultats, comme le montre la simulation de la section 4. Les hypothèses requises sont :

1. Valeur déclarée projetée : soit $\lambda \in s_m$ converti en réponse et soit $y_\lambda^{(\lambda)} = y_\lambda^*$.
2. Paramètres d'imputation projetés : $\forall k \in s_m$, $\hat{\mu}_k^{(\lambda)} = \hat{\mu}_k$ et $\hat{\sigma}_k^{(\lambda)} = \hat{\sigma}_k$.
3. Matrice des relations d'imputation projetées : $\forall k \in s_m$ et $\forall l \in s_r$, $\varphi_{lk}^{(\lambda)} = 0$ si $l = \lambda$ ou si $k = \lambda$ ou $\varphi_{lk}^{(\lambda)} = \varphi_{lk}$ sinon. De même, $\varphi_{0k}^{(\lambda)} = 0$ si $k = \lambda$ ou $\varphi_{0k}^{(\lambda)} = \varphi_{0k}$ sinon.

L'hypothèse 1 suppose que si une unité non répondante, λ, est convertie en unité répondante, sa valeur déclarée est égale à sa valeur imputée. Cela n'est pas vrai de façon générale, mais la valeur imputée est la meilleure estimation que nous obtenons. Cette valeur imputée devrait se rapprocher suffisamment de la valeur déclarée pour permettre d'estimer l'erreur sur les composantes de la variance. Cette hypothèse aura une incidence au moment de la décomposition de la variance d'échantillonnage.

L'hypothèse 2 suppose que les paramètres estimés du modèle d'imputation demeurent inchangés si λ est répondante. Dans le cas d'un estimateur convergent de paramètre de modèle d'imputation, cette hypothèse est plus réaliste lorsque s_r est plus grand.

Enfin, l'hypothèse 3 signifie que la relation d'imputation entre les non-répondants et les répondants demeure inchangée, sauf quand l'unité λ est impliquée. En d'autres termes, l'unité convertie, λ, n'est plus imputée à partir des répondants, mais elle ne servira pas à imputer d'autres unités non répondantes. La figure 3.2 montre comment l'hypothèse 3 est représentée dans la matrice phi.

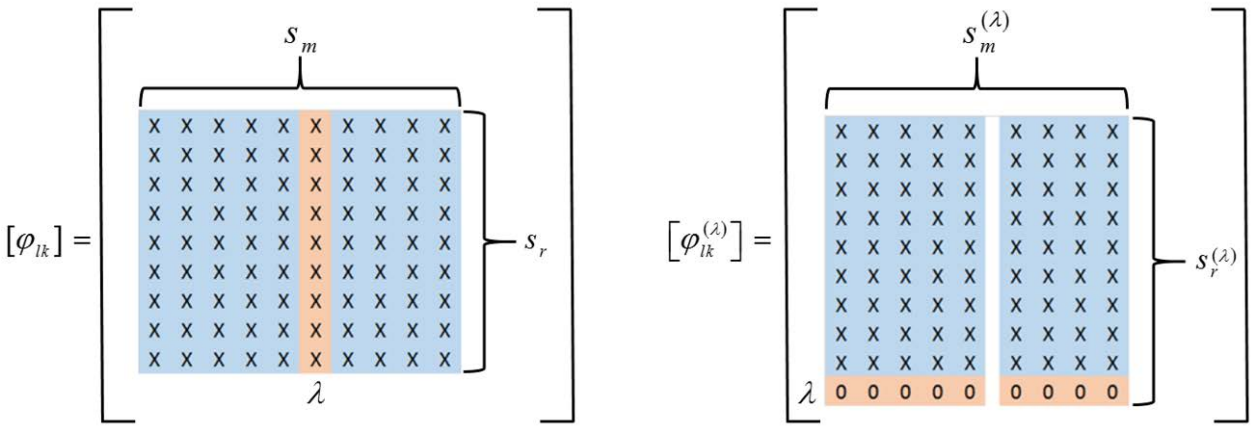

Figure 3.2 Matrice phi de la relation d'imputation initiale et projetée.

Par conséquent, le facteur de pondération de compensation, $W_{dl}^{(\lambda)}$, d'une unité répondante, $\forall l \in s_r$, est projeté comme étant

$$\begin{aligned} W_{dl}^{(\lambda)} &= \sum_{k \in s_m^{(\lambda)}} w_k d_k \varphi_{lk}^{(\lambda)} \\ &= \sum_{k \in s_m} w_k d_k \varphi_{lk} - w_\lambda d_\lambda \varphi_{l\lambda} \\ &= W_{dl} - w_\lambda d_\lambda \varphi_{l\lambda}. \end{aligned} \tag{3.1}$$

On retire le poids marginal de l'unité convertie λ du poids initial de compensation, W_{dl}, pour obtenir le nouveau $W_{dl}^{(\lambda)}$. Notez que $W_{d\lambda}^{(\lambda)} = \sum_{k \in s_m^{(\lambda)}} w_k d_k \varphi_{\lambda k}^{(\lambda)} = 0$ parce que $\varphi_{\lambda k}^{(\lambda)} = 0$ dans l'hypothèse 3. Comme mentionné précédemment, cela signifie que λ n'est pas utilisé pour imputer les non-répondants.

Dans les sous-sections suivantes, la décomposition de l'erreur au niveau de l'unité pour l'unité λ est calculée pour les quatre composantes de la variance, conformément à ce qui est décrit dans la section 2.3.

3.1 Décomposition au niveau de l'unité de l'erreur de la variance naïve d'échantillonnage

La quantité $\hat{V}_{\text{ORD}}(\hat{t}_d)$ dépend des valeurs *y*, des poids finaux et des probabilités de sélection d'ordre un et d'ordre deux. La décomposition de l'erreur au niveau de l'unité de la composante de variance d'échantillonnage naïf $\hat{V}_{\text{ORD}}(\hat{t}_d)$ est triviale puisque l'hypothèse selon laquelle l'unité λ passe de s_m à s_r ne change ni les poids ni les probabilités de sélection. Dans l'hypothèse 1, la valeur déclarée prévue $y_\lambda^{(\lambda)}$ est établie à y_λ^* de sorte que $\hat{V}_{\text{ORD}}^{(\lambda)}(\hat{t}_d) = \hat{V}_{\text{ORD}}(\hat{t}_d)$ lorsque λ est convertie en unité répondante. Par conséquent, la décomposition de $\hat{V}_{\text{ORD}}(\hat{t}_d)$ est obtenue au moyen de

$$\delta_\lambda\left(\hat{V}_{\text{ORD}}(\hat{t}_d)\right) \equiv \hat{V}_{\text{ORD}}(\hat{t}_d) - \hat{V}_{\text{ORD}}^{(\lambda)}(\hat{t}_d) = 0. \tag{3.2}$$

Ce résultat est cohérent avec l'idée que l'estimation ponctuelle de la variance d'échantillonnage naïf changera probablement, mais elle ne devrait pas diminuer avec l'ajout d'une unité répondante.

