Composante du produit nº 12-001-X
Division des méthodes d'enquêtes auprès des entreprises

Article

# Estimation de la variance en présence de non-répondants et de strates à tirage complet

par Jun Shao et Katherine J. Thompson

Décembre 2009

Statistique Canada Statistics Canada

Canada

Techniques d'enquête, décembre 2009
Vol. 35, N° 2, pp. 235-246
Statistique Canada, N° 12-001-X au catalogue



# Estimation de la variance en présence de non-répondants et de strates à tirage complet

**Jun Shao et Katherine J. Thompson** [1]

Résumé

Les enquêtes-entreprises sont souvent réalisées selon un plan d'échantillonnage aléatoire simple stratifié à un degré sans remise comportant certaines strates à tirage complet. Bien que l'on recoure habituellement à l'ajustement de la pondération pour traiter la non-réponse totale, la variabilité due à la non-réponse est parfois omise en pratique quand on estime les variances. Cette situation pose surtout problème lorsqu'il existe des strates à tirage complet. Nous élaborons des estimateurs de variance qui sont convergents quand le nombre d'unités échantillonnées est grand dans chaque classe de pondération, en utilisant les méthodes du jackknife, de la linéarisation et du jackknife modifié. Nous commençons par appliquer les estimateurs ainsi obtenus à des données empiriques provenant de l'Annual Capital Expenditures Survey réalisé par le U.S. Census Bureau, puis nous examinons leur performance dans une étude en simulation.

Mots clés : Non-réponse dépendante d'une covariable ; jackknife ; linéarisation ; ajustement par le quotient ; non-réponse uniforme.

## 1. Introduction

De nombreuses enquêtes-entreprises s'appuient sur un plan d'échantillonnage aléatoire simple stratifié à un degré sans remise. Étant donné l'asymétrie des populations échantillonnées, ces plans comprennent généralement des strates à tirage complet et des strates à tirage partiel. Dans ce genre de plan, les taux d'échantillonnage dans les strates à tirage partiel sont habituellement négligeables (par exemple, moins de 20 % dans toutes les strates). Cependant, si l'unité d'échantillonnage finale est une grande entité commerciale, telle qu'une entreprise, la taille de l'univers est nettement plus petite et, souvent, les fractions d'échantillonnage ne devraient pas être ignorées dans le calcul des estimations de variance.

Des cas de non-réponse se produisent dans la plupart des enquêtes. Nous considérons ici les enquêtes où l'on recourt à l'ajustement de la pondération pour corriger la non-réponse. L'erreur d'échantillonnage n'existant pas dans les strates à tirage complet, les formules de variance classiques ne comprennent aucune composante pour ces strates. En revanche, même dans une strate à tirage complet, la présence d'une non-réponse produit une erreur d'estimation qui est souvent une composante appréciable de l'erreur d'estimation totale.

L'objectif du présent exposé est d'élaborer des méthodes d'estimation de la variance qui tiennent compte de l'ajustement de la pondération pour corriger la non-réponse et de l'existence de strates à tirage complet. À la section 2, nous présentons la notation et les hypothèses, puis nous montrons que les estimateurs de variance obtenu par le jackknife et par linéarisation en ignorant la non-réponse dans les strates à tirage complet, qui sont utilisés à l'heure actuelle dans de nombreuses enquêtes, sous-estiment la variance réelle de l'estimation repondérée du total de population. En établissant directement une formule de variance approximative, nous obtenons deux estimateurs de variance convergents. Ces estimateurs sont également convergents s'il existe des strates à tirage partiel avec fraction d'échantillonnage importante. Nous élaborons également un estimateur de variance jackknife modifié tenant compte de la variabilité due à la non-réponse dans les strates à tirage complet.

À la section 3, nous comparons les estimateurs de variance en utilisant des données couvrant une période de cinq ans provenant de l'Annual Capital Expenditures Survey (ACES) réalisée par le U.S. Census Bureau. À la section 4, nous présentons les résultats de simulations effectuées en utilisant une population créée à partir des données de l'ACES de 2003. Nos résultats de simulation montrent que les estimateurs de variance dans lesquels il n'est pas tenu compte des strates à tirage complet présentent un biais négatif important ; les estimateurs de variance convergents dérivés donnent de bons résultats si les échantillons de strate sont tous de grande taille, mais produisent des résultats non convergents autrement ; et l'estimateur de variance jackknife ne tenant compte d'aucune des fractions d'échantillonnage produit des surestimations. Enfin, à la section 5, nous présentons certaines conclusions.

## 2. Principaux résultats

Considérons un échantillon stratifié sans remise d'une population finie contenant $H$ strates. Soit $n_h$ et $N_h$ les tailles d'échantillon et de population de la strate $h$, respectivement, $y_{hj}$ une variable d'intérêt pour laquelle il peut

1. Jun Shao, University of Wisconsin-Madison et U.S. Census Bureau ; Katherine J. Thompson, U.S. Census Bureau. Courriel : shao@stat.wisc.edu.

exister une non-réponse et $x_{hj}$ une covariable qui prend des valeurs positives et ne présente pas de non-réponse, $j$ étant l'indice de l'unité de population et $h$, l'indice de strate. En suivant la trajectoire échantillon-réponse étudiée par Fay (1991), ainsi que par Shao et Steel (1999), nous considérons la population finie comme un recensement avec des valeurs $y$ et $x$ et des non-répondants, c'est-à-dire que chaque unité $j$ dans la strate $h$ de la population finie est associée à un indicateur $I_{hj}$ (= 1 si $y_{hj}$ est un répondant et = 0 si $y_{hj}$ est un non-répondant). Notre échantillon est tiré de cette population finie et, si l'unité $j$ dans la strate $h$ est comprise dans l'échantillon, $y_{hj}$ est un répondant si $I_{hj} = 1$ et un non-répondant si $I_{hj} = 0$.

Soit $E_s$ et $V_s$ l'espérance et la variance, respectivement, sous l'échantillonnage, ainsi que $E_m, V_m$ et $P_m$ l'espérance, la variance et la probabilité, respectivement, sous le modèle $m$ spécifié dans l'une des hypothèses suivantes.

> *Hypothèse M.* Les valeurs de $(y_{hj}, x_{hj}, I_{hj})$ dans la population finie sont produites indépendamment à partir d'un modèle de superpopulation $m$. La population finie est divisée en $P$ sous-populations telles que, dans la sous-population $p$, la probabilité de réponse est $P_m(I_{hj} = 1 \mid y_{hj}, x_{hj}) = P_m(I_{hj} = 1 \mid x_{hj}) > 0$, $E_m(y_{hj} \mid x_{hj}) = \beta_p x_{hj}$, et $V_m(y_{hj} \mid x_{hj}) = \sigma_p^2 x_{hj}$, où $\beta_p$ et $\sigma_p$ sont des paramètres inconnus qui dépendent de $p$.
>
> *Hypothèse P.* La population finie est divisée en $P$ sous-populations telles que, sous un modèle de superpopulation, $P_m(I_{hj} = 1 \mid y_{hj}, x_{hj}) = \pi_p > 0$ est constante dans la sous-population $p$.

La sous-population considérée dans l'hypothèse M ou dans l'hypothèse P est appelée classe de pondération pour la correction de la non-réponse (ou, brièvement, classe de pondération), puisque nous traitons les non-répondants par ajustement des poids dans chaque classe de pondération. (Si l'on procède à une imputation dans chaque sous-population, les sous-populations sont appelées classes d'imputation.) Dans les applications, les classes de pondération peuvent être des strates ou des unions de strates (les strates sont fusionnées quand elles contiennent un nombre insuffisant de répondants), où elles peuvent recouper diverses strates. L'hypothèse M comporte un modèle de prédiction reliant $y_{hj}$ et $x_{hj}$, ainsi qu'un mécanisme de réponse dépendant d'une covariable à l'intérieur de chaque classe de pondération. Le mécanisme de réponse sous l'hypothèse P est le mécanisme de réponse uniforme à l'intérieur de la cellule de pondération, qui est souvent appelé modèle de réponse presque aléatoire. L'hypothèse P est plus forte que l'hypothèse M en ce qui a trait au mécanisme de réponse. Cependant, l'hypothèse M requiert un modèle explicite de la relation entre $y_{hj}$ et $x_{hj}$ à l'intérieur de chaque classe de pondération. Ici, nous supposons que soit l'hypothèse M, soit l'hypothèse P est vérifiée. Les estimateurs qui peuvent être justifiés sous l'hypothèse P sont appelés estimateurs sous « quasi-randomisation » (Oh et Scheuren 1983).

Quand nous étudions la convergence asymptotique des estimateurs, nous considérons le processus limite de $k_p \to \infty$ pour toute sous-population $p$ avec $H$ et $P$ fixes, où $k_p$ est la taille d'échantillon dans la classe de pondération $p$. Si les classes de pondération sont les mêmes que les strates ou les unions de strates, $k_p \to \infty$ équivaut à $n_h \to \infty$ pour toute strate $h$.

