Statistics Canada   Statistique Canada

Methodology Branch

Household Survey
Methods Division

Direction de la méthodologie

Division des méthodes
d'enquêtes des ménages

Canada

WORKING PAPER
METHODOLOGY BRANCH

VARIANCE ESTIMATION FROM SURVEY DATA UNDER SINGLE VALUE IMPUTATION

HSMD - 2000 - 006E

Hyunshik Lee, Eric Rancourt and Carl-Erik Särndal

Household Survey Methods Division
Statistics Canada

December 2000

The work presented in this paper is the responsibility of the authors and does not necessarily represent the views or policies of Statistics Canada

Variance Estimation from Survey Data Under Single Value Imputation

Hyunshik Lee[1], Eric Rancourt[2] and Carl-Erik Särndal[3]

ABSTRACT

This paper reviews recent contributions to the theory of variance estimation in surveys when single value imputation is used for missing values. Single value imputation, in contrast to multiple imputation implies that a single imputed value is created to take the place of a missing value. A number of contributions to this topic have appeared in the recent literature.

The topic is important since survey nonresponse often reaches high levels, and resources are insufficient to renew contact with respondents or to obtain by other means the desired but missing values. Imputation is then usually resorted to. It is a common practice to use single imputation methods and use ordinary variance estimators as if imputed values were observed. However, this approach could lead to a severe underestimation of the true variance. It is hoped to rectify the situation to the extent possible by providing a review on the topic and useful recommendations.

The paper is developed around three aspects from which this variance estimation problem can be examined: (1) the approach taken to variance estimation; (2) the imputation method(s) used to complete the data set; (3) the sampling design and the prototype estimator used for point estimation. After a theoretical review of the various methods, some empirical results are presented as well as a discussion with recommendations.

KEY WORDS: Bias, Bootstrap, Jackknife, Mean squared error, Model-assisted approach, Two-phase.

[1] Hyunshik Lee, Westat, Inc., USA.
[2] Eric Rancourt, Household Survey Methods Division, Statistics Canada.
[3] Carl-Erik Särndal, Ottawa, Canada.

Estimation de variance à partir de données d'enquêtes sous imputation d'une valeur unique

Hyunshik Lee[1], Eric Rancourt[2] et Carl-Erik Särndal[3]

RÉSUMÉ

Cet article passe en revue les récentes contributions à la théorie de l'estimation de variance dans les enquêtes où l'imputation d'une valeur unique est utilisée pour les données manquantes. L'imputation d'une valeur unique, comparativement à l'imputation multiple, signifie qu'une seule valeur est créée pour remplacer une valeur manquante. Récemment, il y a eu plusieurs contributions à ce domaine dans la littérature.

Ce sujet est important car la non-réponse dans les enquêtes atteint parfois un niveau élevé, et les ressources pour renouer contact avec les répondants ou pour obtenir les valeurs manquantes par d'autres moyens sont insuffisantes. Alors on utilise souvent l'imputation. C'est une pratique courante d'utiliser des méthodes d'imputation d'une valeur unique et d'utiliser les estimateurs de variance habituels comme si les valeurs imputées avaient été fournies par les répondants. Cependant, cette approche peut mener à de sévères sous-estimations de la vraie variance. Nous espérons rectifier la situation le plus possible en fournissant une revue du sujet et en proposant quelques recommandations.

L'article est développé autour de trois aspects à partir desquels on peut examiner le problème d'estimation de la variance : (1) l'approche d'estimation de variance; (2) la (les) méthode(s) d'imputation utilisée(s) pour compléter l'ensemble de données; (3) le plan de sondage et l'estimateur prototype utilisé pour l'estimation ponctuelle. Après une revue théorique des différentes méthodes, on présente quelques résultats empiriques de même qu'une discussion et quelques recommandations.

MOTS CLEFS: Approche assistée d'un modèle, Biais, Bootstrap, Deux phases, Erreur quadratique moyenne, Jackknife.

[1] Hyunshik Lee, Westat, Inc., USA.
[2] Eric Rancourt, Division des méthodes d'enquêtes auprès des ménages, Statistique Canada.
[3] Carl-Erik Särndal, Ottawa, Canada.

1. INTRODUCTION

1.1 Scope of the problem

Imputation is commonly used to fill in substitutes for missing survey data. When this step has been completed, it is also common to treat imputed data as true observations and to use standard variance estimators. However, this approach may lead to severe underestimation of the true variance. This problem was recognized early on and multiple imputation was proposed as a solution in a Bayesian framework (Rubin 1978). Multiple imputation implies that several imputed values are provided for each missing value. This paper is about single value imputation, which implies that a single imputed value is provided for a missing value. Single imputation is widely used, particularly by statistical agencies, for its operational convenience. Therefore, there is a need to provide valid variance estimation techniques for survey data with singly imputed values. In this paper, we focus on the problem of variance estimation of a point estimator of a finite population parameter in the presence of single imputation. Merits and demerits of single and multiple imputation are debated elsewhere (e.g., Fay 1996; Rao 1996; Rubin 1996).

We are interested in estimation of a population parameter θ. To facilitate our discussion, we first introduce some notation. Let s be the probability sample that is selected from the target population $U = \{1, 2, ..., k, ..., N\}$ using a given sampling design under which $p(s)$ is the known probability of realizing the sample s. The inclusion probability of unit k is denoted π_k and $a_k = 1/\pi_k$ is the sampling weight. Let $\hat{\theta}$ denote an estimator of θ that would be appropriate in the ideal case of 100% response. We call it a *prototype estimator* because it will be used to compute an estimate from the imputed data. For example, consider $\theta = Y$, where $Y = \sum_U y_k$ is the population total of the survey variable y taking the value y_k for unit k. (If $A \subseteq U$ is a set of nits, we write simply $\sum_A y_k$ for $\sum_{k\in A} y_k$.)

For example, the prototype could be the Horvitz-Thompson (HT) estimator,

$$\hat{\theta} = \hat{Y}_{\mathrm{HT}} = \sum_s a_k y_k \,. \qquad (1.1.1)$$

Another general and widely used prototype estimator is the generalized regression (GREG) estimator for the total (Särndal, Swensson, and Wretman, 1992). It is given by

$$\hat{\theta} = \hat{Y}_{\mathrm{GR}} = \sum_s a_k g_k y_k \,, \qquad (1.1.2)$$

where

$$g_k = 1 + \lambda_s' \mathbf{x}_k / v_k \qquad (1.1.3)$$

with $\lambda_s' = (\mathbf{X} - \hat{\mathbf{X}}_{\mathrm{HT}})' \mathbf{T}_s^{-1}$, $\mathbf{X} = \sum_U \mathbf{x}_k$, $\hat{\mathbf{X}}_{\mathrm{HT}} = \sum_s a_k \mathbf{x}_k$ and $\mathbf{T}_s = \sum_s a_k \mathbf{x}_k \mathbf{x}_k' / v_k$. The v_k are suitably specified constants. Here g_k, called the g-factor, modifies the sampling weight a_k. For most units, g_k is not far from unity. The function of g_k is to incorporate auxiliary information in the form of the known population total $\mathbf{X}$. The estimator $\hat{Y}_{\mathrm{GR}}$ can be given an interpretation with reference to the linear regression model denoted ς and given by

$$y_k = \mathbf{x}_k' \gamma + \delta_k \,, \qquad (1.1.4)$$

where γ is a κ-dimensional column vector of regression coefficients, $E_{\varsigma}(\delta_k)=0$, $E_{\varsigma}(\delta_k^2)=v_k\sigma^2$, and $E_{\varsigma}(\delta_k\delta_l)=0$ if $k\neq l$. The estimator in (1.1.2) includes many traditional estimators as special cases such as the Horvitz-Thompson estimator with $g_k=1$ for all *k* and the ratio estimator with $g_k=X/\hat{X}_{\mathrm{HT}}$ for all *k*.

Let r be the set of respondents realized from the sample s and let $o=s-r$ be the set of nonrespondents. We consider single imputation and denote by $\hat{y}_k$ the imputed value for unit $k\in o$. The imputation procedure produces a completed data set, $D_I=\{y_k^*:k\in s\}$ where $y_k^*=y_k$ if $k\in r$ (observed values) and $y_k^*=\hat{y}_k$ if $k\in o$ (imputed values). We call D_I the *imputed data* set or simply the *imputed data.*

The usual estimation procedure is to compute the prototype estimator on the imputed data. This results in an *imputed estimator* and it is denoted by $\hat{\theta}_I$. If the Horvitz-Thompson estimator is the prototype, we have the imputed estimator

$$\hat{\theta}_I=\hat{Y}_{HT,I}=\sum\nolimits_s a_k y_k^* \tag{1.1.5}$$

and if the GREG estimator is the prototype,

$$\hat{\theta}_I=\hat{Y}_{GR,I}=\sum\nolimits_s a_k g_k y_k^* . \tag{1.1.6}$$

The total error of the imputed estimator is

$$\hat{\theta}_I-\theta=(\hat{\theta}-\theta)+(\hat{\theta}_I-\hat{\theta}) \tag{1.1.7}$$

The first term on the right hand side is the *sampling error* and the second is the *imputation error*.

A simplistic approach to variance estimation is to take the "usual variance estimator" (the one intended for use with 100% response) and compute it on the imputed data. This approach does not account at all for the imputation error, and often it does not even completely cover the sampling error (see Lee, Rancourt, and Särndal, 1994).

In order to estimate the variance of an imputed estimator correctly, special approaches are needed. In this paper we review various such approaches proposed for this purpose. The imputation variance depends on several factors: the sampling design, the imputation method, the prototype estimator in use, and the response mechanism. These factors are discussed in detail in the following.

1.2 Response mechanism

The approaches proposed for correct estimation of the total variance fall into two broad classes, which differ in the probabilistic set-up governing the sampling and response processes. As in Fay (1991), the two set-ups are depicted in the following diagram:

$$\begin{tikzcd}
\text{Population} \arrow[r, "p"] \arrow[d, "q"] & \text{Complete sample} \arrow[d, "q"] \\
\text{Census with nonrespondents} \arrow[r, "p"] & \text{Sample with nonrespondents}
\end{tikzcd} \tag{1.2.1}$$

where *p* is a known sampling design and *q* is an unknown response mechanism. In order to proceed, one has to make an assumption about the unknown response mechanism.