3.2 Décomposition au niveau de l'unité de la correction de la composante de la variance d'échantillonnage

La décomposition de l'erreur au niveau de l'unité pour l'unité λ de la correction de la composante de la variance d'échantillonnage, $\hat{V}_{\text{DIF}}(\hat{t}_d)$, est obtenue au moyen de

$$\begin{aligned}\delta_\lambda\left(\hat{V}_{\text{DIF}}(\hat{t}_d)\right) &\equiv \hat{V}_{\text{DIF}}(\hat{t}_d) - \hat{V}_{\text{DIF}}^{(\lambda)}(\hat{t}_d) \\ &= \sum_{k \in s_m}(1-\pi_k)\, d_k w_k^2 \hat{\sigma}_k^2 - \sum_{\lambda \in s_m^{(\lambda)}}(1-\pi_k)\, d_k w_k^2 \left(\hat{\sigma}_k^{(\lambda)}\right)^2.\end{aligned}$$

Dans l'hypothèse 2, $\hat{\sigma}_k^{(\lambda)} = \hat{\sigma}_k$, de sorte que

$$\delta_\lambda\left(\hat{V}_{\text{DIF}}(\hat{t}_d)\right) = (1-\pi_\lambda)\, d_\lambda w_\lambda^2 \hat{\sigma}_\lambda^2. \tag{3.3}$$

Le lecteur averti remarquera qu'il ne devrait pas y avoir d'incidence sur la variance d'échantillonnage réel (et non son estimation) qu'une unité soit répondante ou non. Cependant, nous avons décidé d'inclure l'incidence qu'une unité a sur l'estimation de la variance d'échantillonnage afin de traiter de façon cohérente les trois composantes $V_{\text{SAM}}(\hat{t}_d)$, $V_{\text{NR}}(\hat{t}_d)$ et $V_{\text{MIX}}(\hat{t}_d)$.

3.3 Décomposition au niveau de l'unité de la composante de la variance de non-réponse

La décomposition de l'erreur au niveau de l'unité pour l'unité λ de la composante de la variance de non-réponse $\hat{V}_{\text{NR}}(\hat{t}_d)$ est obtenue au moyen de

$$\delta_\lambda\left(\hat{V}_{\mathrm{NR}}\left(\hat{t}_d\right)\right) \equiv \hat{V}_{\mathrm{NR}}\left(\hat{t}_d\right) - \hat{V}_{\mathrm{NR}}^{(\lambda)}\left(\hat{t}_d\right)$$

$$= \left(\sum_{l\in s_r} W_{dl}^2\hat{\sigma}_l^2 + \sum_{k\in s_m} w_k^2 d_k\hat{\sigma}_k^2\right) - \left(\sum_{l\in s_r^{(\lambda)}} \left(W_{dl}^{(\lambda)}\right)^2\left(\hat{\sigma}_l^{(\lambda)}\right)^2 + \sum_{k\in s_m^{(\lambda)}} w_k^2 d_k\left(\hat{\sigma}_k^{(\lambda)}\right)^2\right).$$

Dans les hypothèses 2 et 3, $\hat{\sigma}_k^{(\lambda)} = \hat{\sigma}_k$ et $W_{d\lambda}^{(\lambda)} = 0$. On peut réécrire l'expression précédente comme suit :

$$\delta_\lambda\left(\hat{V}_{\mathrm{NR}}\left(\hat{t}_d\right)\right) = \left(\sum_{l\in s_r} W_{dl}^2\hat{\sigma}_l^2 - \sum_{l\in s_r}\left(W_{dl}^{(\lambda)}\right)^2\hat{\sigma}_l^2\right) + w_\lambda^2 d_\lambda\hat{\sigma}_\lambda^2.$$

Au moyen de la formule (3.1), ceci devient

$$\delta_\lambda\left(\hat{V}_{\mathrm{NR}}\left(\hat{t}_d\right)\right) = \left(\sum_{l\in s_r} W_{dl}^2\hat{\sigma}_l^2 - \sum_{l\in s_r}\left(W_{dl} - w_\lambda d_\lambda\varphi_{l\lambda}\right)^2\hat{\sigma}_l^2\right) + w_\lambda^2 d_\lambda\hat{\sigma}_\lambda^2$$

$$= \left(\sum_{l\in s_r} W_{dl}^2\hat{\sigma}_l^2 - \left(W_{dl}^2 - 2W_{dl}w_\lambda d_\lambda\varphi_{l\lambda} + w_\lambda^2 d_\lambda\varphi_{l\lambda}^2\right)\hat{\sigma}_l^2\right) + w_\lambda^2 d_\lambda\hat{\sigma}_\lambda^2$$

$$= \sum_{l\in s_r}\left(2W_{dl}w_\lambda d_\lambda\varphi_{l\lambda} - w_\lambda^2 d_\lambda\varphi_{l\lambda}^2\right)\hat{\sigma}_l^2 + w_\lambda^2 d_\lambda\hat{\sigma}_\lambda^2. \tag{3.4}$$

3.4 Décomposition au niveau de l'unité de la composante de variance mixte

Enfin, l'incidence de l'unité λ sur le terme de la composante de variance, $\hat{V}_{\mathrm{MIX}}\left(\hat{t}_d\right)$, est obtenue au moyen de

$$\delta_\lambda\left(\hat{V}_{\mathrm{MIX}}\left(\hat{t}_d\right)\right) \equiv \hat{V}_{\mathrm{MIX}}\left(\hat{t}_d\right) - \hat{V}_{\mathrm{MIX}}^{(\lambda)}\left(\hat{t}_d\right)$$

$$= \left(2\sum_{l\in s_r} W_{dl}\left(w_l - 1\right)d_l\hat{\sigma}_l^2 - 2\sum_{k\in s_m} w_k\left(w_k - 1\right)d_k\hat{\sigma}_k^2\right)$$

$$- \left(2\sum_{l\in s_r^{(\lambda)}} W_{dl}^{(\lambda)}\left(w_l - 1\right)d_l\left(\hat{\sigma}_l^{(\lambda)}\right)^2 - 2\sum_{k\in s_m^{(\lambda)}} w_k\left(w_k - 1\right)d_k\left(\hat{\sigma}_k^{(\lambda)}\right)^2\right).$$