Après l'ajustement par le quotient pour corriger la non-réponse, nous considérons l'estimateur qui suit du total des valeurs de $y$ dans la population finie :

$$\hat{Y} = \sum_p \sum_h \sum_{j \in s_h} \left( \frac{\hat{X}_p}{\hat{X}_{pr}} w_{hj} \right) \delta_{phj} I_{hj} y_{hj} = \sum_p \frac{\hat{X}_p}{\hat{X}_{pr}} \hat{Y}_{pr}, \quad (1)$$

où $p$ est l'indice désignant la classe de pondération, $s_h$ est l'échantillon dans la strate $h$, $\delta_{phj}$ est l'indicateur pour la classe de pondération $p$, et $w_{hj}$ est le poids de sondage construit pour l'échantillonnage stratifié,

$$\hat{X}_p = \sum_h \sum_{j \in s_h} w_{hj} \delta_{phj} x_{hj}, \quad \hat{X}_{pr} = \sum_h \sum_{j \in s_h} w_{hj} \delta_{phj} I_{hj} x_{hj},$$

et

$$\hat{Y}_{pr} = \sum_h \sum_{j \in s_h} w_{hj} \delta_{phj} I_{hj} y_{hj}.$$

Dans le cas particulier où les classes de pondération correspondent aux strates,

$$\hat{Y} = \sum_h \frac{\hat{X}_h}{\hat{X}_{hr}} \hat{Y}_{hr}, \quad (2)$$

où

$$\hat{X}_h = \sum_{j \in s_h} w_{hj} x_{hj}, \quad \hat{X}_{hr} = \sum_{j \in s_h} w_{hj} x_{hj} I_{hj},$$

et

$$\hat{Y}_{hr} = \sum_{j \in s_h} w_{hj} y_{hj} I_{hj}.$$

Quand la covariable $x_{hj} \equiv 1$, $\hat{Y}$ est appelée estimateur de comptage. L'estimateur de comptage contrôle les estimations des répondants par rapport aux totaux de population de la base de sondage. Quand les classes de pondération coïncident avec les strates, l'estimateur de comptage utilise les taux de réponse par classe non pondérés, comme le recommandent Vartivarian et Little (2002).

Sous l'hypothèse M ou P,

$$E_m E_s(\hat{Y} - Y) = E_s E_m(\hat{Y} - Y) = 0,$$

où $Y$ est le total de population finie des valeurs de $y$, et la variance totale est donnée par

$$V_{m,s}(\hat{Y} - Y) = E_m[V_s(\hat{Y})] + V_m[E_s(\hat{Y}) - Y].$$

Soit $V_1 = E_m[V_s(\hat{Y})]$ et $V_2 = V_m[E_s(\hat{Y}) - Y]$. Pour estimer $V_1$, il suffit d'estimer la variance d'échantillonnage $V_s(\hat{Y})$. Puisque $\hat{Y}$ défini par (1) est une somme de ratios et que $\hat{X}_p$, $\hat{X}_{pr}$ et $\hat{Y}_{pr}$ sont chacun un total pondéré de variables et d'indicateurs, nous pouvons appliquer l'estimateur de variance jackknife stratifié

$$v_{J1} = \sum_h \left(1 - \frac{n_h}{N_h}\right) \frac{n_h - 1}{n_h} \sum_{j \in s_h} \left(\hat{Y}_{(hj)} - \frac{1}{n_h} \sum_{k \in s_h} \hat{Y}_{(hk)}\right)^2 \quad (3)$$

(voir Wolter 1985 ou Shao et Tu 1995), où $\hat{Y}_{(hj)}$ est l'analogue jackknife de $\hat{Y}$ quand l'unité $j$ est supprimée dans la strate $h$. Notons que les fractions d'échantillonnage sont intégrées dans cette formule. Quand $k_p \to \infty$ pour toutes les classes de pondération, le résultat standard pour les cas de données complètes (voir, par exemple, Krewski et Rao 1981) impliquent que l'estimateur jackknife $v_{J1}$ est convergent pour la variance d'échantillonnage $V_s(\hat{Y})$, sous l'hypothèse M ou P. Puisque $V_1$ est l'espérance de $V_s(\hat{Y})$, $v_{J1}$ est également convergent pour $V_1$ sous certaines conditions mineures.

Puisque, dans (1), la fonction est la somme de ratios et que les données dans les différentes classes de pondération sont indépendantes, il est possible de dériver un estimateur par linéarisation de $V_s(\hat{Y})$ en utilisant le développement en série de Taylor. Quand les classes de pondération coïncident avec les strates, par exemple, $\hat{Y}$ est donné par (2) et est un estimateur par le ratio distinct dont l'estimateur de variance par linéarisation peut être obtenu en utilisant des méthodes standard. Un autre moyen d'établir un estimateur de variance par linéarisation consiste à linéariser l'estimateur $v_{J1}$ (Thompson et Yung 2006). L'estimateur résultant est

$$v_{L1} = \sum_h \frac{n_h}{n_h - 1} \sum_{j \in s_h} \left\{ \sum_p \left[ \frac{\hat{X}_p}{\hat{X}_{pr}} (\bar{e}_{ph} - w_{hj} e_{phj} I_{hj} \delta_{phj}) + \frac{\hat{Y}_{pr}}{\hat{X}_{pr}} (\bar{x}_{ph} - w_{hj} x_{hj} \delta_{phj}) \right] \right\}^2, \quad (4)$$

où $e_{phj} = y_{hj} - (\hat{Y}_{pr}/\hat{X}_{pr}) x_{hj}$, $\bar{e}_{ph} = n_h^{-1} \sum_{j \in s_h} w_{hj} e_{phj} I_{hj} \delta_{phj}$ et $\bar{x}_{ph} = n_h^{-1} \sum_{j \in s_h} w_{hj} x_{hj} \delta_{phj}$. L'estimateur donné en (4) est exactement le même que l'estimateur de variance par linéarisation classique pour l'estimateur par le ratio distinct donné par (2) quand les classes de pondération coïncident avec les strates. Comme $v_{J1}$, $v_{L1}$ est convergent pour $V_1$ quand $k_p \to \infty$ sous l'hypothèse M ou P, ce qui découle du résultat classique pour le cas de données complètes (Krewski and Rao 1981).

Puisque le ratio est une fonction lisse, sous l'hypothèse M ou P,

$$E_s(\hat{Y}) = \sum_p E_s\left(\frac{\hat{X}_p \hat{Y}_{pr}}{\hat{X}_{pr}}\right) \approx \sum_p \frac{E_s(\hat{X}_p) E_s(\hat{Y}_{pr})}{E_s(\hat{X}_{pr})} = \sum_p \frac{X_p Y_{pr}}{X_{pr}},$$

où

$$X_p = \sum_h \sum_{j \in \mathcal{P}_h} \delta_{phj} x_{hj},$$

$$X_{pr} = \sum_h \sum_{j \in \mathcal{P}_h} \delta_{phj} I_{hj} x_{hj},$$

$$Y_{pr} = \sum_h \sum_{j \in \mathcal{P}_h} \delta_{phj} I_{hj} y_{hj},$$

et $\mathcal{P}_h$ est la population finie dans la strate $h$. Posons que $Y_p$ est identique à $X_p$ avec $x_{hj}$ remplacé par $y_{hj}$. Alors

$$V_2 = V_m[E_s(\hat{Y}) - Y] \approx \sum_p V_m\left(\frac{X_p Y_{pr}}{X_{pr}} - Y_p\right).$$

Notons que $V_2$ est petite si le taux de non-réponse est faible ($V_2 = 0$ en l'absence de non-réponse) ou si le modèle sous l'hypothèse M est hautement prédicteur. Si la fraction d'échantillonnage globale, $\sum_h n_h / \sum_h N_h$, converge vers 0, alors $V_2/V_1$ converge vers 0 et, par conséquent, $v_{L1}$ et $v_{J1}$ sont des estimateurs convergents de la variance totale $V_{m,s}(\hat{Y}) = V_1 + V_2 \approx V_1$. Notons que $V_1$ ne contient pas la variation due à la non-réponse émanant des strates à tirage complet. Comme, dans de nombreuses enquêtes, les valeurs de $Y$ provenant des strates à tirage complet sont influentes dans le total $Y$ et qu'il est difficile, dans les applications, de dire à quel point $\sum_h n_h / \sum_h N_h$ doit être petite pour qu'ait lieu la convergence $V_2/V_1 \to 0$, il est nécessaire d'estimer $V_2$.