The upper (*pq*) path is more natural since nonresponse occurs after the sample s is selected (Dalenius 1983). Under this path the response mechanism denoted by $q(r\,|\,s)$ expresses the usually unknown conditional probability that the response set r is realized, given s. This probability may further depend on an auxiliary variable vector $\mathbf{z}_k$ and the survey variable y_k and thus, can be expressed by $q(r\,|\,s) = q(r\,|\,s, \mathbf{z}_s, y_s)$, where $\mathbf{z}_s = \{\mathbf{z}_k : k \in s\}$ is a sample auxiliary data set and $y_s = \{y_k : k \in s\}$. Such auxiliary variables are usually used in imputation if available.

A response mechanism $q(r\,|\,s)$ is said to be *unconfounded* when $q(r\,|\,s, \mathbf{z}_s, y_s) = q(r\,|\,s, \mathbf{z}_s)$ and $\text{Prob}\,(k \in r\,|\,s) > 0$ for all $k \in s$ (see Lee, Rancourt, and Särndal, 1994). Otherwise, it is confounded. As the definition suggests, the response probability can depend on sample auxiliary data but not on sample y-data. The unconfounded (confounded) response mechanism is closely related to the ignorable (nonignorable) response mechanism (Rubin, 1987), in which the response probabilities can depend on observed y-values but not on unobserved y-values.

A stronger one than the unconfounded is the *uniform* mechanism, which occurs when the response probability is constant for all sample units, that is, $\text{Prob}\,(k \in r\,|\,s) = c > 0$ for some constant c and for all $k \in s$. Sometimes it is possible to divide the sample into mutually exclusive and exhaustive classes in such a way that the response probability is believed to be constant within each class. Imputation is carried out class by class assuming the uniform response mechanism within each class and thus, these classes are called imputation classes. This is a special and more restrictive case of the unconfounded mechanism and it will be referred to as the *uniform-within-imputation class* (UWIC) to distinguish from a more general case of the unconfounded mechanism.

Little and Rubin (1987) popularized another terminology, *missing at random* (MAR) response mechanism to describe a "missing-data" mechanism that depends on the auxiliary variable for imputation but not on the y-variable. The opposite is "*not missing at random*" (NMAR). If the mechanism is also independent of the auxiliary variable, it is called *Missing Completely at Random* (MCAR), which is equivalent to the uniform mechanism for practical purposes. The weaker one, the MAR mechanism, is equivalent or so treated in sample surveys to the UWIC when the imputation auxiliary variable is categorical. In general, however, the MAR mechanism is closely related to the unconfounded mechanism.

1.3 Imputation model

There are many imputation methods used in practice. A good survey is provided in Kalton and Kasprzyk (1986) and a recent one for business surveys is given in Kovar and Whitridge (1995), with which our imputation terminology is closely aligned.

Nearly all of the methods used in practice are based on a model even though the model may not be explicitly specified. This model called the *imputation model* (referred to as ξ in this paper) is generally given as

$$y_k = \mathbf{z}'_k \beta + \varepsilon_k, \qquad (1.3.1)$$

where $\mathbf{z}_k$ is a φ-dimensional auxiliary column vector used for imputation, β is a φ-dimensional column vector of regression coefficients, $E_\xi(\varepsilon_k) = 0$ and $E_\xi(\varepsilon_k^2) = c_k \sigma^2$ with suitably specified constants c_k, and $E_\xi(\varepsilon_l \varepsilon_k) = 0$, if $l \neq k$. Note that this model is in general different from the estimation model ς given in (1.1.4) used to formulate the GREG estimator, where the totals of the auxiliary variables x_k are required at the population level as opposed to the case of $\mathbf{z}_k$, which is assumed to be known only at the sample level.

Then, as Kalton and Kasprzyk (1986) pointed out, most imputation methods can be expressed as a form of regression imputation as follows:

$$\hat{y}_k = \mathbf{z}_k' \hat{\mathbf{B}}_r + \hat{e}_k \qquad (1.3.2)$$

where $\hat{\mathbf{B}}_r = \left(\sum_r \omega_k \mathbf{z}_k \mathbf{z}_k' / c_k\right)^{-1} \sum_r \omega_k \mathbf{z}_k y_k / c_k$, with some weight ω_k. Some authors advocate to use the sampling weight w_k for ω_k. The term $\hat{e}_k$ is a sort of error term, which can be set to zero. In this case, the imputation method becomes deterministic regression imputation. If one uses randomly selected estimated residuals ($y_l - \mathbf{z}_l' \hat{\mathbf{B}}_r$) for $\hat{e}_k$, the method becomes random regression imputation. By using a dummy auxiliary variable in (1.3.2), we obtain the class mean imputation or hot deck imputation depending on whether zero or random residuals are used for $\hat{e}_k$. The nearest neighbour imputation method also fits in this expression, where the auxiliary vector of the donor is used instead of the recipient's.

1.4 Variance and Mean Square Error of an imputed estimator

In addition to the sampling variance, imputation generates some, sometimes appreciable, variance. Moreover, the imputed estimator will usually be biased even though the prototype estimator may not, particularly when imputation is carried out under a wrong assumption about the RM. Even under "good imputation," $\hat{\theta}_I$ is not free of bias. We can hope that the bias be small, but since it is non-zero, the mean square error (MSE) is a more relevant indicator of the quality of $\hat{\theta}_I$ than its variance.

We denote, by E_p and V_p, the expectation and variance operators with respect to $p(s)$, and the corresponding operators with respect to $q(r\,|\,s)$ are denoted by E_q and V_q. Here and in the following, $E_q(\cdot)$ is to be interpreted as the conditional expectation $E_q(\cdot\,|\,s)$, and $E_p E_q(\cdot)$ or $E_{pq}(\cdot)$ in short as $E_p E_q(\cdot\,|\,s)$. Using the *pq*-probabilistic path in (1.2.1), the MSE of the imputed estimator $\hat{\theta}_I$ is given by

$$\text{MSE}_{pq}(\hat{\theta}_I) = E_{pq}(\hat{\theta}_I - \theta)^2 = V_p(\hat{\theta}) + E_p V_{cIMP} + E_p(B_c^2) + 2\text{Cov}_p(\hat{\theta}, B_c) \qquad (1.4.1)$$

Here $V_p(\hat{\theta})$ is the variance of the prototype $\hat{\theta}$. The sum of the last three terms of (1.4.1) measures the increase in MSE caused by nonresponse followed by imputation. The first of these involves the *conditional variance* $V_{cIMP} = V_q(\hat{\theta}_I\,|\,s)$; the last two terms contain the *conditional bias*, $B_c = E_q(\hat{\theta}_I\,|\,s) - \hat{\theta}$. The covariance term may be numerically unimportant, but $E_p(B_c^2)$ can represent a large addition to the MSE. If we set $B_c = 0$ for all *s* (which never holds exactly in practice), then (1.4.1) becomes

$$V_{\text{TOT}} = V_{\text{SAM}} + V_{\text{IMP}} \qquad (1.4.2)$$

where $V_{TOT} = V_{pq}(\hat{\theta}_I) = \text{MSE}_{pq}(\hat{\theta}_I)$, $V_{SAM} = V_p(\hat{\theta})$ and $E_p V_{cIMP} = E_p V_q(\hat{\theta}_I\,|\,s)$ are, respectively, the *total variance*, the *sampling variance*, and the *imputation variance* of $\hat{\theta}_I$. In the simulation section (Section 3.2), we evaluate the average performance of the different approaches to variance estimation (presented in Sections 2.1 to 2.6) in relation to the MSE. Comparison with the MSE rather than with the variance is more appropriate because: (i) it gives a reminder that an assumption of zero bias in $\hat{\theta}_I$ (although usually made implicitly by users) is usually untenable; (ii) the MSE is the appropriate indicator of accuracy. In reality, however, there is no choice but to estimate the variance (1.4.2) since the bias cannot be estimated, and this is what is estimated by the approaches that we consider. (The result is often an underestimation of the MSE, as the simulation shows; by contrast, these approaches estimate the variance quite well, even if $\hat{\theta}_I$ is considerably biased.) Formula (1.4.2) represents the total variance as a sum of two components. Estimating the total variance is essential in survey research, but, as explained later, to provide separate estimates of the two components is also important from the survey management point of view.

2. APPROACHES TO VARIANCE ESTIMATION UNDER SINGLE IMPUTATION

Since Särndal (1990), many approaches have been proposed to address the problem of variance estimation for singly imputed data. Earlier, Ford (1983) suggested reimputation for the replication variance estimators under hot-deck imputation, which Burns (1990) used unsuccessfully to address the problem with the jackknife technique. In this section, we review all major approaches found in the literature after 1990.

2.1 The two-phase approach

The probabilistic set-up given in Subsection 1.2 resembles the usual set-up for a two-phase sampling design, where $p(s)$ and $q(r|s)$ respectively correspond to the first and the second phase sampling procedures. The only difference is that in our case the distribution for the second phase is the unknown response mechanism, $q(r|s)$, and we must make some assumption about it in order to proceed with this approach. An often used assumption is that of a uniform response mechanism either throughout the whole population or within subgroups of the population. In the two-phase approach, we need the *pq-expectation*, $E_{pq}(\cdot) = E_p E_q(\cdot|s)$ and the *pq-variance*, $V_{pq}(\cdot)$ to evaluate the bias and the variance of an imputed estimator $\hat{\theta}_I$.

The *pq*-variance of $\hat{\theta}_I$, as given in (1.4.2), can also be written as

$$V_{pq}(\hat{\theta}_I) = V_{\text{TOT}} = V_{\text{SAM}} + E_p(V_{\text{cIMP}}) \qquad (2.1.1)$$

where $V_{\text{SAM}} = V_p(\hat{\theta})$ is the sampling variance and $V_{\text{cIMP}} = E_q[(\hat{\theta}_I - \hat{\theta})^2|s]$ is the conditional imputation variance, given s.

The objective of the two-phase approach is to find estimators, $\hat{V}_{SAM}$ and $\hat{V}_{cIMP}$, of the two variance components in (2.1.1) such that,

$$E_p E_q(\hat{V}_{SAM}) = V_{SAM} \text{ and, for every } s,\ E_q(\hat{V}_{cIMP}) = V_{cIMP}. \qquad (2.1.2)$$

A *pq*-unbiased variance estimator of the total variance V_{TOT} is then obtained by taking $\hat{V}_{TOT} = \hat{V}_{SAM} + \hat{V}_{cIMP}$. The procedure is unbiased with respect to the two phases, that is, $E_p E_q(\hat{V}_{TOT}) = V_{TOT}$.