Cette équation peut être réécrite comme suit, dans les hypothèses 2 et 3 et l'équation (3.1)

$$\delta_\lambda\left(\hat{V}_{\mathrm{MIX}}\left(\hat{t}_d\right)\right) = \left(2\sum_{l\in s_r} W_{dl}\left(w_l - 1\right)d_l\hat{\sigma}_l^2 - 2\sum_{k\in s_m} w_k\left(w_k - 1\right)d_k\hat{\sigma}_k^2\right)$$

$$- \left(2\sum_{l\in s_r}\left(W_{dl} - w_\lambda d_\lambda\varphi_{l\lambda}\right)\left(w_l - 1\right)d_l\hat{\sigma}_l^2 - 2\sum_{k\in s_m^{(\lambda)}} w_k\left(w_k - 1\right)d_k\hat{\sigma}_k^2\right)$$

$$= 2\sum_{l\in s_r} w_\lambda d_\lambda\varphi_{l\lambda}\left(w_l - 1\right)d_l\hat{\sigma}_l^2 - 2w_\lambda\left(w_\lambda - 1\right)d_\lambda\hat{\sigma}_\lambda^2. \tag{3.5}$$

Dans la section 2.3, l'estimation de la variance totale, $\hat{V}_{\mathrm{TOT}}\left(\hat{t}_d\right)$, a été définie comme étant $\hat{V}_{\mathrm{TOT}}\left(\hat{t}_d\right) = \hat{V}_{\mathrm{ORD}}\left(\hat{t}_d\right) + \hat{V}_{\mathrm{DIF}}\left(\hat{t}_d\right) + \hat{V}_{\mathrm{NR}}\left(\hat{t}_d\right) + \hat{V}_{\mathrm{MIX}}\left(\hat{t}_d\right)$. De la même manière, l'incidence de l'unité λ sur $\hat{V}_{\mathrm{TOT}}\left(\hat{t}_d\right)$ est définie comme suit :

$$\delta_\lambda\left(\hat{V}_{\mathrm{TOT}}\left(\hat{t}_d\right)\right)=\delta_\lambda\left(\hat{V}_{\mathrm{ORD}}\left(\hat{t}_d\right)\right)+\delta_\lambda\left(\hat{V}_{\mathrm{DIFF}}\left(\hat{t}_d\right)\right)+\delta_\lambda\left(\hat{V}_{\mathrm{NR}}\left(\hat{t}_d\right)\right)+\delta_\lambda\left(\hat{V}_{\mathrm{MIX}}\left(\hat{t}_d\right)\right),$$

où $\delta_\lambda\left(\hat{V}_{\mathrm{ORD}}\left(\hat{t}_d\right)\right)$, $\delta_\lambda\left(\hat{V}_{\mathrm{DIF}}\left(\hat{t}_d\right)\right)$, $\delta_\lambda\left(\hat{V}_{\mathrm{NR}}\left(\hat{t}_d\right)\right)$, et $\delta_\lambda\left(\hat{V}_{\mathrm{MIX}}\left(\hat{t}_d\right)\right)$ sont respectivement obtenus par les équations (3.2), (3.3), (3.4) et (3.5).

On peut observer (voir les preuves en annexe) que $\hat{V}_{\mathrm{DIF}}\left(\hat{t}_d\right)=\sum_{k\in s_m}\delta_k\left(\hat{V}_{\mathrm{DIF}}\left(\hat{t}_d\right)\right)$ et $\hat{V}_{\mathrm{MIX}}\left(\hat{t}_d\right)=\sum_{k\in s_m}\delta_k\left(\hat{V}_{\mathrm{MIX}}\left(\hat{t}_d\right)\right)$. Cependant, cette relation linéaire ne s'avère pas pour $\hat{V}_{\mathrm{NR}}\left(\hat{t}_d\right)$. Il est important de tenir compte de cette propriété parce que, pour $\hat{V}_{\mathrm{DIF}}\left(\hat{t}_d\right)$ et $\hat{V}_{\mathrm{MIX}}\left(\hat{t}_d\right)$, la somme des erreurs au niveau de l'unité de toutes les unités non répondantes, $k\in s_m$, est égale à la composante de variance estimée correspondante. Dans le cas de la composante de la variance de la non-réponse, la somme des erreurs est différente de $\hat{V}_{\mathrm{NR}}\left(\hat{t}_d\right)$. La différence est obtenue au moyen de

$$\sum_{k\in s_m}\delta_k\left(\hat{V}_{\mathrm{NR}}\left(\hat{t}_d\right)\right)-\hat{V}_{\mathrm{NR}}\left(\hat{t}_d\right)=\sum_{l\in s_r}\left(\left(\sum_{k\in s_m}w_k d_k\varphi_{lk}\right)^2-\sum_{k\in s_m}w_k^2 d_k\varphi_{lk}^2\right)\hat{\sigma}_l^2. \tag{3.6}$$

Cette différence peut être relativement faible, surtout dans les enquêtes auprès des entreprises caractérisées par des données asymétriques. C'est le cas quand $\max_{k\in s_m}\left(w_k d_k\varphi_{lk}\right)\cong\sum_{k\in s_m}w_k d_k\varphi_{lk}$. Cela correspond aux résultats présentés par Mills et coll. 2013.

Dans l'ensemble, on peut considérer la variance totale comme approximativement linéaire pour ce qui est des erreurs au niveau de l'unité, surtout dans le cas des enquêtes par sondage où $\hat{V}_{\mathrm{ORD}}\left(\hat{t}_d\right)$, $\hat{V}_{\mathrm{DIF}}\left(\hat{t}_d\right)$ et $\hat{V}_{\mathrm{MIX}}\left(\hat{t}_d\right)$ sont des facteurs importants de la variance totale.

4 Étude par simulation

La somme des contributions des items devrait se rapprocher suffisamment de la variance estimée due à la non-réponse. On a effectué des simulations pour évaluer la validité du score proposé. L'objectif était alors d'évaluer si la contribution d'item proposée est une bonne approximation de la contribution réelle à la variance totale d'une unité donnée. Pour ce faire, on a comparé les contributions totales d'un sous-ensemble aléatoire de s_m à la différence des variances estimées où ce sous-ensemble est respectivement considéré comme des unités non répondantes et des unités répondantes.

Les étapes qui suivent montrent comment les simulations ont été effectuées.

1. On a créé une population à partir d'une variable auxiliaire x générée suivant une loi gamma avec une moyenne de 48 et une variance de 768. La variable d'intérêt y a été créée conditionnellement sur x à partir d'une loi gamma avec une moyenne de $1{,}5x$ et une variance de $16x$. Ces paramètres sont identiques à ceux établis par Beaumont et Bissonnette (2011).
2. On a sélectionné un échantillon aléatoire simple s dans cette population et un sous-ensemble indépendant de non-réponse s_m a été généré à l'aide d'un échantillonnage de Bernoulli.
    a. Les unités non répondantes de s_m ont été imputées au moyen d'une imputation par le ratio, où $y_k^*=x_k\left(\sum_{l\in s_r}y_l\right)\left(\sum_{l\in s_r}x_l\right)^{-1}$ et

$$\hat{\sigma}_k^2 = x_k \frac{\sum_{l \in s_r} (y_l - y_l^*)^2}{\sum_{l \in s_r} x_l}.$$

b. On a estimé le total de la population $\hat{t}$, les composantes de la variance $\hat{V}_{\bullet}(\hat{t})$, et les décompositions au niveau de l'unité $\delta_k \left(\hat{V}_{\bullet}(\hat{t})\right)$ où l'indice $\bullet$ représente n'importe laquelle des composantes de la variance.