Sous l'hypothèse M, soit $\tilde{E}_m, \tilde{V}_m$ et $\tilde{C}_m$ l'espérance, la variance et la covariance conditionnelles, respectivement, sachant toutes les valeurs de $x$ et tous les indicateurs de réponse. Puisque

$$\tilde{E}_m\left(\frac{X_p Y_{pr}}{X_{pr}} - Y_p\right) = 0,$$

nous obtenons

$$
\begin{aligned}
&V_m\left(\frac{X_pY_{pr}}{X_{pr}}-Y_p\right)\\
&\quad=E_m\left[\tilde{V}_m\left(\frac{X_pY_{pr}}{X_{pr}}-Y_p\right)\right]+V_m\left[\tilde{E}_m\left(\frac{X_pY_{pr}}{X_{pr}}-Y_p\right)\right]\\
&\quad=E_m\left[\tilde{V}_m\left(\frac{X_pY_{pr}}{X_{pr}}-Y_p\right)\right]\\
&\quad=E_m\left[\frac{X_p^2}{X_{pr}^2}\tilde{V}_m(Y_{pr})-2\frac{X_p}{X_{pr}}\tilde{C}_m(Y_{pr},Y_p)+\tilde{V}_m(Y_p)\right]\\
&\quad=E_m\left[\frac{X_p^2}{X_{pr}^2}\sigma_p^2X_{pr}-2\frac{X_p}{X_{pr}}\sigma_p^2X_{pr}+\sigma_p^2X_p\right]\\
&\quad=\sigma_p^2E_m\left(\frac{X_p^2}{X_{pr}}-X_p\right).
\end{aligned}
$$

Sous l'hypothèse P, soit $V_m^I$ la variance par rapport aux $I_{hj}$. Puisque

$$E_m^I\left(\frac{X_pY_{pr}}{X_{pr}}-Y_p\right)\approx 0,$$

nous obtenons

$$
\begin{aligned}
V_m\left(\frac{X_pY_{pr}}{X_{pr}}-Y_p\right)&\approx E_m\left[V_m^I\left(\frac{X_pY_{pr}}{X_{pr}}-Y_p\right)\right]\\
&\approx E_m\left[\frac{1-\pi_p}{\pi_p}\sum_h\sum_{j\in P_h}\delta_{hj}\left(y_{hj}-\frac{Y_p}{X_p}x_{hj}\right)^2\right]\\
&\approx E_m\left[\left(\frac{X_p^2}{X_{pr}}-X_p\right)S_p^2\right],
\end{aligned}
$$

où

$$S_p^2=\frac{1}{X_{pr}}\sum_h\sum_{j\in\mathcal{P}_h}\delta_{hj}\left(y_{hj}-\frac{Y_p}{X_p}x_{hj}\right)^2.$$

Comme $X_p$ et $X_{pr}$ peuvent être estimés par $\hat{X}_p$ et $\hat{X}_{pr}$, respectivement, pour estimer $V_2$ nous devons uniquement trouver une estimateur de $\sigma_p^2$ ou de $S_p^2$. Sous l'hypothèse M, un estimateur par la régression de $\beta_p$ est $\hat{Y}_{pr}/\hat{X}_{pr}$ et un estimateur convergent de $\sigma_p^2$ fondé sur les résidus de régression est

$$\hat{\sigma}_p^2=\frac{1}{\hat{X}_{pr}}\sum_h\sum_{j\in s_h}\delta_{phj}I_{hj}w_{hj}\left(y_{hj}-\frac{\hat{Y}_{pr}}{\hat{X}_{pr}}x_{hj}\right)^2.$$

Selon la théorie de l'échantillonnage, $\hat{\sigma}_p^2$ est aussi un estimateur convergent de $S_p^2$ sous l'hypothèse P. Donc, sous l'hypothèse M ou P, un estimateur convergent de $V_2$ est donné par

$$v_{L2}=\sum_p\hat{\sigma}_p^2\left(\frac{\hat{X}_p^2}{\hat{X}_{pr}}-\hat{X}_p\right). \tag{5}$$

L'indice inférieur $L$ indique que cet estimateur est basé sur la linéarisation.

Dans certaines applications, $\sum_h n_h/\sum_h N_h$ est négligeable et la non-réponse dans les strates à tirage partiel contribue de manière négligeable à la composante de variance $V_2$, c'est-à-dire

$$V_2\approx V_m\left[\sum_p\left(\frac{X_{cp}Y_{cpr}}{X_{cpr}}-Y_{cp}\right)\right], \tag{6}$$

où l'indice inférieur $c$ désigne les strates à tirage complet,

$$
\begin{aligned}
X_{cp}&=\sum_{h\in\mathcal{C}}\sum_{j\in\mathcal{P}_h}\delta_{phj}x_{hj},\quad X_{cpr}=\sum_{h\in\mathcal{C}}\sum_{j\in\mathcal{P}_h}\delta_{phj}I_{hj}x_{hj},\\
Y_{cp}&=\sum_{h\in\mathcal{C}}\sum_{j\in\mathcal{P}_h}\delta_{phj}y_{hj},\quad Y_{cpr}=\sum_{h\in\mathcal{C}}\sum_{j\in\mathcal{P}_h}\delta_{phj}I_{hj}y_{hj},
\end{aligned}
$$

et $\mathcal{C}$ représente la série d'indices des strates à tirage complet. Un estimateur par le jackknife convergent de $V_2$ peut être obtenu comme il suit. Notons que $X_{cp}, X_{cpr}$ et $Y_{cpr}$ sont des estimateurs, puisque $\mathcal{P}_h=s_h$ pour $h\in\mathcal{C}$, mais que $Y_{cp}$ n'est pas un estimateur à cause de la non-réponse. Donc, nous ne pouvons pas appliquer le jackknife à la fonction $X_{cp}Y_{cpr}/X_{cpr}-Y_{cp}$. Partant du calcul précédent, nous notons, sous l'hypothèse M, que

$$
\begin{aligned}
V_2&\approx E_m\tilde{V}_m\left[\sum_p\left(\frac{X_{cp}Y_{cpr}}{X_{cpr}}-Y_{cp}\right)\right]\\
&=E_m\left[\sum_p\left(1-\frac{X_{cpr}}{X_{cp}}\right)\tilde{V}_m\left(\frac{X_{cp}Y_{cpr}}{X_{cpr}}\right)\right].
\end{aligned}
$$

De même, sous l'hypothèse P, le résultat tient quand $\tilde{V}_m$ est remplacé par $V_m^I$. Par conséquent, nous pouvons appliquer le jackknife à l'estimateur $X_{cp}Y_{cpr}/X_{cpr}$. Soit

$$\bar{Y}=\sum_p\sqrt{1-\frac{X_{cpr}}{X_{cp}}}\left(\frac{X_{cp}Y_{cpr}}{X_{cpr}}\right)$$

et $\breve{Y}_{(hj)}$ l'analogue jackknife de $\breve{Y}$ après la suppression de l'unité $j$ dans $h \in \mathcal{C}$, quand nous traitons $X_{cp}Y_{cpr}/X_{cpr}$ comme des estimateurs. Alors, un estimateur jackknife de $V_2$ est donné par

$$v_{J2} = \sum_{h \in C} \frac{N_h - 1}{N_h} \sum_{j \in \mathcal{P}_h} \left( \breve{Y}_{(hj)} - \frac{1}{N_h} \sum_{k \in P_h} \breve{Y}_{(hk)} \right)^2$$

($n_h = N_h$ et $s_h = \mathcal{P}_h$ quand $h \in \mathcal{C}$). Dans la formule de $\tilde{Y}$, le facteur $\sqrt{1 - X_{cpr}/X_{cp}}$ fait la correction appropriée pour la non-réponse. Sous l'hypothèse P, $X_{cpr}/X_{cp} \approx \pi_p$ est le taux de réponse, qui peut être considéré comme une fraction d'« échantillonnage » pour les strates à tirage complet.

L'estimateur jackknife résultant de la variance totale $V_1 + V_2$ est alors $v_{J1} + v_{J2}$. Puisque $n_h = N_h$ (c'est-à-dire $1 - n_h / N_h = 0$) si la strate $h$ est à tirage complet, il est facile de voir que $v_{J1} + v_{J2}$ est égal à

$$v_J = \sum_h \frac{n_h - 1}{n_h} \sum_{j \in s_h} \left( \tilde{Y}_{(hj)} - \frac{1}{n_h} \sum_{k \in s_h} \tilde{Y}_{(hk)} \right)^2, \qquad (7)$$

où

$$\tilde{Y}_{(hj)} = \begin{cases} \breve{Y}_{(hj)} & \text{si la strate } h \text{ est à tirage complet} \\ \hat{Y}_{(hj)} \sqrt{1 - \dfrac{n_h}{N_h}} & \text{si la strate } h \text{ n'est pas à tirage complet.} \end{cases}$$

Comparativement à l'estimateur de variance jackknife $v_{J1}$ donné en (3), $v_J$ donné en (7) tient compte de la variabilité due à la non-réponse dans les strates à tirage complet, tandis que $v_{J1}$ ne le fait pas. Sous (6) et l'hypothèse M ou P, $v_J$ est convergent.

Enfin, l'estimateur jackknife où sont ignorées toutes les fractions d'échantillonnage est donné par :

$$\tilde{v}_J = \sum_h \frac{n_h - 1}{n_h} \sum_{j \in s_h} \left( \hat{Y}_{(hj)} - \frac{1}{n_h} \sum_{k \in s_h} \hat{Y}_{(hk)} \right)^2. \qquad (8)$$

Cette estimateur semble être prudent, bien qu'il ne soit pas justifié théoriquement.

Brièvement, nous avons les estimateurs de la variance totale $V_{m,s}(\hat{Y})$ qui suivent :

1. l'estimateur jackknife $v_{J1}$ défini en (3), qui produit une sous-estimation quand $V_2/V_1$ n'est pas négligeable ;
2. l'estimateur par linéarisation $v_{L1}$ défini en (4), qui est asymptotiquement équivalent à $v_{J1}$ ;
3. $v_L = v_{L1} + v_{L2}$ avec $v_{L2}$ défini en (5), qui est convergent ;
4. $v_{JL} = v_{J1} + v_{L2}$, qui est asymptotiquement équivalent à $v_L$ ;
5. l'estimateur de variance jackknife $v_J$ défini en (7), qui est convergent quand l'expression (6) est vérifiée ;
6. l'estimateur jackknife $\tilde{v}_J$.