An advantage of the two-phase approach is that it uses a rather simple probabilistic base involving only two distributions: the sampling design and the response mechanism. A weakness is that the results of the approach depend on how well the assumption made about $q(r|s)$ describes the true, unknown response distribution. The variance estimators will be biased when the assumption does not hold. Another weakness is that it is not apparent how one should deal with certain types of frequently used imputation methods such as nearest neighbour imputation.

Consider the case where the target parameter is the population total of the variable y, that is, $\theta = Y = \sum_U y_k$ and the sampling design is any design with finite sampling weights, $a_k = 1/\pi_k$ and $a_{k\ell} = 1/\pi_{k\ell}$, and the prototype estimator is $\hat{Y}_{GR}$ given in (1.1.2), with a specified auxiliary vector $\mathbf{x}_k$ with known population totals. In the case of full response, the usual variance estimator for $\hat{Y}_{GR}$ is given by

$$\hat{V}(\hat{Y}_{GR}) = \sum_s \sum_s A_{k\ell} g_k (y_k - \mathbf{x}_k' \hat{\mathbf{G}}_s) g_\ell (y_\ell - \mathbf{x}_\ell' \hat{\mathbf{G}}_s) \qquad (2.1.3)$$

with $A_{k\ell} = a_k a_\ell - a_{k\ell}$ and $\hat{\mathbf{G}}_s = \mathbf{T}_s^{-1} \sum_s a_k \mathbf{x}_k y_k / v_k$. The corresponding imputed estimator, $\hat{Y}_{GR.I}$, is obtained by computing (1.1.6) on the imputed data set, that is, $\hat{Y}_{GR.I} = \sum_s a_k g_k y_k^*$. Assuming a uniform nonresponse mechanism throughout the population, the sampling variance of $\hat{Y}_{GR.I}$ can be computed as

$$\hat{V}_{SAM} = \sum_r \sum_r A_{k\ell} D_{k\ell} g_k (y_k - \mathbf{x}_k' \hat{\mathbf{G}}_r) g_\ell (y_\ell - \mathbf{x}_\ell' \hat{\mathbf{G}}_r) \qquad (2.1.4)$$

where $D_{kk} = n/m$ for all $k = \ell$ and $D_{k\ell} = n(n-1)/m(m-1)$ for all $k \neq \ell$, and m is the number of respondents (the size of r), and $\hat{\mathbf{G}}_r = \mathbf{T}_r^{-1} \sum_r a_k \mathbf{x}_k y_k / v_k$ is the analogue of $\hat{\mathbf{G}}_s$, based on the respondents only. The technique is illustrated by an example later.

Note that (2.1.4) respects the restriction that y-values are available only for $k \in r$. But we can use auxiliary data known for the whole sample s to improve on (2.1.4) as we now describe.

Suppose that the imputation vector value $\mathbf{z}_k$ and the auxiliary vector value $\mathbf{x}_k$ are known for every $k \in s$. Denote by $\mathbf{u}_k = (\mathbf{x}_k', \mathbf{z}_k')'$ the combined predictor, after elimination of any variables that may be common to $\mathbf{x}_k$ and $\mathbf{z}_k$. Consider the regression of y_k on $\mathbf{u}_k$ for $k \in s$; let the resulting residuals be $y_k - \mathbf{u}_k' \hat{\mathbf{H}}_s$. (This regression is conceptual only, because y-values are present for $k \in r$ only.) Then $y_k - \mathbf{x}_k' \hat{\mathbf{G}}_s = \mathbf{u}_k' \hat{\mathbf{H}}_s + \tilde{e}_{ks}$, where $\tilde{e}_{ks} = y_k - \mathbf{u}_k' \hat{\mathbf{H}}_s - \mathbf{x}_k' \hat{\mathbf{G}}_s$. Now inserting $y_k - \mathbf{x}_k' \hat{\mathbf{G}}_s = \mathbf{u}_k' \hat{\mathbf{H}}_s + \tilde{e}_{ks}$ into (2.1.3) and developing, we obtain one term that is strengthened by knowing $\mathbf{u}_k$ for the whole sample s. In the rest of expression, we must replace the sums over s by estimating counterparts over r, since y_k is observed only for $k \in r$.

To construct the imputation variance component $\hat{V}_{cIMP}$, consider deterministic regression imputation given by (1.3.2) with $\hat{e}_k = 0$ for all k. Let $\hat{\mathbf{B}}_s$ be the full sample analogue of $\hat{\mathbf{B}}_r$. We can choose ω_k and c_k, so that $\sum_r w_k (y_k - \mathbf{z}_k' \hat{\mathbf{B}}_r) = \sum_s w_k (y_k - \mathbf{z}_k' \hat{\mathbf{B}}_s) = 0$ (for example, by taking $\omega_k = w_k$ and $c_k = \lambda' \mathbf{z}_k$ for any constant vector λ). The imputation error is then given by $\hat{Y}_{GR,I} - \hat{Y}_{GR} = -\sum_o w_k (y_k - \hat{y}_k) = \hat{\mathbf{Z}}_{GR}' (\hat{\mathbf{B}}_r - \hat{\mathbf{B}}_s)$, where $\hat{\mathbf{Z}}_{GR} = \sum_s w_k \mathbf{z}_k$. The problem of estimating the conditional imputation variance $\hat{V}_{cIMP}$ is thus reduced to that of estimating the variance of the regression coefficient $\hat{\mathbf{B}}_r$, given s. Under a UWIC mechanism, this can be done, using Taylor linearization.

The two-phase approach for SRS and with the Horvitz-Thompson estimator as the prototype was first studied by Rao (1990) and it was refined and extended to more complex situations by Rao and Sitter (1995). The procedure that we have described extends their reasoning in two aspects: the sampling design is arbitrary and the prototype is the GREG. The following example illustrates the technique in a simple case.

Example 2.1.1. Consider the following conditions: the sampling design is SRS with n units drawn from N so that $a_k \equiv 1/f = n/N$ for all k; the prototype is the Horvitz-Thompson estimator $\hat{Y} = N\bar{y}_s$ (the special case of $\hat{Y}_{GR}$ with $g_k = 1$ and $x_k = 0$ for all k); response mechanism is uniform throughout the population; ratio imputation is used. Then (2.1.4) becomes

$$\hat{V}_{SAM} = N^2(1/n - 1/N)\ S^2_{yr} \qquad (2.1.5)$$

where $S^2_{yr} = \sum_r (y_k - \bar{y}_r)^2/(m-1)$. Rao (1990) suggested a better alternative, derivable with the procedure outlined above, namely,

$$\hat{V}_{SAM} = N^2(1/n - 1/N)\{\ \hat{B}^2_r\ S^2_{zs} + 2\,\hat{B}_r S_{zer} + S^2_{er}\ \} \qquad (2.1.6)$$

where $S^2_{zs} = \sum_s (z_k - \bar{z}_s)^2/(n-1)$; $S^2_{er} = \sum_r e^2_{kr}/(m-1)$; $S_{zer} = \sum_r z_k e_{kr}/(m-1)$; $e_{kr} = y_k - z_k\hat{B}_r$ with $\hat{B}_r = \sum_r y_k / \sum_r z_k$. Note that S^2_{zs} is computed on the values z_k known for the entire s; all other terms involve y_k-values and, consequently, the sums appearing in these terms must be made over r.

We now derive the imputation variance. The ratio imputed values are $\hat{y}_k = z_k\hat{B}_r$, where we have used $\omega_k = N/n$ and $c_k = z_k$. The imputation error is then

$$-(N/n)\sum_o (y_k - \hat{y}_k) = (N/n)(\sum_s z_k)(\hat{B}_r - \hat{B}_s)$$

with $\hat{B}_s = \sum_s y_k / \sum_s z_k$. Consequently, given s the conditional imputation variance is $V_{cIMP} = N^2(1/m - 1/n)\sum_s e^2_{ks}/(n-1)$ with $e_{ks} = y_k - z_k\hat{B}_s$, which leads to the imputation variance estimator

$$\hat{V}_{cIMP} = N^2(1/m - 1/n)\ (\bar{z}_s/\bar{z}_r)^2 S^2_{er} \qquad (2.1.7)$$

where $S^2_{er} = \sum_r e^2_{kr}/(m-1)$ with $e_{kr} = y_k - z_k\hat{B}_r$. Standard sampling theory recommends to include the factor $(\bar{z}_s/\bar{z}_r)^2$ but without any numerical consequence, it can be replaced by unity to obtain the formula given in Rao (1990). The estimated total variance is then $\hat{V}_{TOT} = \hat{V}_{SAM} + \hat{V}_{cIMP}$, where $\hat{V}_{SAM}$ is given by (2.1.5), or preferably by (2.1.6), and $\hat{V}_{cIMP}$ is given by (2.1.7).

2.2 The model-assisted approach

The probabilistic base for the model-assisted approach consists of three distributions: the sampling design $p(s)$, the response mechanism $q(r|s)$, and the imputation model ξ given by (1.3.1). The approach involving these three distributions is called *model-assisted.*

In this setting, the appropriate variance concept is the anticipated pq-variance, or the ξpq-variance, of the imputed estimator $\hat{\theta}_I$, denoted by $E_\xi V_{pq}(\hat{\theta}_I)$. Taking the model expected value of both sides of (2.1.1), we obtain

$$E_\xi V_{pq}(\hat{\theta}_I) = E_\xi V_{TOT} = E_\xi V_{SAM} + E_\xi E_p(V_{cIMP}) \qquad (2.2.1)$$

As in the two-phase approach, we seek estimators of the two components of the total variance, V_{SAM} and $E_p(V_{cIMP})$. The model serves as an instrument in deriving the component estimators, $\hat{V}_{SAM}$ and $\hat{V}_{cIMP}$, such that,

$$E_\xi\left\{E_p E_q(\hat{V}_{SAM}) - V_{SAM}\right\} = 0; \text{ and for every } s,\ E_\xi\left\{E_q(\hat{V}_{cIMP}) - V_{cIMP}\right\} = 0. \qquad (2.2.2)$$

That is, we have $E_\xi\{E_p E_q(\hat{V}_{TOT}) - V_{TOT}\} = 0$. Then a ξpq-unbiased estimator of the total variance is obtained by taking $\hat{V}_{TOT} = \hat{V}_{SAM} + \hat{V}_{cIMP}$.