3. Un sous-ensemble, Λ, d'unités, λ, de s_m, sélectionné indépendamment dans une expérience de Bernoulli, a été déplacé de s_m à s_r de façon à simuler la conversion de non-réponse. Par conséquent, nous avons une nouvelle partition, $P_s^{(\Lambda)}$, avec $s_m^{(\Lambda)} = s_m \setminus \Lambda$ et $s_r^{(\Lambda)} = s_r \cup \Lambda$.

   a. Les unités non répondantes k de $s_m^{(\Lambda)}$ ont été imputées de nouveau à l'aide d'un modèle de ratio donné par $y_k^{**} = x_k \left(\sum_{l \in s_r^{(\Lambda)}} y_l\right)\left(\sum_{l \in s_r^{(\Lambda)}} x_l\right)^{-1}$ et

$$\hat{\sigma}_k^{*2} = x_k \frac{\sum_{l \in s_r^{(\Lambda)}} (y_l - y_l^{**})^2}{\sum_{l \in s_r^{(\Lambda)}} x_l}.$$

   b. Le total de la population, $\hat{t}^{(\Lambda)}$, et les composantes de la variance, $\hat{V}_{\bullet}\left(\hat{t}^{(\Lambda)}\right)$, ont été estimés.

4. Le total des décompositions au niveau de l'unité, $\sum_{\lambda \in \Lambda} \delta_\lambda \left(\hat{V}_{\bullet}(\hat{t})\right)$, pour les unités λ de Λ a été comparé à la différence dans les estimations de composante de la variance, $\hat{V}_{\bullet}(\hat{t}) - \hat{V}_{\bullet}\left(\hat{t}^{(\Lambda)}\right)$. La différence relative dans l'erreur de décomposition, DRel, a été calculée comme étant

$$\text{DRel} = \frac{\left(\hat{V}_{\bullet}\left(\hat{t}^{(\Lambda)}\right) + \sum_{\lambda \in \Lambda} \delta_\lambda \left(\hat{V}_{\bullet}(\hat{t})\right)\right) - \hat{V}_{\bullet}(\hat{t})}{\hat{V}_{\bullet}(\hat{t})}. \tag{4.1}$$

On a répété les étapes 1 à 4 indépendamment avec différentes combinaisons de taille de population, de taille d'échantillon, de taux de réponse et de taux de conversion, comme le décrivent les sections 4.1, 4.2 et 4.3.

4.1 Scénario de simulation 1 : paramètres fixes

Dans le scénario 1, la taille de la population, la taille de l'échantillon, le taux de réponse et le taux de conversion ont été fixés respectivement à 400, 100, 70 % et 33,3 %, avec 200 itérations indépendantes. Les résultats sont présentés dans les figures 4.1 et 4.2.

Les figures 4.1 et 4.2 montrent que la somme de la décomposition au niveau de l'unité est une bonne variable explicative de la variation des estimations de la composante de non-réponse. La différence relative moyenne des estimations de la variance est faible à 2,1 %, mais l'erreur-type est importante à 5,8 %. Sur les 200 différences relatives, seules 19 ne sont pas dans la fourchette de +/– 10 %, mais elles sont toutes supérieures à 10 %. Si un non-répondant est converti en répondant, nous concluons que la composante de non-réponse de la variance sera approximativement réduite par la contribution mesurée de cette unité.

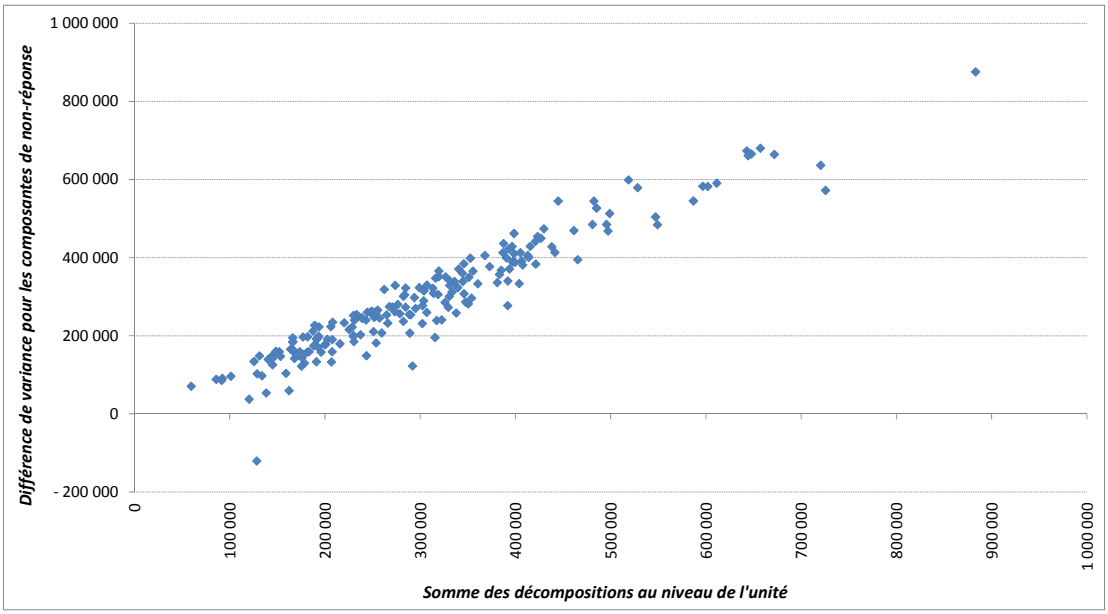

Figure 4.1 Différence de variance entre les composantes de non-réponse et les décompositions totales au niveau de l'unité avec des paramètres fixes.