Sous échantillonnage aléatoire simple stratifié et l'hypothèse P, $v_L$ est approximativement le même que l'estimateur de variance obtenu en traitant l'ensemble des répondants comme une phase supplémentaire de l'échantillon aléatoire simple stratifié (c'est-à-dire un plan d'échantillonnage à deux phases) et en appliquant la formule de variance classique (quand $x_{hj} \equiv 1$) ou la formule de variance pour les estimateurs par calage (Kott 1994, Särndal, Swensson et Wretman 1992, et Hidiroglou et Särndal 1998). Cet estimateur de variance n'est toutefois pas convergent quand l'hypothèse P n'est pas vérifiée.

## 3. Comparaisons empiriques

À la présente section, nous appliquons les estimateurs de variance décrits à la section 2 à des données empiriques couvrant une période de cinq ans provenant de la composante de l'employeur de l'ACES présentée à la section 1. La section 3.1 donne le contexte des variables d'analyse, du plan d'échantillonnage et des méthodes d'estimation de l'ACES. La section 3.2 présente les comparaisons empiriques.

### 3.1 Contexte de l'ACES

L'ACES est conçue pour recueillir des données sur la nature et sur le niveau des dépenses en immobilisations des entreprises non agricoles exploitées aux États-Unis. Les répondants déclarent les dépenses en immobilisations, ventilées selon le type (dépenses en bâtiments et dépenses en équipement) pour l'année civile, dans toutes les succursales et divisions, pour toutes les opérations ayant lieu aux États-Unis. L'univers de l'ACES englobe deux sous-populations, à savoir les entreprises employeuses (ACES-1) et les entreprises non employeuses (ACE-2). (Une entreprise non employeuse est une entreprise qui ne possède pas d'employés rémunérés, dont les recettes d'affaires annuelles sont égales ou supérieures à 1 000 $ (1 $ ou plus dans le secteur de la construction) et qui est assujettie aux lois sur l'impôt fédérales. La plupart des non-employeurs sont des travailleurs autonomes exploitant de très petites entreprises non constituées en société qui peuvent ou non être la source principale de revenu du propriétaire.) Divers questionnaires sont envoyés par la poste aux unités échantillonnées selon qu'il s'agit d'une entreprise ACE-1 ou ACE-2. De nouveaux échantillons ACE-1 et ACE-2 sont sélectionnés chaque

année, dans les deux cas selon un plan d'échantillonnage aléatoire simple stratifié sans remise. L'échantillon ACE-1 comprend environ 75 % de l'échantillon de l'ACES (à peu près 46 000 entreprises sélectionnées par année pour l'ACE-1 et 15 000 sélectionnées par année pour l'ACE-2). Dans le plan d'échantillonnage ACE-1, les unités sont stratifiées par catégorie de taille dans chaque branche d'activités figurant dans la base de sondage. Il existe cinq strates ACE-1 distinctes dans chaque branche d'activités, soit une strate à tirage complet (appelée strate 10) et quatre strates à tirage partiel définies selon la taille de l'entreprise dans la branche d'activités (désignées par 2A à 2D, classées de la plus grande à la plus petite dans la branche d'activités), avec environ 500 strates à tirage partiel dans le plan de sondage de chaque année. Les fractions d'échantillonnage dans les strates correspondant aux unités de grande taille dans la branche d'activités (2A) peuvent être assez élevées : la plupart des années, environ 55 % de l'échantillon compris dans les strates 2A est tiré à un taux variant de 0,5 à 1. Dans les strates correspondant aux trois autres catégories de taille dans la branche d'activités, les fractions d'échantillonnage sont habituellement inférieures à 0,20. Les poids de sondage varient de 1 à 1 000, selon la branche d'activités et la catégorie de taille. La composante ACE-2 est nettement moins stratifiée, le nombre de strates selon la catégorie de taille utilisées chaque année variant d'un total de 6 à 8, et les fractions d'échantillonnage étant inférieure à 0,01 dans toutes les strates. Notre analyse empirique est limitée à la composante ACE-1 de l'enquête qui répond à toutes les conditions décrites à la section précédente.

Le programme de l'ACES publie des estimations des totaux et des variations d'une année à l'autre. Ces estimations sont publiées pour l'ensemble de l'échantillon et selon le code de branche d'activités indiqué par les unités répondantes (qui ne correspond pas nécessairement au code de branche d'activités qui figure dans la base de sondage). En cas de non-réponse, les variances sont estimées en utilisant l'estimateur jackknife avec suppression d'un groupe (Kott 2001). Afin de tenir compte de la non-réponse totale, nous utilisons pour la composante ACE-1 la méthode d'ajustement par le quotient présentée à la section 2 en nous servant de données administratives sur la paie comme variable auxiliaire $x$.

Les classes de pondération sont les strates du plan de sondage, à condition qu'il y ait au moins un répondant dans la classe. La fusion des classes est extrêmement rare et est ignorée dans la suite du présent exposé. Des renseignements plus détaillés concernant le plan de sondage, la méthodologie et les limites des données de l'ACES sont disponibles en ligne à http://www. census.gov/csd/ace.

Même si le plan de sondage ACE-1 est relativement typique d'une enquête-entreprise, les données recueillies ne le sont pas. Les petites entreprises déclarent souvent des dépenses en immobilisations nulles qui sont des valeurs légitimes et, par conséquent, la majorité des estimations sont fréquemment obtenues d'après les données des entreprises des strates à tirage complet et des grandes entreprises des strates à tirage partiel (2A). À mesure que se poursuit la ventilation des dépenses en immobilisations, l'incidence des valeurs nulles déclarées (surtout chez les petites entreprises) augmente.

### 3.2 Comparaisons

Afin d'évaluer l'effet qu'a la méthode d'ajustement de la pondération pour corriger la non-réponse totale sur les erreurs-types pour la composante ACE-1, nous avons calculé des estimations de variance en utilisant des données ACE-1 corrigées de la non-réponse totale à l'aide de l'estimateur par le ratio en prenant la paie comme variable auxiliaire, dans quatre branches d'activités pour chacune desquelles les taux d'échantillonnage étaient élevés dans les strates à tirage partiel de grandes entreprises (2A). Les branches d'activités choisies sont représentatives des secteurs représentés dans l'ACES. Ces branches d'activités et leurs codes du Système de classification des industries de l'Amérique du Nord (SCIAN) sont : Extraction de pétrole et de gaz (211100), Extraction de minerais non métalliques (212300), Autres activités diverses de fabrication (339900), et Architecture, génie et services connexes (541300). Dans les tableaux et discussions qui suivent, les branches d'activités sont désignées par leur code du SCIAN.

Au tableau 1, nous comparons les estimations de variance produites en utilisant les données de l'enquête ACE-1 couvrant une période de cinq ans pour trois caractéristiques, à savoir les dépenses totales en immobilisations (total), les immobilisations en bâtiments (bâtiments) et les immobilisations en équipement (équipement). Pour la comparaison, les estimations de la variance sont données sous forme du ratio par rapport à $v_{J1}$ dans le tableau 1. Les totaux estimés sont également inclus. (Il convient de souligner que ces totaux ne sont pas identiques aux estimations publiées, car ils sont calculés en utilisant la classification des branches d'activités qui figure dans la base de sondage et non celle fournie par le répondant.)

Comme prévu, l'estimateur jackknife $v_{J1}$ et l'estimateur jackknife par linéarisation $v_{L1}$ sont très proches pour toutes les variables. Les estimateurs de variance convergents ($v_L$ et $v_{JL}$) produisent tous deux une valeur appréciablement plus grande que leurs analogues jackknife respectifs ($v_{L1}$ et $v_{J1}$). En général, la plupart des dépenses en immobilisations sont déclarées par les entreprises des strates à tirage complet et des strates des grandes entreprises à tirage partiel, de sorte que l'introduction de la composante des non-répondants dans l'estimateur de variance a un effet

appréciable sur l'estimation de la variance. L'estimateur jackknife $v_J$, qui est corrigé de l'effet des strates à tirage complet, se situe généralement entre $v_{J1}$ et $v_{JL}$. Dans certains cas, $v_J$ est égal à $v_{J1}$ ou en est très proche, indiquant que la variabilité due à la non-réponse provient principalement des strates à tirage partiel dont la fraction d'échantillonnage est grande. L'estimation jackknife $\tilde{v}_J$, qui ne tient pas compte des fractions d'échantillonnage, est beaucoup plus grande que n'importe quelle autre estimation.

## 4. Résultats des simulations

Ici, nous présentons une étude en simulation effectuée en nous servant de données modélisées d'après les branches d'activités de l'ACE-1 présentées à la section précédente. À la section 4.1, nous décrivons les conditions de simulation et à la section 4.2, nous présentons et résumons les résultats.

### 4.1 Conditions de simulation

Nous avons modélisé notre population en utilisant les données des répondants recueillies pendant le cycle de collecte de 2003 pour les trois principaux items de dépenses visés par l'enquête (total, bâtiments et équipement). Les données pour la variables auxiliaire (paie) étaient disponibles pour toutes les unités dans la base de sondage. Les données pour la population complète ont été produites au moyen de l'algorithme SIMDAT (Thompson 2000) en choisissant comme cellules de modélisation les strates d'échantillonnage et comme taille de population, la taille originale dans la base de sondage pour chaque cellule. Le tableau 2 donne les fractions d'échantillonnage et les coefficients de corrélation avec la paie pour les données modélisées dans chaque strate.