Note that (2.2.2) is the model-assisted analogue of (2.1.2). It is assumed here that the response mechanism is unconfounded (but otherwise unknown). This assumption allows changing the order of the operators, $E_\xi E_p E_q$ into $E_p E_q E_\xi$, and back, without affecting the value of the expectation. We construct $\hat{V}_{cIMP}$ in such a way that $E_\xi(\hat{V}_{cIMP}) = V_{IMP\xi}$, where $V_{IMP\xi} = E_\xi[(\hat{\theta}_I - \hat{\theta})^2 | s, r]$ is the imputation variance under the model, given s and r. Such a choice satisfies the second part of equation (2.2.2) since $E_\xi(V_{\mathrm{cIMP}}) = E_q(V_{\mathrm{IMP}\xi})$. Although our notation, $\hat{V}_{SAM}$ and $\hat{V}_{cIMP}$, is the same as in the two-phase approach, the computed variance component estimates will usually be different between these approaches.

To derive $\hat{V}_{cIMP}$, we need a model unbiased estimator $\hat{\sigma}^2$ of the unknown σ^2 in (1.3.1). This is because the σ^2 will usually factor out when we take the model expectation $V_{IMP\xi} = E_\xi[(\hat{\theta}_I - \hat{\theta})^2 | s, r]$. If $\hat{\sigma}^2$ satisfies $E_\xi(\hat{\sigma}^2 | s, r) = \sigma^2$, we can thus obtain a $\hat{V}_{cIMP}$ with the required property that $E_\xi(\hat{V}_{cIMP}) = V_{IMP\xi}$, for any fixed s and r.

In the following, we show how the variance component estimators $\hat{V}_{SAM}$ and $\hat{V}_{cIMP}$ are constructed, when the target parameter is the population total, $Y = \sum_U y_k$.

Example 2.2.1. Let the conditions be as in Example 2.1.1, except that we relax the assumption about the response mechanism. We now assume it to be unconfounded, which is weaker than uniform. As derived in Särndal (1990, 1992), the estimated sampling variance is

$$\hat{V}_{\mathrm{SAM}} = N^2\left(\frac{1}{n} - \frac{1}{N}\right)\left(S^2_{y\bullet s} + c_0\hat{\sigma}^2\right) \quad \text{with} \quad S^2_{y\bullet s} = \sum_s \left(y^\bullet_k - \bar{y}^\bullet_s\right)^2 / (n-1) \qquad (2.2.3)$$

where $\bar{y}^\bullet_s = \sum_s y^\bullet_k / n$, c_0 is a constant defined in terms of z_k, of which a close approximation is given by $c_0 \approx (1 - m/n)\bar{z}_o$ with $\bar{z}_o = \sum_o z_k / (n - m)$, and $\hat{\sigma}^2 = C_r \sum_r e^2_{kr} / \sum_r z_k$ with $C_r = \{m/(m-1)\}\{1 - S^2_{zr}/(m\bar{z}^2_r)\}^{-1}$ and $e_{kr} = y_k - z_k\hat{B}_r$. The imputation variance estimator is given by

$$\hat{V}_{\mathrm{cIMP}} = N^2(1/m - 1/n)(\bar{z}_o\bar{z}_s / \bar{z}_r)\hat{\sigma}^2. \qquad (2.2.4)$$

Note that the term $N^2(1/n - 1/N)S^2_{y\bullet s}$ in $\hat{V}_{\mathrm{SAM}}$ is the standard variance estimator (normally used for the prototype estimator) applied to the imputed data. The correction $N^2(1/n - 1/N)c_0\hat{\sigma}^2$ is needed since there is not enough variability in the imputed data.

An alternative to adding a corrective term is an "amended data approach." It entails changing the imputed data so that they will contain sufficient variability to give a "corrective level" when the standard formula is computed on the the amended imputed data. For deterministic regression imputation, this can be achieved by adding a randomly selected residual to the imputed value. This procedure can be implemented within imputation classes. Stochastic imputation methods usually do not require this type of amendment to estimate the sampling variance by the standard formula. The same is true for nearest neighbour imputation.

The amendment approach becomes especially important in statistical agencies that use a modern software package for variance calculation. Examples are Statistics Canada's GES (see Esteveo, Hidiroglou, and Särndal, 1995) and Statistics Sweden's CLAN (see Anderson and Nordberg, 1994). These contain a "standard formula," designed to give variance estimates at a correct level for the prototype estimator in the case of 100% response. Applying the existing software to the amended imputed data will then give an essentially correct estimate of the sampling variance except perhaps for very high rates of nonresponse.

This approach was first proposed by Särndal (1990, 1992). It was studied by Deville and Särndal (1991, 1994) for the regression imputed Horvitz-Thompson estimator, by Gagnon et al. (1996) for the imputed GREG estimator, by Rancourt, Särndal, and Lee (1994) for the nearest neighbour imputed Horvitz-Thompson estimator.

For example, for the nearest neighbour imputed GREG estimator, we get under the ratio imputation model the following estimator of the imputation variance:

$$\hat{V}_{\text{cIMP}} = \left(\sum_{o} w_k^2 z_k + \sum_{\ell \in r} S_\ell^2 z_\ell \right) \hat{\sigma}^2 \qquad (2.2.5)$$

where $S_\ell = \sum_{k \in o_\ell} a_k g_k$ with $o_\ell = \{k : k \in o \text{ and } k \text{ uses } \ell \text{ as donor}\}$. Here $\hat{\sigma}^2$ is as in Example 2.2.1. It can be seen from this expression that multiple utilization of the same donor has a tendency to increase the imputation variance.

It is an advantage of the model-assisted approach that even a relatively complex imputation method such as nearest neighbour imputation is easy to handle (whereas it is unclear how the two-phase approach would work for nearest neighbour imputation). A weakness of the model-assisted approach is its sensitivity to the imputation model assumptions.

2.3 Replication approach: Jackknife and BRR

This section discusses the approaches based on replication techniques: jackknife and balanced repeated replication (BRR). The adjustments required to apply these techniques to the imputed data are explained. The linearized version of the jackknife is also presented.

2.3.1 Jackknife technique

The jackknife technique is a replication approach designed to obtain variance estimates without having to derive a closed expression (see Wolter, 1985). A set of replicates is created by removing one unit or a set of units at a time from the full sample and replicate estimates are computed by applying the estimator $\hat{\theta}$ to each replicate. A variance estimate of $\hat{\theta}$ is obtained by computing the variance among the replicate estimates. The technique is illustrated in the following with a stratified single-stage sampling design where sampling is done with replacement. We assume that jackknife replicates are created by removing one unit (rather than a set of units) at a time. In this case, a replicate can be identified by the unit removed. Let there be H strata and n_h units selected from stratum h. In this case, L replicates can be created, where $L = \sum_{h=1}^{H} n_h$. Then the jackknife variance estimator is given by

$$\hat{V}_J = \sum_{h=1}^{H} \frac{n_h - 1}{n_h} \sum_{j \in s_h} (\hat{\theta}^{(j)} - \hat{\theta})^2 \qquad (2.3.1)$$

where $\hat{\theta}^{(j)}$ is the replicate estimate analogously calculated as $\hat{\theta}$ using the replicate created by removing unit j from the h-th stratum sample s_h, $h = 1, 2, 3, \ldots, H$.

When the jackknife technique is naively applied to imputed data, the variance is underestimated. Burns (1990) tried to correct the underestimation for hot-deck imputation using re-imputation. In this procedure, the missing values are re-imputed by the hot-deck method within each replicate using the respondent data in the replicate and the replicate estimate is computed using the re-imputed replicate data. However, he found that the re-imputation led to an over-estimation of the variance. Rao and Shao (1992) proved this theoretically and proposed an adjustment approach as an alternative. Rao (1992) extended the approach to mean and ratio imputation. The technique given in Rao and Shao (1992), Zanutto (1993) and Rao (1996) is described below.

The basic principle is that when deleted unit j in a given replicate is a respondent, the imputed values in the replicate are adjusted. Otherwise, they remain unchanged. If we let $a_k^{(j)}$ denote the adjustment, then the adjusted value for $k \in s_h$, $y_k^{*(a,j)}$ is given by

$$y_k^{*(a,j)} = \begin{cases} y_k & \text{if } k \in r_h \\ \hat{y}_k + a_k^{(j)} & \text{if } k \in o_h \text{ and } j \in r_h \\ \hat{y}_k & \text{if } k \in o_h \text{ and } j \in o_h \end{cases} \qquad (2.3.2)$$

where r_h and o_h denote the respondent and nonrespondent sets within the stratum, respectively. The adjustment is defined by $a_k^{(j)} = E_I^{(j)}(\hat{y}_k) - E_I(\hat{y}_k)$, where $E_I^{(j)}$ is the expectation with respect to the imputation procedure applied in replicate j and E_I is that for the full sample. For deterministic imputation, $E_I(\hat{y}_k) = \hat{y}_k$ and thus the adjusted value is equal to $E_I^{(j)}(\hat{y}_k)$, which is the re-imputed value. The adjusted jackknife variance estimator for the imputed data is then obtained by applying the jackknife variance estimator to the adjusted replicates, namely,

$$\hat{V}_J^{*} = \sum_{h=1}^{H} \frac{n_h - 1}{n_h} \sum_{j \in s_h} (\hat{\theta}_I^{(aj)} - \bar{\hat{\theta}}_I^{(a.)})^2. \qquad (2.3.3)$$

where $\hat{\theta}_I^{(aj)}$ is the imputed estimator computed using the adjusted replicate and $\bar{\hat{\theta}}_I^{(a.)}$ is the mean of $\hat{\theta}_I^{(aj)}$'s. We may use $\hat{\theta}_I$ instead.