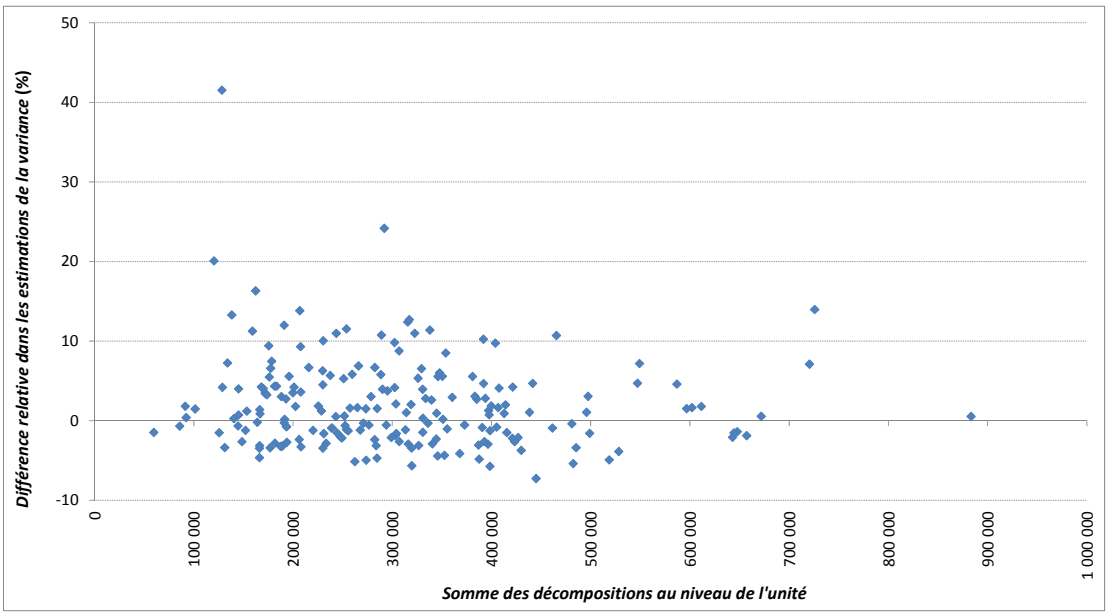

Figure 4.2 Différence relative entre les estimations de la variance et les décompositions totales au niveau de l'unité avec des paramètres fixes.

4.2 Scénario de simulation 2 : variation de la population et de la taille des échantillons

Dans le scénario 2, la taille de la population variait de 100 à 50 000, le taux d'échantillonnage, le taux de réponse et la conversion étant fixés à 20 %, 70 % et 33,3 % respectivement. Un plus grand nombre

d'itérations (40) a été créé pour la plus petite population $(N = 100)$, et moins (10) pour la plus grande $(N = 50\ 000)$, pour des questions opérationnelles. Les résultats sont présentés dans les figures 4.3 et 4.4.

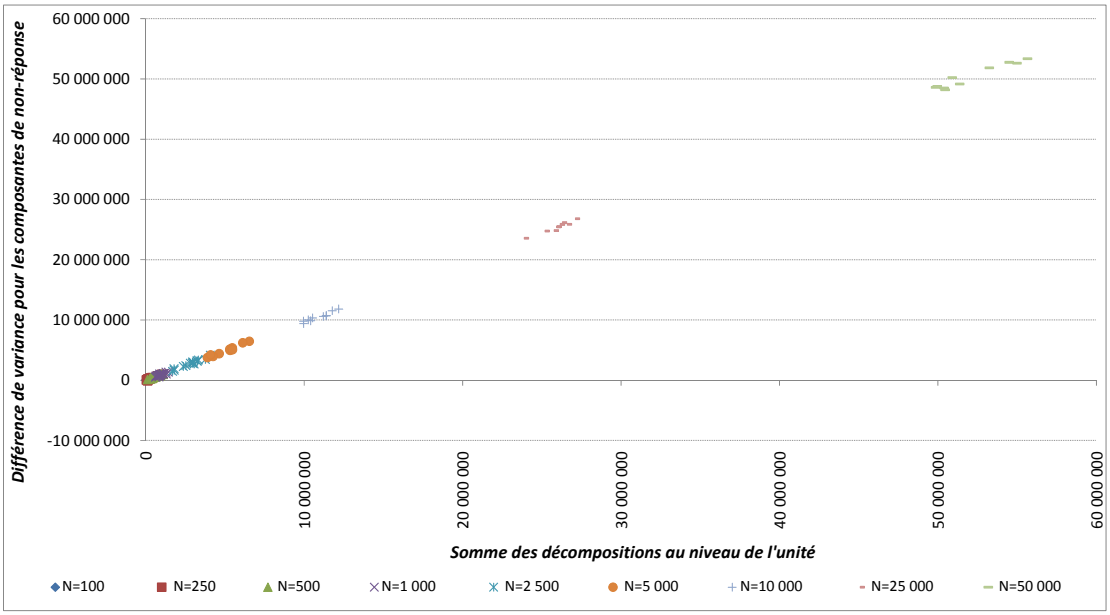

Figure 4.3 Différence de variance entre les composantes de non-réponse et les décompositions totales au niveau de l'unité, selon la taille de la population.

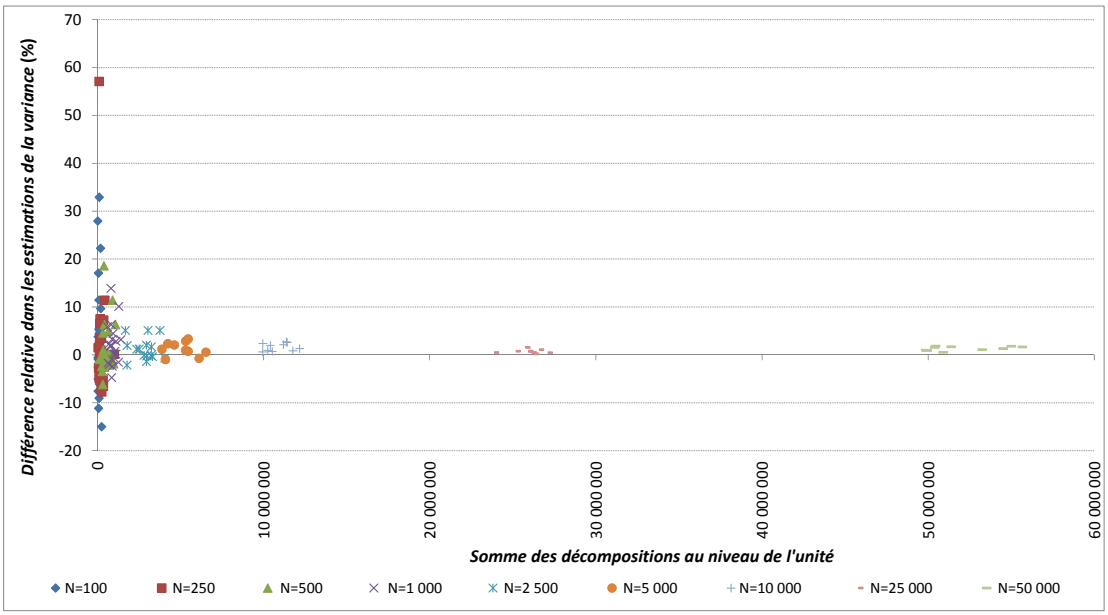

Figure 4.4 Différence relative entre les estimations de la variance et les décompositions au niveau de l'unité totale, selon la taille de la population.

Les figures 4.3 et 4.4 montrent que les différences relatives des erreurs de décomposition sont plus volatiles pour les populations plus petites, mais qu'elles convergent rapidement vers 0 à mesure que la taille de population et la taille d'échantillonnage augmentent. Cela est confirmé dans le tableau 4.1.