Dans la simulation, des échantillons aléatoires simples stratifiés ont été tirés de la population produite. Nous examinons les propriétés statistiques des estimateurs de variance décrits à la section 2 appliqués à des échantillons répétés sous les deux mécanismes de réponse qui suivent, appliqués aux données de l'échantillon :

1. le mécanisme de réponse dépendant de la covariable obtenu par application aléatoire des propensions à répondre modélisées d'après les données d'enquête en prenant la paie comme covariable, ce qui produit de très fortes probabilités de réponse pour les grandes unités et de très faibles probabilités pour les petites unités (strates à tirage partiel) ;
2. le mécanisme de réponse uniforme à l'intérieur de la strate obtenu en utilisant les taux de réponse observés durant l'enquête comme probabilité de réponse à l'intérieur de la strate.

En moyenne, les probabilités de réponse par strate individuelle dans la branche d'activités était de 0,85, 0,76. 0,77, 0,76 et 0,68 pour les strates 10, 2A, 2B, 2C et 2D, respectivement.

Nous avons tiré 5 000 échantillons de la population, calculé $\hat{Y}$ donné en (1) à partir de chaque échantillon en présence de non-réponse et d'ajustement des poids, et calculé la moyenne et la variance empirique des 5 000 valeurs de $\hat{Y}$. Nous avons répété l'exercice pour chaque branche d'activités et chaque item, en appliquant deux méthodes d'ajustement, à savoir l'estimateur par le ratio et l'estimateur de comptage. Quand $\hat{Y}$ est l'estimateur par le ratio en utilisant la paie comme variable auxiliaire, la valeur absolue du biais relatif empirique est inférieure à 1,4 % et plus petite que 1 % dans la plupart des cas. Pour l'estimateur de comptage sous le mécanisme de réponse uniforme à l'intérieur de la strate, la valeur absolue du biais relatif empirique est inférieure à 0,5 %. L'estimateur de comptage n'est pas approximativement sans biais en théorie sous le mécanisme de réponse dépendant de la covariable. Toutefois, dans la simulation, la valeur absolue de son biais relatif empirique est inférieure à 1 % dans la plupart des cas et a une valeur maximale de 2,7 %. Nous avons utilisé la variance empirique de 5 000 valeurs de $\hat{Y}$ comme « valeur réelle » de la variance de $\hat{Y}$.

### 4.2 Résultats

Dans 2 000 des 5 000 échantillons, nous avons calculé les six estimations de variance différentes pour les trois items, quatre branches d'activités et deux méthodes d'ajustement des poids. Nous avons examiné les propriétés statistiques de chaque estimateur de variance sur les échantillons répétés en nous servant du biais relatif (BR) défini comme étant

$$\frac{\text{moyenne de 2 000 estimations de variance}}{\text{variance réelle}} - 1,$$

la stabilité (ST) définie comme étant

$$\frac{\sqrt{\text{erreur quadratique moyenne empirique de l'estimation de variance}}}{\text{variance réelle}},$$

et le taux d'erreur (ER) défini comme étant la proportion empirique des intervalles de confiance à 90 % approximatifs $(\hat{Y} \pm 1,645\sqrt{\text{estimation de la variance}})$ obtenus pour 2 000 échantillons qui ne contient pas le total de population réel.

**Tableau 1**
**Estimations de la variance pour $\hat{Y}$ avec ajustement par le quotient dans l'enquête ACE-1**

| Branche d'activités | Item | Année | $\hat{Y}$ | $v_{J1}$ | $\frac{v_{L1}}{v_{J1}}$ | $\frac{v_L}{v_{J1}}$ | $\frac{v_{JL}}{v_{J1}}$ | $\frac{v_J}{v_{J1}}$ | $\frac{\tilde{v}_J}{v_{J1}}$ |
|---|---|---|---|---|---|---|---|---|---|
| 211000 | Total | 2002 | 1,63E+7 | 4,63E+11 | 0,97 | 1,14 | 1,17 | 1,00 | 17,3 |
| | | 2003 | 2,28E+7 | 6,87E+12 | 0,95 | 1,21 | 1,26 | 1,00 | 2,81 |
| | | 2004 | 2,30E+7 | 2,45E+12 | 0,98 | 1,23 | 1,25 | 1,00 | 4,77 |
| | | 2005 | 3,08E+7 | 4,29E+12 | 0,98 | 1,22 | 1,24 | 1,19 | 4,77 |
| | | 2006 | 4,18E+7 | 6,29E+12 | 0,99 | 1,17 | 1,19 | 1,00 | 8,78 |
| | Bâtiments | 2002 | 1,31E+7 | 3,99E+11 | 0,97 | 1,14 | 1,17 | 1,00 | 15,3 |
| | | 2003 | 1,86E+7 | 5,78E+12 | 0,94 | 1,22 | 1,27 | 1,00 | 2,78 |
| | | 2004 | 1,70E+7 | 8,39E+11 | 0,99 | 1,42 | 1,43 | 1,00 | 11,3 |
| | | 2005 | 2,64E+7 | 3,84E+12 | 0,98 | 1,22 | 1,24 | 1,16 | 4,64 |
| | | 2006 | 3,55E+7 | 5,41E+12 | 0,99 | 1,19 | 1,21 | 1,00 | 8,76 |
| | Équipement | 2002 | 3,20E+6 | 6,14E+10 | 0,98 | 1,15 | 1,17 | 1,00 | 7,26 |
| | | 2003 | 4,18E+6 | 8,39E+11 | 0,97 | 1,22 | 1,24 | 1,00 | 1,70 |
| | | 2004 | 6,01E+6 | 1,54E+12 | 0,97 | 1,13 | 1,16 | 1,00 | 1,39 |
| | | 2005 | 4,33E+6 | 1,34E+11 | 0,97 | 1,22 | 1,25 | 1,15 | 6,17 |
| | | 2006 | 6,31E+6 | 7,14E+11 | 0,99 | 1,12 | 1,13 | 1,00 | 2,68 |
| 212300 | Total | 2002 | 1,56E+6 | 4,14E+10 | 0,81 | 1,06 | 1,24 | 1,20 | 3,19 |
| | | 2003 | 1,33E+6 | 1,21E+10 | 0,94 | 1,18 | 1,24 | 1,36 | 5,43 |
| | | 2004 | 2,01E+6 | 2,86E+10 | 0,96 | 1,60 | 1,65 | 2,20 | 6,04 |
| | | 2005 | 1,96E+6 | 1,93E+10 | 0,98 | 1,12 | 1,14 | 2,30 | 6,04 |
| | | 2006 | 2,28E+6 | 2,19E+10 | 0,96 | 1,26 | 1,30 | 3,22 | 11,7 |
| | Bâtiments | 2002 | 2,22E+5 | 4,36E+8 | 1,00 | 1,11 | 1,11 | 1,64 | 8,61 |
| | | 2003 | 1,49E+5 | 2,27E+8 | 0,96 | 1,28 | 1,32 | 1,48 | 7,32 |
| | | 2004 | 4,14E+5 | 1,03E+8 | 0,96 | 46,6 | 46,6 | 75,3 | 426 |
| | | 2005 | 2,23E+5 | 9,33E+8 | 0,99 | 1,12 | 1,13 | 1,32 | 1,88 |
| | | 2006 | 2,20E+5 | 1,88E+9 | 0,97 | 1,20 | 1,22 | 1,19 | 2,29 |
| | Équipement | 2002 | 1,33E+6 | 4,05E+10 | 0,81 | 1,06 | 1,25 | 1,15 | 2,86 |
| | | 2003 | 1,18E+6 | 1,13E+10 | 0,94 | 1,20 | 1,26 | 1,32 | 5,07 |
| | | 2004 | 1,60E+6 | 2,82E+10 | 0,96 | 1,40 | 1,44 | 1,53 | 3,30 |
| | | 2005 | 1,73E+6 | 1,62E+10 | 0,97 | 1,16 | 1,19 | 2,33 | 6,69 |
| | | 2006 | 2,06E+6 | 2,14E+10 | 0,96 | 1,26 | 1,30 | 2,94 | 10,8 |
| 339900 | Total | 2002 | 1,75E+6 | 1,94E+10 | 0,99 | 1,27 | 1,29 | 1,10 | 3,71 |
| | | 2003 | 1,58E+6 | 2,99E+10 | 0,98 | 1,24 | 1,27 | 1,10 | 1,60 |
| | | 2004 | 1,70E+6 | 1,00E+10 | 0,99 | 1,40 | 1,40 | 1,69 | 4,61 |
| | | 2005 | 1,77E+6 | 2,55E+10 | 0,99 | 1,28 | 1,29 | 1,25 | 3,02 |
| | | 2006 | 1,94E+6 | 5,51E+10 | 0,99 | 1,23 | 1,25 | 1,12 | 2,15 |
| | Bâtiments | 2002 | 2,99E+5 | 1,21E+9 | 0,99 | 1,24 | 1,24 | 1,09 | 3,55 |
| | | 2003 | 1,93E+5 | 8,54E+8 | 0,99 | 1,27 | 1,28 | 1,09 | 1,75 |
| | | 2004 | 2,10E+5 | 2,00E+8 | 0,99 | 1,86 | 1,87 | 2,08 | 5,89 |
| | | 2005 | 2,56E+5 | 5,07E+8 | 0,99 | 1,80 | 1,81 | 1,97 | 9,61 |
| | | 2006 | 5,97E+5 | 4,93E+10 | 0,99 | 1,19 | 1,20 | 1,01 | 1,16 |
| | Équipement | 2002 | 1,45E+6 | 1,62E+10 | 0,99 | 1,27 | 1,28 | 1,07 | 3,02 |
| | | 2003 | 1,39E+6 | 2,71E+10 | 0,97 | 1,24 | 1,27 | 1,09 | 1,58 |
| | | 2004 | 1,49E+6 | 9,14E+9 | 0,99 | 1,40 | 1,41 | 1,62 | 4,61 |
| | | 2005 | 1,51E+6 | 2,45E+10 | 0,99 | 1,22 | 1,23 | 1,15 | 2,12 |
| | | 2006 | 1,34E+6 | 5,65E+9 | 0,99 | 1,42 | 1,43 | 1,60 | 6,20 |
| 541300 | Total | 2002 | 3,38E+6 | 2,32E+10 | 0,99 | 1,47 | 1,48 | 1,67 | 5,02 |
| | | 2003 | 3,09E+6 | 2,61E+10 | 0,99 | 1,26 | 1,27 | 1,05 | 1,62 |
| | | 2004 | 3,97E+6 | 1,12E+11 | 1,00 | 1,23 | 1,23 | 1,03 | 1,37 |
| | | 2005 | 4,94E+6 | 2,54E+11 | 1,00 | 1,20 | 1,20 | 1,04 | 1,71 |
| | | 2006 | 4,96E+6 | 2,82E+10 | 1,00 | 1,40 | 1,40 | 1,75 | 8,36 |
| | Bâtiments | 2002 | 7,41E+5 | 6,32E+9 | 1,00 | 1,70 | 1,71 | 1,64 | 7,47 |
| | | 2003 | 4,29E+5 | 3,32E+9 | 1,00 | 1,29 | 1,29 | 1,01 | 1,33 |
| | | 2004 | 6,96E+5 | 4,38E+10 | 1,00 | 1,22 | 1,22 | 1,00 | 1,40 |
| | | 2005 | 7,12E+5 | 9,00E+9 | 1,00 | 1,25 | 1,25 | 1,08 | 2,08 |
| | | 2006 | 8,73E+5 | 3,44E+9 | 1,00 | 1,58 | 1,59 | 1,63 | 9,88 |
| | Équipement | 2002 | 2,96E+6 | 1,39E+10 | 0,99 | 1,37 | 1,38 | 1,54 | 3,95 |
| | | 2003 | 2,66E+6 | 1,94E+10 | 0,99 | 1,25 | 1,26 | 1,05 | 1,59 |
| | | 2004 | 3,27E+6 | 5,83E+10 | 1,00 | 1,22 | 1,23 | 1,04 | 1,29 |
| | | 2005 | 4,23E+6 | 2,40E+11 | 1,00 | 1,19 | 1,20 | 1,03 | 1,59 |
| | | 2006 | 4,09E+6 | 2,35E+10 | 1,00 | 1,27 | 1,28 | 1,49 | 5,47 |