Different imputation methods require different adjustments. However, according to the model given by (1.3.1), the adjustment can be expressed in a unified form as follows:

$$a_k^{(j)} = \mathbf{z}_k' \hat{\mathbf{B}}_r^{(j)} - \mathbf{z}_k \hat{\mathbf{B}}_r \text{ for } k \in o \qquad (2.3.4)$$

and $\hat{\mathbf{B}}_r^{(j)}$ is computed for replicate j analogously as $\hat{\mathbf{B}}_r$. The adjustment is calculated within each imputation class separately but for ease of notation the imputation class indicator is suppressed. Note that a deterministic imputation method and its stochastic counterpart use the same adjustment (e.g., mean imputation and hot deck imputation). For nearest neighbour imputation, donor's $\mathbf{z}_k$ should be used in the recipient's place.

The approach was studied by Rao and Shao (1992) for mean and hot-deck imputation, by Rao and Sitter (1992) for ratio imputation, and by Sitter and Rao (1997) for ratio imputation in case the imputation auxiliary variable is not available for the full sample. For nearest neighbour imputation, Kovar and Chen (1994) used the same adjustment appropriate for ratio imputation, while Rancourt (1999) proposed the one above.

Rao and Sitter (1992) developed a linearized jackknife variance formula for the Horvitz-Thompson estimator as the prototype under SRSWOR as given in the following:

$$\hat{V}_{\text{LINJ}} = N^2\left(\frac{1}{n}-\frac{1}{N}\right)\left(\hat{B}_r^2 S_{zs}^2 + 2\frac{\bar{z}_s}{\bar{z}_r}\hat{B}_r S_{zer}\right) + N^2\left(\frac{1}{m}-\frac{1}{N}\right)\left(\frac{\bar{z}_s}{\bar{z}_r}\right)^2 S_{er}^2 \qquad (2.3.5)$$

where S_{zs}^2, S_{zer}, and S_{er}^2 are defined as in (2.1.6). The authors note that the formula can be viewed as a linearized variance estimator and in this form, it can be used directly, without replication.

In the linearized form given in (2.3.5), the finite population correction (fpc) for without-replacement sampling is applied. However, the application of fpc is not trivial for the adjusted jackknife variance estimator given in (2.3.3). The formula is valid for with-replacement sampling or when the fpc is negligible. When it is appreciable, the variance estimator can overestimate substantially. However, if the fpc is applied naively, the variance estimator underestimates because the fpc application reduces not only the sampling variance but also the imputation variance, which does not need the fpc correction. One option may be to use the correction proposed by Lee, Rancourt and Särndal (1995) where the jackknife is replaced by $\hat{V}_J^{**} = \hat{V}_J^{*} - \sum_{h=1}^{H} N_h \hat{S}_{yU_h}^2$, where $\hat{S}_{yU_h}^2$ is an unbiased estimator of $S_{yU_h}^2$ such as $S_{yr_h}^2$ under uniform response mechanism within each stratum. Also of note is an approach by Steel and Fay (1995), where two nearest neighbours along with a model are used to obtain a corrected formula.

The versatility of the jackknife variance estimator is an advantage, which has been exploited by Skinner and Rao (1993) to extend the jackknife technique to multivariate statistics in the presence of imputation. It is also evidenced in Rancourt, Lee and Särndal (1994), who applied the jackknife technique for situations where more than one imputation method is used within a given data set. Further, Rao (1993) presented the jackknife technique for complex sample designs as a tool that can be implemented into an estimation system such as Statistics Canada's Generalized Estimation System. A comprehensive account of the jackknife approach for imputed data is provided in Rao (1996).

2.3.2 Balanced repeated replication (BRR)

The BRR is another variance estimation approach that uses replication. It was originally proposed by McCarthy (1969) for a special case of the stratified sampling design, where two first-stage units are selected per stratum (i.e., $n_h = 2$, $h =$ 1, 2,..., H). A replicate is created by choosing one unit from each stratum and thus it results in a half-sample. A set of L replicates (half-samples) is said to be balanced if pairs of two units from two different strata appear in the same frequency in the set. There are 2^H possible half-samples and this set is balanced. However, a much smaller number of half-samples ($H+1 \le L \le H+4$) can satisfy this condition and this set of balanced half-sample is used. Generalizing the same idea for a general case with $n_h \ge 2$, a set of B balanced replicates can be constructed by choosing ℓ_h ($\le n_h/2$) units from each stratum but its construction is more difficult than the special case. For a more detailed discussion, see Wolter (1985), and also see Gupta and Nigam (1987), Gurney and Jewett (1975), Sitter (1993), and Wu (1991), Rao and Shao (1996).

Now consider a prototype estimator $\hat{\theta} = \sum_{h=1}^{H}\sum_{i \in s_h} w_{hi} y_{hi}$. From each balanced replicate, replicate estimate is computed by $\hat{\theta}^{(\ell)} = \sum_{h=1}^{H}\sum_{i \in s_h} w_{hi}^{(\ell)} y_{hi}$ using appropriately modified weights $w_{hi}^{(\ell)}$, $\ell = 1, 2, \ldots, L$. Then the BRR variance estimator is given by

$$\hat{V}_{\text{BRR}}(\hat{\theta}) = \frac{1}{L}\sum_{\ell=1}^{L}(\hat{\theta}^{(\ell)} - \hat{\theta})^2 \qquad (2.3.6)$$

If, within stratum, sampling is with replacement, this variance estimator is consistent. Otherwise, it will be somewhat conservative but the bias will be small when the sampling fractions are small.

Turning to the variance estimation problem for imputed data, when the approach is naively applied to the imputed data, it underestimates the variance. Shao, Chen, and Chen (1998) proposed an adjusted BRR approach to correct this problem by using similar adjustment as for the adjusted jackknife.

The adjustment proposed by Shao, Chen, and Chen (1998) under the uniform response mechanism is

$$y_{hi}^{*(a\ell)} = \begin{cases} y_{hi} & \text{if } y_{hi} \text{ is observed} \\ \hat{y}_{hi} + E_I^{(\ell)}(\hat{y}_{hi}) - E_I(\hat{y}_{hi}) & \text{if } y_{hi} \text{ is imputed} \end{cases} \qquad (2.3.7)$$

where E_I is the expectation under the imputation procedure for the full sample and $E_I^{(\ell)}$ is that for ℓ th replicate. This adjustment is similar to the one used for the adjusted jackknife approach described in the previous subsection.

For deterministic imputation methods, $E_I(\hat{y}_{hij}) = \hat{y}_{hij}$ and $E_I^{(\ell)}(\hat{y}_{hij})$ is the re-imputed value (using the same imputation procedure) within the ℓ th replicate. In this case, the adjusted imputed value becomes the re-imputed value. For example, with the ratio imputation method, the adjusted imputed value is equal to the re-imputed value given by $\hat{y}_{hi}^{(a\ell)} = \hat{B}_r^{(\ell)} x_{hi}$ where $\hat{B}_r^{(\ell)} = \sum_r w_{hij}^{(\ell)} y_{hij} \big/ \sum_r w_{hij}^{(\ell)} x_{hij}$.

For stochastic imputation methods, we can often compute the adjustment by a closed expression. For instance, the adjustment for the weighted hot deck imputation method (Rao and Shao, 1992) is given by

$$y_{hi}^{*(a\ell)} = \begin{cases} y_{hi} & \text{if } y_{hi} \text{ is observed} \\ \hat{y}_{hi} + \dfrac{\sum_r w_{hi}^{(\ell)} y_{hi}}{\sum_r w_{hi}^{(\ell)}} - \dfrac{\sum_r w_{hi} y_{hi}}{\sum_r w_{hi}} & \text{if } y_{hi} \text{ is imputed.} \end{cases} \qquad (2.3.8)$$

Using the adjusted imputed data, the adjusted replicate estimate $\hat{\theta}_I^{(a\ell)} = \sum_{h=1}^{H}\sum_{i\in s_h} w_{hi}^{(\ell)} y_{hi}^{*(a\ell)}$ is obtained and then the adjusted BRR variance estimator is computed as

$$\hat{V}_{\text{ABRR}}(\hat{\theta}_I) = \frac{1}{L}\sum_{\ell=1}^{L}(\hat{\theta}_I^{(a\ell)} - \hat{\theta}_I)^2 \, . \qquad (2.3.9)$$

The approach can be applied to a UWIC response mechanism by computing the adjustments within each imputation class separately.

For the prototype estimator considered above, the adjusted replicate estimate can be written as

$$\hat{\theta}_I^{(a\ell)} = \sum_{h=1}^{H}\sum_{i\in s_h} w_{hi}^{(\ell)} y_{hi}^{*(a\ell)} = \hat{\theta}_I^{(\ell)} + \delta^{(\ell)} \qquad (2.3.10)$$

where $\delta^{(\ell)} = \sum_{h=1}^{H}\sum_{i\in r_h} w_{hi}^{(\ell)}\{E_I^{(\ell)}(\hat{y}_{hi}) - E_I(\hat{y}_{hi})\}$. Then the adjusted BRR (ABRR) can be decomposed as

$$\hat{V}_{\text{ABRR}} = \frac{1}{L}\sum_{\ell=1}^{L}(\hat{\theta}_I^{(a\ell)} - \hat{\theta}_I)^2 = \frac{1}{L}\sum_{\ell=1}^{L}(\hat{\theta}_I^{(\ell)} - \hat{\theta}_I)^2 + \frac{1}{L}\sum_{\ell=1}^{L}(\delta^{(\ell)})^2 + \frac{2}{L}\sum_{\ell=1}^{L}(\hat{\theta}_I^{(\ell)} - \hat{\theta}_I)\delta^{(\ell)} . \quad (2.3.11)$$

The first term on the left is the standard BRR estimator applied to the imputed data. The second term is a variance due to imputation. The third term is the covariance between the first component and the second component. As discussed in Shao, Chen, and Chen (1998), this decomposition can be viewed from the multiple imputation perspective. The first term corresponds to the "within imputation variance" of the multiple imputation variance estimator and the second term corresponds to the "between imputation variance." The third term is missing from the multiple imputation variance estimator in the case of improper imputation (see also Kott, 1995). Since the ABRR captures all these terms, the approach can be applied to a multiply imputed data set whether the imputation procedure is proper or not. It is conjectured that this is the case for all other valid variance estimator for singly imputed data.

It was shown by Chen (1993) that the two replication approaches, jackknife and BRR, are asymptotically equivalent up to the order of $n^{-3/2}$. The two approaches also require about the same amount of computation and thus the choice between the two approaches depends on the ease of implementation or availability of software. The adjusted jackknife approach, however, does not work well for sample quintiles for which the adjusted BRR can be applied (see the paper for details).