Tableau 4.1
Nombre, moyenne et écart-type des différences relatives des estimations de la variance selon la taille de la population

Taille de la population (N)	Différences relatives dans les estimations de la variance en pourcentage		
	Nombre	Moyenne	Écart-type
100	33(*)	2,2	10,6
250	30	1,6	11,4
500	25	1,0	5,3
1 000	20	2,2	4,4
2 500	10	1,2	2,3
5 000	10	1,2	1,4
10 000	10	1,6	0,8
25 000	10	0,7	0,4
50 000	10	1,3	0,4
Total	163	1,6	7,3

(*) : sur les 40 répétitions créées, seules 33 avaient des unités converties.

Pour déterminer les sources d'instabilité, les différences relatives entre la variance d'imputation estimée, $\mathrm{DRel}\left(\hat{\sigma}_k^2\right) = \left(\hat{\sigma}_k^2 - \hat{\sigma}_k^{2(\Lambda)}\right) \Big/ \hat{\sigma}_k^{2(\Lambda)}$, et la différence relative entre l'élément de relation d'imputation estimée, $\mathrm{DRel}\left(\varphi_{lk}\right) = \left(\varphi_{lk} - \varphi_{lk}^{(\lambda)}\right) \Big/ \varphi_{lk}^{(\lambda)}$, ont été mesurées pour toutes les unités, $k \notin \Lambda$ et $l \notin \Lambda$. Il faut noter que dans le modèle d'imputation par le ratio, les deux sont constantes pour une répétition donnée, c'est-à-dire, $\mathrm{DRel}\left(\hat{\sigma}_k^2\right) = \mathrm{DRel}\left(\hat{\sigma}^2\right)$ et $\mathrm{DRel}\left(\varphi_{lk}\right) = \mathrm{DRel}\left(\varphi\right)$. Après la suppression des deux répétitions extrêmes, la corrélation entre la différence relative des estimations de la variance DRel et $\mathrm{DRel}\left(\hat{\sigma}^2\right)$ est égale à 0,78 alors que la corrélation entre DRel et $\mathrm{DRel}\left(\varphi\right)$ est égale à 0,01. Cela montre que la principale cause de l'instabilité est la variabilité des estimations de $\hat{\sigma}_l^{(\lambda)} = \hat{\sigma}_l$. À partir de ce scénario, nous obtenons les conclusions suivantes :

- L'hypothèse 2 devient valide pour des tailles d'échantillon suffisamment grandes et elle conduit à une décomposition au niveau de l'unité plus exacte pour des estimateurs de variance du modèle d'imputation convergents.
- La décomposition au niveau de l'unité est robuste pour la validité de l'hypothèse 3.

4.3 Scénario de simulation 3 : variation des taux de conversion

Dans le scénario 3, la taille de la population et la taille de l'échantillon ont été fixées à 2 500 et 500, respectivement, et le taux de réponse est fixé à 50 %. On faisait varier les taux de conversion (TC) de 10 % à 100 % par intervalle de 10 %, afin de générer différentes tailles de sous-ensemble Λ, avec 15 itérations chacune. Les résultats sont présentés dans les figures 4.5 et 4.6.

Les figures 4.5 et 4.6 montrent que la différence relative des erreurs de décomposition présente un biais à mesure que la taille de Λ augmente, comme le confirme le tableau 4.2. Cela s'explique principalement

par la non-linéarité de $\hat{V}_{\mathrm{NR}}\left(\hat{t}_d\right)$, comme le montre l'équation (3.6). La nature monotone de la relation de la figure 4.5 semble indiquer qu'il n'y a pas d'incidence sur l'ordre des unités contribuant aux erreurs, ce qui signifie que les grands contributeurs estimés auront un effet plus important sur la variance que ceux dont la contribution estimée est faible.

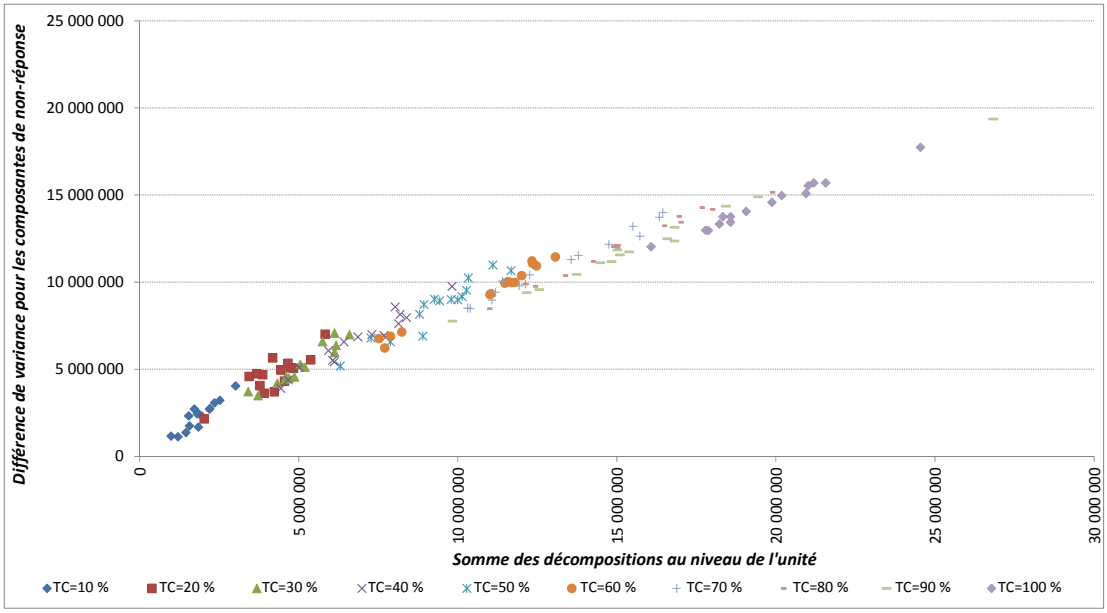

Figure 4.5 Différence de variance entre les composantes de non-réponse et les décompositions totales au niveau de l'unité, selon le taux de conversion (TC).