**Tableau 2**
**Caractéristiques de la population pour l'étude par simulation**

| Branche d'activités | Strate | Taille de la population | Fraction d'échantillonnage | Corrélation avec la paie | | |
|---|---|---|---|---|---|---|
| | | | | Total | Branche d'activités | Strate |
| 211000 | 10 | 26 | 1,00 | 0,65 | 0,53 | 0,95 |
| | 2A | 128 | 0,77 | 0,68 | 0,66 | 0,22 |
| | 2B | 372 | 0,11 | 0,57 | 0,51 | 0,51 |
| | 2C | 1 800 | 0,02 | -0,07 | 0,00 | -0,10 |
| | 2D | 10 406 | 0,00 | 0,28 | 0,00 | 0,28 |
| 212300 | 10 | 30 | 1,00 | 0,96 | 0,95 | 0,94 |
| | 2A | 108 | 0,37 | 0,85 | 0,74 | 0,77 |
| | 2B | 414 | 0,07 | 0,03 | 0,76 | -0,03 |
| | 2C | 1 310 | 0,03 | 0,42 | 0,13 | 0,43 |
| | 2D | 4 762 | 0,01 | 0,44 | -0,22 | 0,44 |
| 339900 | 10 | 158 | 1,00 | 0,80 | 0,40 | 0,80 |
| | 2A | 498 | 0,26 | 0,40 | 0,04 | 0,51 |
| | 2B | 2 048 | 0,05 | 0,20 | 0,24 | 0,18 |
| | 2C | 6 310 | 0,02 | 0,19 | 0,48 | 0,09 |
| | 2D | 25 288 | 0,00 | 0,37 | 0,67 | 0,36 |
| 541300 | 10 | 160 | 1,00 | 0,60 | 0,56 | 0,59 |
| | 2A | 959 | 0,38 | 0,20 | 0,39 | 0,06 |
| | 2B | 4 531 | 0,06 | 0,28 | 0,13 | 0,27 |
| | 2C | 17 913 | 0,01 | 0,08 | 0,06 | 0,08 |
| | 2D | 67 440 | 0,00 | 0,13 | -0,01 | 0,15 |

Les tableaux 3 et 4 donnent, respectivement, les résultats de simulation sous les deux mécanismes de réponse. Ces résultats se résument comme il suit.

1. Deux estimateurs de variance ne tenant pas compte de $V_2, v_{J1}$ et $v_{L1}$, possèdent en général un grand biais relatif négatif. Le taux d'erreur des intervalles de confiance connexe est également grand.
2. Deux estimateurs de variance convergents, $v_L$ et $v_{JL}$, produisent des résultats fort semblables et sont généralement nettement meilleurs que $v_{J1}$ et $v_{L1}$ en ce qui concerne le biais relatif et le taux d'erreur des intervalles de confiance connexes.
3. L'estimateur de variance jackknife $v_J$ donne de bons résultats pour les branches d'activités 339900 et 541300, mais peut présenter un grand biais relatif positif dans les branches d'activités 211000 et 212300. Nous pensons qu'il s'agit d'un effet de « petit échantillon », puisque $v_J$ est justifié par la convergence asymptotique et que la taille de la strate à tirage complet dans les branches d'activités 211000 et 212300 est égale à 26 et 30, respectivement (tableau 2). Pour les deux autres branches d'activités, la taille de la strate à tirage complet est de 158 et 160, respectivement. En fait, la performance de $v_L$ et $v_{JL}$ est généralement meilleure pour les branches d'activités 339900 et 541300.
4. Dans certains cas, $v_J$ possède un biais relatif négatif de plus de 10 %, qui est dû au fait que la fraction d'échantillonnage de certaines strates à tirage partiel est grande, autrement dit que l'approximation (6) ne tient pas suffisamment.
5. L'estimateur de variance jackknife $\tilde{v}_J$ dans lequel sont ignorées toutes les fractions d'échantillonnage présente un biais relatif positif très grand et est trop prudent.

## 5. Conclusion

En présence de non-réponse dans les strates à tirage complet (ou dans les strates dont la fraction d'échantillonnage est grande), l'estimateur de variance jackknife et l'estimateur de variance par linéarisation qui ne tiennent pas compte des strates à tirage complet (ou des strates dont la fraction d'échantillonnage est grande) ne sont pas acceptables à cause de leur biais négatif important. Nous dérivons deux estimateurs de variance asymptotiquement sans biais et convergents en ajoutant un terme supplémentaire qui tient compte de la variabilité due à la non-réponse dans les strates à tirage complet (ou dans les strates dont la fraction d'échantillonnage est grande). Nous dérivons également un estimateur jackknife modifié qui est convergent quand les strates à tirage complet sont les seules strates qui contribuent à la variance due à la non-réponse (c'est-à-dire quand l'hypothèse (6) est vérifiée).

Nos résultats de simulation montrent que les trois estimateurs de variance donnent de bons résultats quand les tailles d'échantillon de strate sont grandes et des résultats inconsistants autrement, et que l'estimateur de variance jackknife ne tenant compte d'aucune des fractions d'échantillonnage est très prudent.

Comparativement à la méthode de linéarisation, le jackknife requiert plus de ressources informatiques, mais il offre d'autres avantages. Il est facile à programmer, ne nécessite qu'une seule recette pour divers problèmes et ne requiert pas de calculs compliqués ou distincts pour divers estimateurs. Notre estimateur de variance par linéarisation donné en (4) est en fait obtenu par linéarisation de l'estimateur jackknife donné en (3).