2.4 Resampling approach: Bootstrap

Another popular but computer-intensive variance estimation approach is the bootstrap, which was originally proposed for non-survey sampling cases (i.e., independently and identically distributed cases) by Efron (1979).

The basic principle of the bootstrap approach mimics the sampling behavior of the prototype estimator $\hat{\theta}$ by simulating the conditional sampling behavior of the bootstrap prototype estimator $\hat{\theta}^*$. From the estimated population distribution based on sample s, a number (say, L) of bootstrap samples s_l^* ($l = 1, \ldots, L$) are generated and then $\hat{\theta}^{*(l)}$ is computed using the bootstrap sample s_l^*. The variance of $\hat{\theta}$ is then estimated by the bootstrap variance

$$V_{\text{BOOT}}(\hat{\theta}) = \frac{1}{L}\sum_{\ell=1}^{L}(\hat{\theta}^{*(\ell)} - \bar{\hat{\theta}}^*)^2 , \quad (2.4.1)$$

where $\bar{\hat{\theta}}^*$ is the average of L bootstrap estimates $\hat{\theta}^{*(\ell)}$.

When the standard bootstrap procedure is applied to survey data, the variance estimator given in (2.4.1) can be inconsistent (see Shao and Tu, 1995, pp. 246-247). The procedure has been adopted for various sampling designs typically used in surveys. These include the with-replacement bootstrap described by McCarthy and Snowden (1985), the re-scaling method of Rao and Wu (1988), the mirror-match bootstrap proposed by Sitter (1992a), and the without-replacement bootstrap of Sitter (1992b), which is used in the simulation.

The Bootstrap variance estimator also underestimates the variance when applied to imputed data. Shao and Sitter (1996) proposed a way to correct the underestimation. The basic idea is to use re-imputation for each bootstrap sample applying the same imputation procedure used for the original sample. The steps of the proposed procedure are given below.

1) Draw a bootstrap sample using the imputed data D_I as normally done in the case of no missing survey data. This bootstrap sample is a mix of observed values (the bootstrap response set, which is denoted by r^*) and imputed values (the bootstrap nonresponse set denoted by o^*);

2) Treat imputed values in the bootstrap sample as missing and impute them using the same procedure used to produce the original imputed data D_I but using r^*. Let the completed bootstrap sample be denoted by D_I^*;
3) Calculate the bootstrap estimate $\hat{\theta}_l^*$ by applying $\hat{\theta}$ to D_I^*;
4) Repeat above steps (1-3) L times and compute the bootstrap variance estimator using $\hat{\theta}_l^*$'s.

Assuming that the response indicator is a population characteristic, Shao and Sitter (1996) showed the consistency of the modified bootstrap variance estimator for commonly used imputation methods under complex sample designs. It boasts of generality of the approach irrespective of the sampling design, the imputation method, and the type of point estimate $\hat{\theta}$. However, a drawback is its huge computational burden. The bootstrap approach is already a computer-intensive method. Performing imputation for every bootstrap sample increases the computational burden even more. One way of reducing this burden is to use an adjustment, instead of performing re-imputation for each bootstrap sample. The adjustment is added to the imputed value $\hat{y}_k^*$ as follows:

$$\hat{y}_k^* + E_I^{(\ell)}(\hat{y}_k^*) - E_I(\hat{y}_k^*) \qquad (2.4.2)$$

where E_I is the expectation under the imputation procedure and $E_I^{(\ell)}$ is the expectation of the same imputation procedure but performed using the ℓ-th bootstrap sample. This modified values are used in variance estimation (only) instead of the original imputed values. The adjustment resembles the one used for the adjusted jackknife and the adjusted BRR, and enables us to avoid re-imputation for every bootstrap sample. In fact, for deterministic imputation, the adjusted bootstrap imputed value is exactly the same as the re-imputed value, where no random number generation is necessary. For stochastic imputations, the adjusted bootstrap imputed value can be calculated without random number generation. The resulting variance estimator is still asymptotically valid for estimating totals and a function of totals. However, the adjusted bootstrap cannot handle the case of quantile estimation although the original approach can. Other ways of reducing the computational burden under particular cases are discussed more fully in Shao and Sitter (1996).

2.5 All cases imputation approach

The All Cases Imputation (ACI) approach was proposed by Montaquila and Jernigan (1997). The idea of the approach is to apply the imputation method to the respondents and then the imputation variance is estimated using both imputed and reported values of the respondents. The sampling variance is estimated by directly applying the ordinary variance estimator to the original imputed data used for point estimation.

The imputation variance is estimated using the residuals $\hat{e}_k = \hat{y}_k - y_k$, $k \in r$. For the case where the Horvitz-Thompson estimator is the prototype, the sample design is simple random sampling, and missing data are imputed by hot-deck imputation, the imputation variance is estimated by

$$\hat{V}_{\text{IMP}} = \frac{N^2 l}{n^2(m-1)} \sum_r (\hat{e}_k - \bar{\hat{e}}_r)^2 + \frac{2N^2 l(l-1)}{n^2 m(m-1)} \sum_{k \in r} \sum_{\substack{i,j \in r \\ j>i}} I_{k(i,j)} (\hat{e}_i - \bar{\hat{e}}_r)(\hat{e}_j - \bar{\hat{e}}_r), \quad (2.5.1)$$

where m is the number of respondents, $l = n - m$, $\bar{\hat{e}}_r = \sum_r \hat{e}_k / m$, and $I_{k(i,j)}$ is equal to 1 if respondent k is the donor for both respondents i and j in all case imputation, and is equal to 0 otherwise. The second term will be zero if donors are selected independently. The formula is valid under the uniform response mechanism with $S^2_{y\bullet s}$ defined in (2.2.3). The ACI variance estimator is then given by

$$N^2(1/n-1/N)S^2_{y\bullet s}+\hat{V}_{\mathrm{IMP}}. \qquad (2.5.2)$$

The ACI approach is straightforwardly extended to the stratified simple random sampling design for the stratified sample mean prototype estimator as shown in Montaquila and Jernigan (1997) if imputation classes coincide with strata. This estimator for random regression imputation was studied by Krenzke, Mohadjer and Montaquila (1998).

The approach implicitly assumes that the ordinary variance estimator is unbiased for the sampling variance $V_{\mathrm{SAM}} = V_p(\hat{\theta}_I)$. This can hold when the nonresponse mechanism is UWIC and a stochastic imputation method such as hot-deck is used. It works well also for nearest neighbour imputation as demonstrated in the simulation study shown later. On the other hand, under deterministic imputation methods such as mean, ratio and regression, it underestimates the variance. The underestimation can be corrected by using amended imputed data explained in Section 2.2 to estimate the sampling variance by an ordinary variance estimator.

Montaquila and Jernigan (1997) indicated that the approach can be extended to more complicated situations in which a more complex sample design and a nonlinear prototype estimator are used.

2.6 Other approaches

Some authors used the *qp*-path depicted in (1.2.1) to formulate variance estimation procedure for imputed data. The order of sampling and response mechanism is reversed in this path and the response mechanism is no longer conditional on the sample.

Tollefson and Fuller (1992) used this path to derive a variance estimator for the Horvitz-Thompson estimator with hot-deck imputed data. They assumed the MAR response mechanism under a superpopulation structure.

Shao and Steel (1999) also used the *qp*-path for a complicated problem. Their approach is reviewed in more detail in the following section.

2.6.1 Approach by Shao and Steel

Shao and Steel (1999) was motivated by the fact that some imputation used in practice is composite in the sense that more than one imputation method is used and/or imputed values are in turn used for imputation of other variables. The situation becomes more complicated when the sampling fraction is not negligible. Some approaches can more easily handle the latter situation (e.g., the model-assisted, two-phase, and ACI) and some other approaches are more adaptable for composite imputation (e.g., the replication approaches). However, these approaches are difficult to apply when both conditions hold.

Assuming that the imputed estimator is (nearly) unbiased, the total variance is given by

$$V(\hat{\theta}_I-\theta)=E_qV_p(\hat{\theta}_I)+V_qE_p(\hat{\theta}_I-\theta) \qquad (2.6.1)$$

Note that E_q and V_q are defined at the population level and no longer conditional on sample s. The authors particularly considered the Horvitz-Thompson estimator for the population total assuming a stratified multi-stage sampling. That is, $\hat{\theta}_I=\hat{Y}_I=\sum_s a_k y_k^{\bullet}$.

Let there be J variables (denoted by u_j) involved in imputation to obtain imputed data set D_I from the y-variable. Then, using response indicator variable ι_j for each variable u_j, the imputed estimator can be expressed as a smooth known function of estimated totals. In the case of deterministic imputation, $\hat{Y}_I = \varphi(\hat{\mathbf{T}})$ where $\hat{\mathbf{T}} = \left(\sum_s a_k \iota_{1k} u_{1k}, ..., \sum_s a_k \iota_{Jk} u_{Jk}\right)$ and φ is known. Let

$$V_1 \equiv V_p(\hat{Y}_I) = V_p[\varphi(\hat{\mathbf{T}})]. \qquad (2.6.2)$$

After conditioning on a set of respondents for variable u_j defined for the whole population, the response indicator ι_{jk} is fixed for every unit and can be treated as a population characteristic. The conditional variance given in (2.6.2) can be estimated using the usual variance estimation methods such as the Taylor, jackknife, BRR, etc. In the case of random imputation, the random component should be included by writing $\hat{Y}_I = \varphi(\hat{\mathbf{T}}) + T^*$ where $\varphi(\hat{\mathbf{T}}) = E_*(\hat{Y}_I)$, $T^* = \hat{Y}_I - E_*(\hat{Y}_I)$, and E_* is the expectation with respect to the random imputation. Hence,

$$V_1 = V_p E_*[\varphi(\hat{\mathbf{T}})] + E_p V_*(T^*) = V_p[\varphi(\hat{\mathbf{T}})] + E_p V_*(T^*) \qquad (2.6.3)$$

The first term on the right can be estimated as before. The second term can also be estimated according to the particular imputation method employed. Let an estimate of V_1 in (2.6.2) or in (2.6.3) be denoted as v_1. Note that if the fpc is non-negligible, it can easily be incorporated in the estimation of v_1.