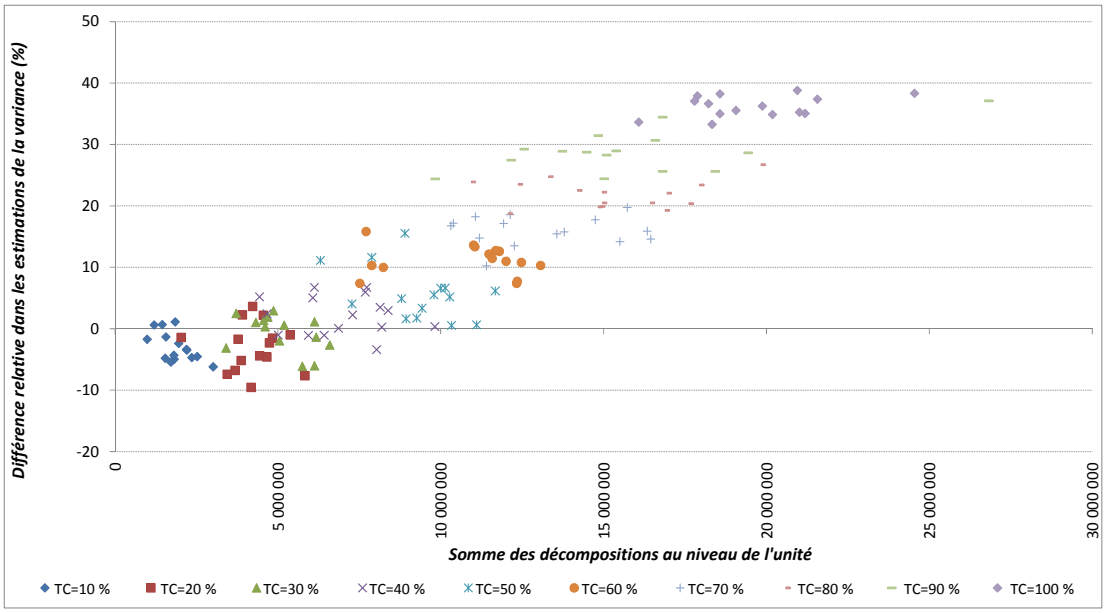

Figure 4.6 Différence relative entre les estimations de la variance et les décompositions au niveau de l'unité totale, selon le taux de conversion (TC).

Tableau 4.2
Nombre, moyenne et écart-type des différences relatives des estimations de la variance selon le taux de conversion (TC)

Taux de conversion (TC)	Différences relatives dans les estimations de la variance en pourcentage		
	Nombre	Moyenne	Écart-type
10 %	15	-3,0	2,4
20 %	15	-3,0	3,9
30 %	15	-0,5	3,0
40 %	15	2,2	3,3
50 %	15	5,7	4,3
60 %	15	11,1	2,4
70 %	15	16,0	2,4
80 %	15	21,9	2,3
90 %	15	28,9	3,5
100 %	15	36,2	1,7
Total	150	11,5	13,5

Bien que les différences relatives dans les estimations de la variance ne soient pas nulles en moyenne, on peut utiliser la décomposition des erreurs proposée pour déterminer les plus grandes sources de variance, surtout dans les populations asymétriques. Mills et coll. (2013) montrent, au moyen d'une simulation, comment cela peut être adapté dans une stratégie efficace de collecte active.

5 Conclusion

Le score au niveau de l'unité proposé est une bonne approximation de l'incidence de l'unité sur la variance due à la non-réponse. Il s'applique à différents plans de sondage, est conforme aux estimateurs de calage pour les totaux de domaine et fonctionne avec de nombreuses méthodes d'imputation courantes. Les hypothèses sur lesquelles repose la décomposition sont généralement valides dans les enquêtes courantes utilisant des méthodes d'imputation sans biais et des estimateurs convergents de paramètres de modèle d'imputation. Les résultats de la simulation montrent que cette méthode gagne en exactitude quand la taille des échantillons augmente. La décomposition de la variance de non-réponse est biaisée en raison de sa non-linéarité. Toutefois, le biais est plus petit dans les populations asymétriques et lorsqu'on se concentre sur un petit nombre d'unités non répondantes. Le fait que l'ordre des unités au moyen de la contribution estimée à la variance due à la non-réponse est semblable à l'ordre réel est un aspect important lorsque la priorité consiste à déterminer les facteurs contribuant le plus à l'erreur totale, et non pas nécessairement à déterminer leurs contributions réelles.

L'article a présenté la méthode dans un contexte univarié, mais on peut facilement l'appliquer à un cadre multivarié, en utilisant une fonction de distance pour combiner les contributions des items dans une contribution d'unité. L'idée consiste toujours à concentrer notre attention sur les traitements de collecte ou la vérification manuelle des cas dans lesquels les scores d'unité sont les plus élevés. Dans ce cas, le traitement du suivi de la non-réponse pourrait différer en cas de non-réponse d'unité ou de non-réponse partielle. Par exemple, un suivi téléphonique pourrait servir à recueillir tous les items pour les unités non répondantes ayant le score le plus élevé; et les non-répondants partiels au score élevé pourraient être envoyés

à un analyste à des fins d'examen, selon le budget alloué au suivi. De plus, si on peut calculer ce score plusieurs fois pendant la période de collecte, le suivi des cas de non-réponse gagnera en efficacité, car le score d'unité sera plus exact et la qualité pourrait alors être satisfaisante pour certaines estimations. Les résultats de simulation montrent que le score proposé est une bonne approximation de la contribution d'une unité à la variance due à la non-réponse. Par la suite, ce score pourrait servir à déterminer le nombre d'unités non répondantes et les unités non répondantes dont il faut assurer le suivi afin d'atteindre un coefficient de variation estimé donné.

À l'origine, ce travail visait à prioriser les unités non-répondantes dans le cadre du processus de plan adaptatif itératif d'estimations en continu pour le PISE. À la suite du plan initial, des estimations d'item clés seraient calculées avec les indicateurs de qualité connexes à plusieurs moments précis de la période de collecte. Après chaque moment fixé, les unités ayant les contributions les plus grandes selon notre méthode seraient classées par ordre de priorité de suivi.

Remerciements

Les auteurs tiennent à remercier les lecteurs critiques (Cynthia Bocci et Jessica Andrews), le rédacteur associé, les examinateurs et le rédacteur adjoint pour leurs précieux commentaires.

Annexe

Preuve 1

$$\sum_{k\in s_m} \delta_k \left(\hat{V}_{\mathrm{DIF}}\left(\hat{t}_d\right)\right) = \sum_{k\in s_m} (1-\pi_k)\, w_k^2 d_k \hat{\sigma}_k^2 = \hat{V}_{\mathrm{DIF}}\left(\hat{t}_d\right).$$