**Tableau 3**
**Résultats des simulations (en %) pour l'estimation de la variance sous mécanisme de réponse dépendant de la covariable**

| Estimation | Item | Branche d'activités | | $v_{J1}$ | $v_{L1}$ | $v_L$ | $v_{JL}$ | $v_J$ | $\tilde{v}_J$ |
|---|---|---|---|---|---|---|---|---|---|
| Ratio | Total | 211000 | BR | -35,8 | -38,1 | -10,3 | -8,0 | 39,1 | 113,9 |
| | | | ST | 49,8 | 50,4 | 47,4 | 48,6 | 252,9 | 182,9 |
| | | | ER | 19,6 | 19,8 | 12,2 | 11,8 | 10,7 | 1,1 |
| | | 212300 | BR | -20,4 | -22,2 | -4,48 | -2,69 | 54,8 | 266,4 |
| | | | ST | 30,3 | 31,1 | 26,8 | 27,3 | 139,1 | 268,8 |
| | | | ER | 12,6 | 12,6 | 9,9 | 9,6 | 6,3 | 0,1 |
| | | 339900 | BR | -21,2 | -22,5 | 0,26 | 1,55 | -5,34 | 52,5 |
| | | | ST | 47,3 | 47,0 | 55,0 | 56,0 | 43,9 | 67,8 |
| | | | ER | 14,3 | 14,6 | 10,4 | 10,3 | 10,0 | 2,6 |
| | | 541300 | BR | -20,7 | -21,0 | 3,83 | 4,08 | -11,6 | 18,4 |
| | | | ST | 32,7 | 32,8 | 34,9 | 35,0 | 29,4 | 32,0 |
| | | | ER | 12,6 | 12,7 | 8,6 | 8,6 | 10,7 | 6,2 |
| | Bâtiments | 211000 | BR | -38,0 | -40,1 | -11,9 | -9,59 | 33,9 | 108,1 |
| | | | ST | 51,3 | 51,9 | 48,5 | 49,6 | 244,4 | 180,8 |
| | | | ER | 20,9 | 21,1 | 12,9 | 12,6 | 11,1 | 1,1 |
| | | 212300 | BR | -23,2 | -23,9 | -12,4 | -11,6 | 33,2 | 341,5 |
| | | | ST | 31,5 | 32,0 | 27,1 | 27,0 | 95,0 | 344,3 |
| | | | ER | 12,3 | 12,3 | 10,4 | 10,3 | 6,9 | 0,1 |
| | | 339900 | BR | -20,0 | -20,4 | -6,31 | -5,88 | -10,9 | 39,8 |
| | | | ST | 42,5 | 42,7 | 42,3 | 42,3 | 39,9 | 64,0 |
| | | | ER | 15,9 | 16,0 | 12,7 | 12,6 | 13,2 | 5,4 |
| | | 541300 | BR | -20,0 | -20,1 | 0,09 | 0,33 | -15,9 | 15,7 |
| | | | ST | 42,6 | 42,5 | 50,5 | 50,7 | 41,1 | 42,7 |
| | | | ER | 13,1 | 13,2 | 9,9 | 9,9 | 12,0 | 6,5 |
| | Équipement | 211000 | BR | -15,0 | -17,3 | 14,1 | 16,4 | -9,37 | 27,9 |
| | | | ST | 63,9 | 62,6 | 87,7 | 90,0 | 64,1 | 69,6 |
| | | | ER | 16,2 | 16,7 | 13,3 | 13,0 | 14,7 | 6,7 |
| | | 212300 | BR | -21,4 | -23,3 | -4,13 | -2,21 | 39,7 | 201,1 |
| | | | ST | 31,7 | 32,5 | 28,4 | 29,0 | 113,7 | 204,4 |
| | | | ER | 13,3 | 13,5 | 10,2 | 10,1 | 7,7 | 0,2 |
| | | 339900 | BR | -21,4 | -22,8 | 1,18 | 2,57 | -7,29 | 50,8 |
| | | | ST | 51,2 | 50,9 | 60,8 | 61,9 | 47,9 | 69,2 |
| | | | ER | 15,5 | 15,8 | 11,6 | 11,4 | 11,0 | 2,3 |
| | | 541300 | BR | -19,7 | -19,9 | 6,16 | 6,43 | -11,9 | 12,8 |
| | | | ST | 33,8 | 33,9 | 38,4 | 38,5 | 31,0 | 30,9 |
| | | | ER | 12,5 | 12,5 | 8,9 | 8,9 | 11,0 | 7,0 |
| Comptage | Total | 211000 | BR | -30,1 | -31,9 | 0,05 | 1,85 | 1,05 | 103,1 |
| | | | ST | 50,4 | 50,5 | 55,9 | 57,3 | 46,7 | 113,4 |
| | | | ER | 15,3 | 15,6 | 9,0 | 8,8 | 8,7 | 1,0 |
| | | 212300 | BR | -33,2 | -34,6 | -6,30 | -4,96 | 17,6 | 204,5 |
| | | | ST | 38,7 | 39,6 | 27,7 | 27,8 | 42,8 | 208,6 |
| | | | ER | 14,1 | 14,7 | 9,1 | 8,7 | 6,9 | 0,4 |
| | | 339900 | BR | -23,9 | -24,6 | 1,73 | 2,44 | -14,2 | 46,4 |
| | | | ST | 47,5 | 47,4 | 55,2 | 55,7 | 43,4 | 62,4 |
| | | | ER | 13,4 | 13,5 | 9,1 | 9,1 | 10,7 | 2,5 |
| | | 541300 | BR | -22,9 | -23,2 | 1,68 | 1,94 | -18,8 | 15,4 |
| | | | ST | 32,9 | 33,0 | 32,0 | 32,2 | 30,2 | 28,9 |
| | | | ER | 11,5 | 11,6 | 7,2 | 7,1 | 10,6 | 5,2 |
| | Bâtiments | 211000 | BR | -30,3 | -32,2 | -0,15 | 1,65 | -1,27 | 101,5 |
| | | | ST | 51,3 | 51,3 | 57,3 | 58,7 | 46,7 | 112,3 |
| | | | ER | 15,8 | 16,3 | 9,6 | 9,4 | 9,2 | 0,8 |
| | | 212300 | BR | -37,4 | -38,0 | -13,5 | -12,9 | 3,53 | 250,2 |
| | | | ST | 41,6 | 42,0 | 28,9 | 28,8 | 32,2 | 254,8 |
| | | | ER | 15,4 | 15,6 | 9,6 | 9,5 | 8,1 | 0,4 |
| | | 339900 | BR | -20,0 | -20,3 | -4,33 | -4,00 | -14,5 | 38,6 |
| | | | ST | 42,3 | 42,4 | 42,4 | 42,4 | 40,1 | 62,8 |
| | | | ER | 14,6 | 14,7 | 11,9 | 11,8 | 13,6 | 5,0 |
| | | 541300 | BR | -20,9 | -21,2 | -0,54 | -0,32 | -18,9 | 14,5 |
| | | | ST | 41,6 | 41,6 | 47,8 | 48,0 | 40,6 | 40,9 |
| | | | ER | 12,5 | 12,5 | 9,1 | 9,1 | 12,1 | 6,0 |
| | Équipement | 211000 | BR | -17,8 | -20,0 | 11,2 | 13,3 | -13,0 | 26,6 |
| | | | ST | 58,9 | 58,0 | 76,4 | 78,4 | 57,7 | 64,1 |
| | | | ER | 15,7 | 15,8 | 12,5 | 12,3 | 14,5 | 6,1 |
| | | 212300 | BR | -30,7 | -32,2 | -4,74 | -3,27 | 12,1 | 164,3 |
| | | | ST | 37,6 | 38,6 | 29,1 | 29,3 | 38,7 | 168,9 |
| | | | ER | 14,1 | 14,5 | 9,6 | 9,5 | 7,9 | 0,6 |
| | | 339900 | BR | -24,1 | -24,9 | 2,52 | 3,27 | -15,2 | 45,0 |
| | | | ST | 51,2 | 51,1 | 61,0 | 61,5 | 47,7 | 64,1 |
| | | | ER | 14,8 | 15,1 | 9,9 | 9,8 | 11,9 | 2,3 |
| | | 541300 | BR | -21,6 | -21,9 | 4,10 | 4,39 | -18,1 | 10,1 |
| | | | ST | 33,6 | 33,7 | 35,2 | 35,3 | 31,5 | 28,2 |
| | | | ER | 11,1 | 11,1 | 7,2 | 7,1 | 10,3 | 5,9 |

**Tableau 4**
**Résultats des simulations (en %) pour l'estimation de la variance sous mécanisme de réponse uniforme à l'intérieur de la strate**