$E_p(\hat{Y}_I - Y)$ in the second term of (2.6.1), regardless whether the imputation method is random or deterministic, can be written as $E_p(\hat{Y}_I) - Y \approx \varphi[E_p(\hat{\mathbf{T}})] - Y = \phi(\tilde{\mathbf{T}})$, $\tilde{\mathbf{T}} = \left(Y, \sum_U \iota_{1k} u_{1k}, ..., \sum_U \iota_{Jk} u_{Jk}\right)$ for a smooth function ϕ. Then its variance with respect to q can be written using the Taylor expansion as

$$V_2 \equiv V_q\left[E_p(\hat{Y}_I) - Y\right] \approx V_q\left[\phi(\tilde{\mathbf{T}})\right] = \left[\nabla\phi(E_q\tilde{\mathbf{T}})\right]' C\left[\nabla\phi(E_q\tilde{\mathbf{T}})\right] \qquad (2.6.4)$$

where $\nabla\phi$ denotes a partial derivative of a vector variable and C is a $(J+1)\times(J+1)$ matrix of covariances of components of ϕ. The expression in (2.6.4) is a population quantity, which can be estimated by substituting the population values by estimates. The evaluation and estimation of C is easier under the design-based approach than under other approaches. Let an estimate of V_2 in (2.6.4) be denoted as v_2.

The resulting variance estimator of the total variance is given by $v_1 + v_2$. Its asymptotic unbiasedness and consistency under the UWIC mechanism can be easily established since the two variance components (V_1 and V_2) are smooth functions of estimated totals.

Applying the approach to a simple case in which the sample design is SRSWOR and missing values are imputed by ratio imputation. We obtain the following variance estimation formula for the Horvitz-Thompson estimator using Taylor linearization technique:

$$\hat{V}_{\text{POP}} = N^2\left(\frac{1}{n} - \frac{1}{N}\right)S_{us}^2 + \frac{Nn(m-1)}{m}\left(\frac{1}{m} - \frac{1}{n}\right)\left(\frac{\bar{z}_s}{\bar{z}_r}\right)^2\left(S_{er}^2\right) \qquad (2.6.5)$$

where $S_{us}^2 = \sum_s (u_k - \bar{u})^2/(n-1)$, $u_k = \hat{B}_r z_k + (n\bar{z}_s / m\bar{z}_r)\iota_k e_{kr}$, and ι_k is the response indicator variable. Other terms in the formula are defined as in Example 2.1.1.

This approach will be referred to as the POP approach later since it is based on a population level response mechanism. It has some important advantages: (1) it can handle complicated situations as long as the imputed estimator can be expressed as a smooth function of estimated totals; (2) the v_1 term is robust since it does not depend on the response mechanism or the model; (3) V_2/V_1 is $O(n/N)$ where n/N is the overall sampling fraction. Therefore, if the overall sampling fraction is small, the second term can be ignored, which compensates the difficulty involved in getting v_2; and (4) incorporation of the sampling fractions in v_1 is straightforward.

3. COMPARATIVE DISCUSSION OF THE DIFFERENT APPROACHES

All the variance estimation approaches presented in Section 2 provide a valid variance estimate by taking into account the variance due to imputation. However, they have different characteristics and some are more advantageous than others for a specific situation. In this section, we present a qualitative comparison of the approaches with respect to various aspects that should be considered for variance estimation.

3.1 Theoretical comparisons

Variance components In (1.4.2), the total variance was decomposed into two main components: the sampling variance and the variance due to imputation. Not all approaches can provide these components separately. The model-assisted approach provides the components rather naturally since the variance formula is written in the form of (1.4.2). However, replication and re-sampling approaches (jackknife, BRR, and bootstrap) do not provide them. All other approaches (two-phase, all-case-imputation (ACI), linearized jackknife), except the POP approach can provide the two components as well. The POP approach gives different variance components whose interpretation is different.

Minimization of the sampling variance is one of the key goals of designing an efficient survey. However, the ultimate goal should be to reduce the total variance (that is, including the imputation variance not only the sampling variance). If the imputation variance is a significant portion of the total variance, then one should pay more attention to this variance component. Therefore, from the survey management point of view, it is necessary to estimate the two variance components separately. It is then possible to allocate the survey resources in a cost efficient manner. Not many authors realize this very important point (see Gagnon et al., (1997) for more discussion).

Computational burden Even though computers are always becoming more powerful, the computational burden is still an important consideration in many situations. For example, it arises when a data analyst has limited computing power or has to handle a large data set with a large number of variables. Approaches based on explicit formulae such as the model-assisted, two-phase, linearized jackknife, and POP are less burdensome in computation. Replication approaches (jackknife and BRR) are more computer-intensive but the bootstrap approach is the most computer-intensive since re-imputation for each bootstrap sample adds more burdens. As noted earlier, Shao and Sitter (1996) proposed an approach for reducing the computational burden.

Some approaches (jackknife, BRR, and bootstrap) require or may use adjusted imputed values for variance estimation. The ACI approach requires imputation for respondents. All these add computational burden. If we order the approaches in term of computational burden from low to high, we get:

(Two-phase, model-assisted, linearized jackknife, POP) → ACI → jackknife → BRR → bootstrap

A parallel can be made between the replication approaches and the multiple imputation approach. The replication approaches do not have much computational advantage over multiple imputation unless the number of multiple imputations is large, but they have the important advantage that they can be applied to situations where imputation is not proper (see Rubin 1987 for a definition of proper imputation). In fact, any single imputation variance estimator can be used for multiply-imputed data sets regardless whether the imputation method is proper or not. This was pointed out by Shao, Chen and Chen (1998) for BRR but applies for any other valid single imputation variance estimator.

Information needs for variance estimation Besides the usual information needed for the standard variance estimators, all approaches require various pieces of information at the time of variance estimation. Most of all, an imputation flag is necessary to indicate the response/nonresponse status of all records in that data file. This crucial link between imputation and estimation is pointed out and discussed in Rancourt (1996). Information on the imputation method and the imputation auxiliary variables is also necessary. Information on donors (through a record identifier) in the case of nearest neighbour imputation is needed as well. If imputation classes are used, information on the classes must be available.

All the approaches discussed in this chapter usually require the information described above; some need more and others need less. For example, the ACI approach does not require information on donors. The resampling and replication approaches (bootstrap, BRR, jackknife) may or may not require information on donors depending on the imputation method and the way in which the approach is applied. The jackknife approach requires it for nearest neighbour imputation but does not for hot-deck or other random imputation methods. The BRR and bootstrap approaches do require it also if the same adjustment as for the jackknife is used for nearest neighbour imputation but do not if re-imputation is used. There have been some attempts to provide a valid variance estimator that does not require an imputation flag (see, Shao and Sitter, 1996). However, to the best of our knowledge, no such method has yet been made available for single imputation. Therefore, it is important to pass all the required information to the variance estimator when single imputation is used.

Adaptability Any approach discussed here is based on some assumptions about various factors that affect the variance. These factors include the sample design, prototype estimator, parameter of interest, imputation method, and response mechanism. It would be very helpful to know which approaches are available for a particular situation. We focus on the first three factors. Other factors are discussed under separate headings.

In general, replication or resampling approaches are more flexible to different prototype estimators. Most prototype estimators can be expressed as a smooth function of estimated totals by Horvitz-Thompson type of estimators and those approaches can handle fairly easily the variance estimation of such estimators. However, the jackknife technique is more sensitive to the smoothness of the function than the BRR and bootstrap approaches. For example, the jackknife approach does not work well for estimation of quintiles, while the other two approaches do. The latter two (BRR and bootstrap) can also handle more easily a complicated situation where more than one imputation method is used for a single variable. This flexibility comes from the re-imputation principle used in the approaches. However, for these approaches, including the jackknife (except for simple cases), it is not always clear how to incorporate the finite population correction (fpc) correctly and thus they can be seriously biased when the fpc is non-negligible. This difficulty is due to their inability to decompose the total variance into the sampling and imputation variances. Note that the fpc must be applied to the sampling variance component only. The sample design can be quite general (i.e., stratified multi-stage sampling) as long as the first stage sampling is with replacement or of a negligible fpc.

For all other approaches, a variance formula needs to be worked out for each prototype estimator with respect to the sample design, the imputation method employed, and the assumed response mechanism. This was done for simpler situations. For more complex situations, the derivation of the formula can be quite involved. On the other hand, these approaches can incorporate the fpc easily.

In this discussion, we have assumed that any value involved in imputation is observed. However, imputed values are often used for imputation of other variables. This is referred to as composite imputation. Currently, only the POP approach specifically addresses this situation. However, it seems that the BRR and bootstrap approaches should be able to handle this case as well, at least in principle if the composite imputation procedure can be replicated and the fpc is negligible. If the fpc is not negligible, the BRR and bootstrap approaches are biased and need a correction. The POP approach can handle both situations.

Response mechanism All approaches described in Section 2 have been designed to work at least for the uniform (MCAR) case. In fact, the two-phase approach was originally developed under this mechanism, but we show that this condition can be relaxed to the uniform-within-imputation-class (UWIC) mechanism. In fact, all approaches are applicable under one form of the UWIC mechanism. The ACI approach, however, uses a more restrictive form of the UWIC, where imputation classes are defined within the stratum boundaries. All other approaches except the two mentioned above are applicable under a more relaxed unconfounded mechanism than the UWIC.

When the underlying mechanism is confounded (or NMAR), all approaches are vulnerable (not robust) to the misspecification of the response mechanism and the variance estimators are not valid. Moreover, the point estimators are biased and correction of the bias is a more pressing issue than estimation of the variance as studied in Rancourt, Lee, and Särndal (1994).

Imputation model All imputation methods assume a model, either explicitly or implicitly. The model-assisted approach uses the model assumption explicitly in the derivation of the variance formula and therefore, it is sensitive to the model assumption. On the other hand, if the model is correct, the approach provides more precise variance estimates. Note that Hidiroglou (1989) had also used a model to explore variance estimation under mean and ratio imputation. All other approaches are design-based or can be applied under the design-based framework and are robust to the misspecification of the imputation model. Rao (1992) proved that the jackknife is design and model unbiased (ξp-unbiased) under a linear imputation model and the UWIC (MAR) mechanism.