Preuve 2

$$\begin{aligned}
\sum_{k\in s_m} \delta_k \left(\hat{V}_{\mathrm{MIX}}\left(\hat{t}_d\right)\right) &= \sum_{k\in s_m}\left(2\sum_{l\in s_r} w_k d_k \varphi_{lk}\,(w_l-1)\, d_l \hat{\sigma}_l^2 - 2w_k\,(w_k-1)\, d_k \hat{\sigma}_k^2\right) \\
&= 2\sum_{k\in s_m}\sum_{l\in s_r} w_k d_k \varphi_{lk}\,(w_l-1)\, d_l \hat{\sigma}_l^2 - 2\sum_{k\in s_m} w_k\,(w_k-1)\, d_k \hat{\sigma}_k^2 \\
&= 2\sum_{l\in s_r}\sum_{k\in s_m} w_k d_k \varphi_{lk}\,(w_l-1)\, d_l \hat{\sigma}_l^2 - 2\sum_{k\in s_m} w_k\,(w_k-1)\, d_k \hat{\sigma}_k^2 \\
&= 2\sum_{l\in s_r}\left(\sum_{k\in s_m} w_k d_k \varphi_{lk}\right)(w_l-1)\, d_l \hat{\sigma}_l^2 - 2\sum_{k\in s_m} w_k\,(w_k-1)\, d_k \hat{\sigma}_k^2 \\
&= 2\sum_{l\in s_r} W_{dl}\,(w_l-1)\, d_l \hat{\sigma}_l^2 - 2\sum_{k\in s_m} w_k\,(w_k-1)\, d_k \hat{\sigma}_k^2 \\
&= \hat{V}_{\mathrm{MIX}}\left(\hat{t}_d\right).
\end{aligned}$$

Preuve 3

$$
\begin{aligned}
\sum_{k\in s_m} \delta_k \left(\hat{V}_{\mathrm{NR}}\left(\hat{t}_d\right)\right) &= \sum_{k\in s_m}\left(\sum_{l\in s_r}\left(2W_{dl}w_k d_k \varphi_{lk} - w_k^2 d_k \varphi_{lk}^2\right)\hat{\sigma}_l^2 + w_k^2 d_k \hat{\sigma}_k^2\right) \\
&= \sum_{k\in s_m}\left(\sum_{l\in s_r}\left(2W_{dl}w_k d_k \varphi_{lk} - w_k^2 d_k \varphi_{lk}^2\right)\hat{\sigma}_l^2\right) + \sum_{k\in s_m} w_k^2 d_k \hat{\sigma}_k^2 \\
&= \sum_{l\in s_r}\left(\sum_{k\in s_m}\left(2W_{dl}w_k d_k \varphi_{lk} - w_k^2 d_k \varphi_{lk}^2\right)\hat{\sigma}_l^2\right) + \sum_{k\in s_m} w_k^2 d_k \hat{\sigma}_k^2 \\
&= \sum_{l\in s_r}\left(2W_{dl}\sum_{k\in s_m} w_k d_k \varphi_{lk}\hat{\sigma}_l^2 - \sum_{k\in s_m} w_k^2 d_k \varphi_{lk}^2 \hat{\sigma}_l^2\right) + \sum_{k\in s_m} w_k^2 d_k \hat{\sigma}_k^2 \\
&= \sum_{l\in s_r}\left(2W_{dl}^2\hat{\sigma}_l^2 - \sum_{k\in s_m} w_k^2 d_k \varphi_{lk}^2 \hat{\sigma}_l^2\right) + \sum_{k\in s_m} w_k^2 d_k \hat{\sigma}_k^2 \\
&= \sum_{l\in s_r} 2W_{dl}^2\hat{\sigma}_l^2 - \sum_{l\in s_r}\sum_{k\in s_m} w_k^2 d_k \varphi_{lk}^2 \hat{\sigma}_l^2 + \sum_{k\in s_m} w_k^2 d_k \hat{\sigma}_k^2 \\
&= \hat{V}_{\mathrm{NR}}\left(\hat{t}_d\right) + \sum_{l\in s_r} W_{dl}^2\hat{\sigma}_l^2 - \sum_{l\in s_r}\sum_{k\in s_m} w_k^2 d_k \varphi_{lk}^2 \hat{\sigma}_l^2 \\
&= \hat{V}_{\mathrm{NR}}\left(\hat{t}_d\right) + \sum_{l\in s_r}\left(W_{dl}^2 - \sum_{k\in s_m} w_k^2 d_k \varphi_{lk}^2\right)\hat{\sigma}_l^2 \\
&= \hat{V}_{\mathrm{NR}}\left(\hat{t}_d\right) + \sum_{l\in s_r}\left(\left(\sum_{k\in s_m} w_k d_k \varphi_{lk}\right)^2 - \sum_{k\in s_m} w_k^2 d_k \varphi_{lk}^2\right)\hat{\sigma}_l^2 .
\end{aligned}
$$

Bibliographie

Beaumont, J.-F., et Bissonnette, J. (2011). Estimation de la variance sous imputation composite : méthodologie programmée dans le SEVANI. *Techniques d'enquête*, 37, 2, 183-192. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/pub/12-001-x/2011002/article/11605-fra.pdf.

Beaumont, J.-F., et Bocci, C. (2009). Variance estimation when donor imputation is used to fill in missing values. *Canadian Journal of Statistics*, 37, 400-416.

Beaumont, J.-F., Bocci, C. et Haziza, D. (2014). An adaptive data collection procedure for call prioritization. *Journal of Official Statistics*, 30, 607-621.

Beaumont, J.-F., Haziza, D. et Bocci, C. (2011). On variance estimation under auxiliary value imputation in sample surveys. *Statistica Sinica*, 21, 515-537.

Biemer, P.P. (2010). Total survey error: Design, implementation, and evaluation. *Public Opinion Quarterly*, 74, 5, 817-848.

Bosa, K., et Godbout, S. (2014). *IBSP Quality Measures – Methodology Guide*. Division des méthodes d'enquêtes auprès des entreprises. Document interne.

Godbout, S., Beaucage, Y. et Turmelle, C. (2011). Achieving quality and efficiency using a top-down approach in the Canadian integrated business statistics Program. *Proceedings of the Conference of European Statisticians*. United Nations Statistical Commission and Economic Commission for Europe. Work Session on Statistical Data Editing. Ljubljana, Slovénie, 9-11 mai 2011.

Groves, R.M., et Heeringa, S.G. (2006). Responsive design for household surveys: Tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society, Series A*, 169, No. 3, 439-457.

Mills, F., Godbout, S., Bosa, K. et Turmelle, C. (2013). Multivariate selective editing in the integrated business statistics program. *Proceedings of the Joint Statistical Meeting 2013 – Survey Research Methods Section*. Août 2013. Montréal, Canada.

Särndal, C.-E. (1992). Méthodes pour estimer la précision des estimations d'une enquête ayant fait l'objet d'une imputation. *Techniques d'enquête*, 18, 2, 257-268. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/pub/12-001-x/1992002/article/14483-fra.pdf.

Schouten, B., Calinescu, M. et Luiten, A. (2013). Optimiser la qualité de la réponse au moyen de plans de collecte adaptatifs. *Techniques d'enquête*, 39, 1, 33-66. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/pub/12-001-x/2013001/article/11824-fra.pdf.

Statistique Canada (2015). *Aperçu du Programme Intégré de la Statistique des Entreprises*. Statistique Canada, nº 68-515-X au catalogue. Ottawa.

Turmelle, C., Godbout, S. et Bosa, K. (2012). Methodological challenges in the development of Statistics Canada's new integrated business statistics program. *Proceedings of the International Conference on Establishment Surveys IV*. Montréal, Canada.