| Estimation | Item | Branche d'activités | | $v_{J1}$ | $v_{L1}$ | $v_L$ | $v_{JL}$ | $v_J$ | $\tilde{v}_J$ |
|---|---|---|---|---|---|---|---|---|---|
| Ratio | Total | 211000 | BR | -49,2 | -50,4 | -17,2 | -16,0 | 89,2 | 138,4 |
| | | | ST | 55,4 | 56,1 | 43,7 | 43,9 | 310,5 | 258,3 |
| | | | ER | 26,9 | 27,1 | 13,9 | 13,8 | 9,10 | 1,80 |
| | | 212300 | BR | -5,42 | -7,99 | 16,1 | 18,7 | 111,7 | 337,2 |
| | | | ST | 28,9 | 28,5 | 37,4 | 39,6 | 179,2 | 341,7 |
| | | | ER | 13,5 | 13,8 | 9,85 | 9,50 | 5,85 | 0,05 |
| | | 339900 | BR | -9,37 | -10,5 | 18,0 | 19,2 | 16,5 | 83,8 |
| | | | ST | 45,8 | 45,4 | 59,0 | 60,1 | 48,4 | 95,6 |
| | | | ER | 14,5 | 14,7 | 9,55 | 9,55 | 8,60 | 2,65 |
| | | 541300 | BR | -8,83 | -9,03 | 18,2 | 18,4 | 6,62 | 44,5 |
| | | | ST | 26,7 | 26,8 | 36,6 | 36,7 | 28,2 | 52,3 |
| | | | ER | 12,6 | 12,6 | 8,45 | 8,45 | 9,70 | 5,35 |
| | Bâtiments | 211000 | BR | -52,6 | -53,7 | -19,2 | -18,0 | 78,4 | 128,0 |
| | | | ST | 58,0 | 58,7 | 45,3 | 45,4 | 290,8 | 248,5 |
| | | | ER | 28,7 | 29,0 | 15,1 | 14,9 | 9,80 | 2,25 |
| | | 212300 | BR | -16,2 | -18,1 | 9,92 | 11,9 | 63,5 | 356,2 |
| | | | ST | 32,0 | 32,4 | 37,1 | 38,5 | 108,6 | 361,9 |
| | | | ER | 15,5 | 16,0 | 10,9 | 10,7 | 6,65 | 0,35 |
| | | 339900 | BR | -13,2 | -13,6 | 13,9 | 14,3 | 1,15 | 54,9 |
| | | | ST | 47,8 | 47,8 | 59,2 | 59,5 | 46,3 | 82,7 |
| | | | ER | 17,2 | 17,2 | 12,6 | 12,5 | 13,8 | 6,40 |
| | | 541300 | BR | -8,9 | -9,2 | 19,2 | 19,5 | -2,22 | 36,0 |
| | | | ST | 39,4 | 39,3 | 53,7 | 54,0 | 38,6 | 55,9 |
| | | | ER | 12,9 | 12,9 | 8,85 | 8,85 | 11,3 | 6,85 |
| | Équipement | 211000 | BR | -1,1 | -2,75 | 27,5 | 29,1 | 12,8 | 60,1 |
| | | | ST | 64,4 | 63,2 | 88,6 | 90,3 | 71,1 | 89,8 |
| | | | ER | 15,3 | 15,6 | 12,0 | 12,0 | 12,7 | 5,10 |
| | | 212300 | BR | -6,3 | -8,96 | 16,8 | 19,4 | 90,0 | 263,1 |
| | | | ST | 30,3 | 29,8 | 39,3 | 41,6 | 148,6 | 269,1 |
| | | | ER | 13,9 | 14,2 | 10,1 | 9,60 | 6,75 | 0,15 |
| | | 339900 | BR | -8,84 | -10,1 | 19,5 | 20,7 | 15,8 | 84,6 |
| | | | ST | 50,8 | 50,3 | 65,7 | 66,9 | 52,9 | 98,8 |
| | | | ER | 15,1 | 15,3 | 10,3 | 10,3 | 9,50 | 2,45 |
| | | 541300 | BR | -6,89 | -7,1 | 19,9 | 20,1 | 6,76 | 38,5 |
| | | | ST | 28,6 | 28,6 | 40,0 | 40,1 | 30,1 | 48,3 |
| | | | ER | 12,4 | 12,4 | 8,60 | 8,55 | 10,3 | 5,90 |
| Comptage | Total | 211000 | BR | -27,8 | -29,0 | 14,2 | 15,4 | 16,3 | 149,5 |
| | | | ST | 47,4 | 47,5 | 53,1 | 54,2 | 44,4 | 158,1 |
| | | | ER | 16,0 | 16,2 | 8,30 | 8,30 | 7,45 | 1,85 |
| | | 212300 | BR | -33,5 | -34,9 | 15,3 | 16,7 | 23,9 | 219,9 |
| | | | ST | 40,0 | 40,9 | 38,5 | 39,5 | 39,4 | 228,5 |
| | | | ER | 18,8 | 19,3 | 9,80 | 9,65 | 8,55 | 1,90 |
| | | 339900 | BR | -16,5 | -17,1 | 20,2 | 20,8 | 4,21 | 75,7 |
| | | | ST | 45,1 | 44,0 | 57,9 | 58,5 | 42,4 | 87,6 |
| | | | ER | 15,6 | 15,8 | 9,40 | 9,35 | 10,8 | 3,20 |
| | | 541300 | BR | -9,61 | -9,81 | 18,9 | 19,1 | -0,77 | 45,0 |
| | | | ST | 26,5 | 26,5 | 36,0 | 36,1 | 24,7 | 52,2 |
| | | | ER | 12,4 | 12,4 | 8,55 | 8,55 | 11,3 | 4,85 |
| | Bâtiments | 211000 | BR | -27,5 | -28,7 | 14,5 | 15,7 | 14,6 | 149,0 |
| | | | ST | 48,1 | 48,1 | 54,5 | 55,6 | 45,1 | 157,6 |
| | | | ER | 17,1 | 17,5 | 9,05 | 9,00 | 8,50 | 2,05 |
| | | 212300 | BR | -39,4 | -40,4 | 11,6 | 12,6 | 10,1 | 238,5 |
| | | | ST | 44,8 | 45,5 | 38,8 | 39,6 | 32,1 | 248,4 |
| | | | ER | 20,2 | 20,7 | 9,95 | 9,85 | 10,3 | 1,80 |
| | | 339900 | BR | -14,2 | -14,6 | 13,6 | 14,0 | -3,55 | 53,5 |
| | | | ST | 47,1 | 47,0 | 57,3 | 57,6 | 45,1 | 80,5 |
| | | | ER | 17,6 | 17,7 | 12,1 | 12,1 | 14,7 | 6,30 |
| | | 541300 | BR | -9,54 | -9,76 | 20,0 | 20,2 | -5,32 | 36,0 |
| | | | ST | 39,1 | 39,0 | 53,3 | 53,5 | 38,3 | 55,8 |
| | | | ER | 12,6 | 12,6 | 9,05 | 9,05 | 11,9 | 6,55 |
| | Équipement | 211000 | BR | -8,12 | -9,64 | 22,7 | 24,2 | 1,56 | 54,3 |
| | | | ST | 58,0 | 57,1 | 76,2 | 77,7 | 57,5 | 82,1 |
| | | | ER | 16,5 | 16,7 | 12,4 | 12,4 | 14,6 | 6,45 |
| | | 212300 | BR | -28,5 | -30,0 | 17,1 | 18,6 | 21,5 | 189,7 |
| | | | ST | 37,6 | 38,4 | 40,9 | 42,0 | 38,2 | 198,9 |
| | | | ER | 18,1 | 18,5 | 9,95 | 9,80 | 9,25 | 1,75 |
| | | 339900 | BR | -15,8 | -16,4 | 21,8 | 22,5 | 4,69 | 76,8 |
| | | | ST | 49,5 | 49,3 | 64,6 | 65,2 | 47,4 | 91,1 |
| | | | ER | 16,4 | 16,5 | 9,45 | 9,40 | 11,4 | 3,20 |
| | | 541300 | BR | -7,53 | -7,74 | 20,2 | 20,4 | 0,26 | 38,8 |
| | | | ST | 28,2 | 28,2 | 39,2 | 39,4 | 27,2 | 48,0 |
| | | | ER | 12,7 | 12,7 | 8,30 | 8,25 | 11,3 | 5,65 |

## Remerciements

Le présent rapport est diffusé afin de tenir les parties intéressées au courant des travaux de recherche et de favoriser les discussions. Les opinions exprimées concernant les questions statistiques et méthodologiques sont celles des auteurs et ne reflètent pas forcément celles du U.S. Census Bureau. Les auteurs remercient deux examinateurs et un rédacteur associé de leurs commentaires et suggestions constructifs, ainsi que Carol Caldwell, Rita Petroni et Mark Sands de leurs commentaires utiles concernant une version antérieure du présent article. Les travaux de recherche de Jun Shao ont été financés en partie par la bourse SES-0705033 de la NSF.

## Bibliographie

Fay, R.E. (1991). A design-based perspective on missing data variance. *Proceedings of the 1991 Annual Research Conference*, U.S. Bureau of the Census, 429-440.

Hidiroglou, M.A., et Särndal, C.-E. (1998). Emploi des données auxiliaires dans l'échantillonnage à deux phases. *Techniques d'enquête*, 24, 11-20.

Kott, P. (1994). A note on handling nonresponse in sample surveys. *Journal of the American Statistical Association*, 89, 693-696.

Kott, P. (2001). The delete-a-group jackknife. *Journal of Official Statistics*, 17, 521-526.

Krewski, D., et Rao, J.N.K. (1981). Inference from stratified samples: Properties of the linearization, jackknife and balanced repeated replication methods. *Annals of Statistics*, 9, 1010-1019.

Oh, H.L., et Scheuren, F.J. (1983). Weighting adjustment of unit nonresponse. *Incomplete Data in Sample Surveys*. New York : Academic Press, 20, 143-184.

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

Shao, J., et Steel, P. (1999). Variance estimation for survey data with composite imputation and nonnegligible sampling fractions. *Journal of the American Statistical Association*, 94, 254-265.

Shao, J., et Tu, D. (1995). *The Jackknife and Bootstrap*. New York : Springer-Verlag.

Thompson, J.R. (2000). *Simulation*: *A Modeler's Approach*. New York : John Wiley & Sons, Inc.

Thompson, K.J., et Yung, W. (2006). To replicate (a weight adjustment procedure) or not to replicate? An analysis of the variance estimation effects of a shortcut procedure using the stratified jackknife. *Proceedings of the Section of Survey Research Methods*, American Statistical Association, 3772-3779.

Vartivarian, S., et Little, R.J. (2002). On the formation of weighting adjustment cells for unit non-response. *Proceedings of the Section of Survey Research Methods*, American Statistical Association, 3553-3558.

Wolter, K.M. (1985). *Introduction to Variance Estimation*. New York : Springer-Verlag.