Public use of the imputed data set If a survey data file with imputed values is made available for public-use, some information required for variance estimation may not be available for confidentiality reasons. This makes estimation of the total variance difficult. Without consideration of the imputation variance, Yung (1997) used the bootstrap technique to produce confidentiality-protected public-use micro data files with bootstrap weights. A similar approach could be considered for the creation of public-use imputed data. The multiple imputation (Rubin, 1987) has an edge in this regard because it was conceived for the creation of public-use micro data files, which does not require information on imputation for variance estimation.

Summary Table 1 below presents a summary of the discussion given above.

Table 1. Summary of the Characteristics of the Variance Estimation Approaches

Approach	Var. Comp.	Comp. Burden	Imputation Flag	Adapta-bility	Response Mech.	Model
Two-phase	Yes	Low	Yes	Medium	UWIC	No
Model-assisted	Yes	Low	Yes	Medium	Unconf.	Yes
Jackknife	No	Medium	Yes	Medium	Unconf.	No
Lin. Jackknife	Yes	Low	Yes	Low	Unconf.	No
Bootstrap	No	High	Yes	High	Unconf.	No
BRR	No	High	Yes	High	Unconf.	No
ACI	Yes	Medium	Yes	Medium	UWIC	No
POP	No	Low	Yes	Medium	Unconf.	No

3.2 Empirical comparisons

In this section, the variance estimators presented in Section 2 are empirically compared with each other through a simulation study. Twelve populations (of size $N = 100$) were artificially generated, as described in Lee, Rancourt and Särndal (1994) for the simulation. Four different super-population models were used for this purpose: ratio, simple regression with an intercept, and two nonlinear regression models with a mild second-degree term. For each model type, three error variance structures were considered. Then, a simple random sample without replacement of size $n = 30$ was drawn and nonresponse was simulated using an expected nonresponse rate of 30% and five different response mechanisms (uniform, unconfounded with the response probability increasing or decreasing with z, and confounded with the response probability increasing or decreasing with y). The prototype estimator is the Horvitz-Thompson estimator $\hat{Y}_{\text{HT}} = (N/n)\sum_{k=1}^{n} y_k$ for the population total. Finally, two imputation methods were studied, namely, ratio and nearest neighbour imputation. The imputation model behind ratio and nearest neighbour imputation is the ratio model as given in (1.3.1) with a scalar imputation auxiliary variable z_k and $c_k = z_k$. The variance estimation formulae presented below are those that apply to the Horvitz-Thompson imputed estimator with ratio imputation.

1) Ordinary variance estimator (ORD): $N^2(1/n - 1/N)S_{y\bullet s}^2$ with $S_{y\bullet s}^2$ as given in (2.2.3).
2) Two-phase approach (2PH): $\hat{V}_{\text{SAM}} + \hat{V}_{\text{cIMP}}$ with components as given in (2.1.6) and (2.1.7).
3) Model assisted approach (MOD): $\hat{V}_{\text{SAM}} + \hat{V}_{\text{cIMP}}$ with components as given in (2.2.3) and (2.2.4)
4) Jackknife technique (JKNF): $\{(n-1)/n\}\sum_{j \in s}(\hat{Y}_I^{(aj)} - \hat{Y}_I^{(a)})^2 - NS_{yr}^2$, which is the fpc-corrected formula proposed by Lee, Rancourt, and Särndal (1995).
5) Linearized jackknife approach (LINJ): The formula given in (2.3.5).
6) Balanced repeated replication (BRR): To implement this approach under SRSWOR, two pseudo clusters were created by randomly dividing the sample into two equal-sized groups. Then the adjusted BRR variance estimator (assuming $H = 1$ and $n_h = 2$) with $L = 2$ balanced replicates was applied with repeating the procedure K (= 50) times to stabilize the variance of the variance estimator. An fpc-corrected estimator was then obtained as $(1/K)\sum_{k=1}^{K}\sum_{\ell=1}^{L}(\hat{Y}_I^{(kal)} - \hat{Y}_I^{(k)})^2 - NS_{yr}^2$.
7) Bootstrap approach (BOOT): $(1/L)\sum_{\ell=1}^{L}(\hat{Y}_I^{*(l)} - \overline{\hat{Y}}_I^{*})^2$.
8) All-case-imputation approach (ACI): Not implemented since the formula as given in (2.5.2) is not applicable for ratio imputation.
9) Approach based on the population level response mechanism (POP): The formula is given in (2.6.5).
10) Multiple imputation was carried out with $M = 2, 5 \text{ and } 50$ for ratio imputation and $M = 2$ for nearest neighbour imputation in the same way as given in Lee, Rancourt, and Särndal (1994).

Not all approaches provide a variance estimation formula for nearest neighbour imputation. One of those approaches is the model-assisted and the formula is $N^2(1/n - 1/N)S_{y\bullet s}^2 + \hat{V}_{\text{cIMP}}$ with $\hat{V}_{\text{cIMP}}$ given in (2.2.5). Rancourt (1999) proposed a method to use the JKNF approach for NN imputation, which was implemented in the simulation not only for the JKNF but also for the BRR and BOOT approaches. The ACI formula given in (2.5.2) is also appropriate for NN imputation. For other approaches, there is no specific formula available for NN imputation but we used the same formulae used for ratio imputation as given above.

The sampling experiment was repeated 50,000 times. The performance of the variance estimators was compared using three measures: relative bias (RB), root mean squared error (RMSE), and coverage rate (COVR) of the 95% confidence interval. They are defined as follows:

$$\text{RB}\,[\hat{V}(\hat{Y}_I)] = 100\big(E_M(\hat{V}) - V\big)/V \quad \text{and} \quad \text{RMSE}\,[\hat{V}(\hat{Y}_I)] = \sqrt{E_M(\hat{V} - V)^2} \qquad (3.2.1)$$

where E_M is the Monte Carlo average over all iterations and V is the Monte Carlo variance of the imputed estimator $\hat{Y}_I$. The COVR is defined as the ratio of the number of times that the 95% confidence interval $\hat{Y}_I \pm 1.96\sqrt{\hat{V}}$ contains the true population total Y to the total number of iterations of the simulation (50,000).

The purpose of the simulation was two-fold: (i) evaluate the approaches under the ideal situation where both assumed imputation model and assumed response mechanism are correct; and (ii) observe the sensitivity of the approaches to some violations of these assumptions.

Without affecting the conclusions, we present the results for only two populations (Ratio and Concave) to save space. The super-population model used to generate these populations is given by

$$y_k = a + bz_k + cz_k^2 + \varepsilon_k, \quad E(\varepsilon_k^2) = \sigma^2 z_k, \quad E(\varepsilon_k \varepsilon_l) = 0, \text{if } k \neq l. \qquad (3.2.2)$$

The ratio population was generated from this model with $a = 0$, $b = 1.5$, $c = 0$, and $\sigma = 13.78$; for the Concave population $a = 0$, $b = 3$, $c = -0.1$, and $\sigma = 5.6$ were used(see Lee, Rancourt, and Särndal (1994) for more detail). The value of σ was chosen to have a correlation coefficient of 0.75 between the z and y variables.

The three response mechanisms used in the simulation are: (1) uniform with 70 % response probability for any sample unit; (2) unconfounded mechanism with a decreasing response probability according to $\exp(-\kappa z_k)$ as z_k increases with a constant κ determined in such a way that the overall response rate is 70%; (3) confounded mechanism with a decreasing response probability according to $\exp(-\kappa y_k)$ as y_k increases with a constant κ determined in such a way that the overall response rate is 70%. Mechanisms (2) and (3) are non-uniform but (1) and (2) are unconfounded or MAR, while (3) is confounded or NMAR.

The Ratio population is ideal for the ratio imputation model, while the Concave represents the non-ideal populations. Table 2 presents the results and discussion of the results follows.

Under the Ratio population

Case RR1: Ratio imputation under uniform response mechanism.

All variance estimators designed for ratio imputation work reasonably well with the RB contained within 10%. As expected the ORD approach greatly underestimated the variance by about 30% and must be ruled out under any circumstances when the nonresponse rate is appreciable. For the JKNF and BRR approaches, the small positive RBs (4.6 and 5.2, respectively) could have been much larger if the fpc-correction had not been used. The BOOT approach underestimated the variance slightly (-6% RB). However, the negative bias is related to the fpc. The bootstrap procedure appropriate for simple random sampling without replacement mimics the finite population sampling procedure and therefore, the fpc is automatically incorporated in the bootstrap procedure. When the variance of imputed estimator is estimated by the BOOT approach of Shao and Sitter (1996), the imputation variance is also affected by the fpc incorporated in the bootstrap procedure, which results in a negative bias.

The RMSE's are fairly close for all approaches except the BRR which may be due to the way it was implemented. The BRR approach is meant for stratified sampling. Nonetheless, we applied the approach by creating pseudo clusters for a single stratum design as described earlier. All the variance estimators (except ORD) give fairly good coverage rates achieving close to the nominal 95%.

Case RR2: Ratio imputation under unconfounded response mechanism.

In terms of RB, all approaches have a limited bias within 10% range. All approaches also have very good coverage rates. The JNKF and BRR approaches have somewhat larger RMSE.

Case RR3: Ratio imputation under confounded response mechanism.

All variance estimators are severely biased with RB ranging between –34% and –22% because the point estimator is negatively biased (-7.7% of the population total). This is a confirmation of the well known fact that the nonresponse methodology is very sensitive to the important difference between the unconfounded and confounded response mechanisms.

Case RN1: Nearest neighbour (NN) imputation under uniform response mechanism.

The ORD approach is even more biased with –35% RB since a larger imputation variance than for ratio imputation is missed. The variance formulae (MOD, JKNF, BRR, and, BOOT) designed explicitly or implicitly for NN imputation worked well with a negligible negative bias except for the BOOT approach. The bias of the BOOT is somewhat large with RB of -14% for the reason explained before but in the case of NN imputation, the negative bias is larger because the variance is larger than for ratio imputation. The 2PH and LINJ approaches have a moderate negative bias mainly because NN imputation was treated like ratio imputation. The POP approach is also negatively biased for the same reason but much less so than the 2PH and LINJ approaches.

The replication approaches tend to have larger RMSE. The COVR's for all approaches including those with a moderate bias are reasonable, even though they are a little bit on the low side.

Case RN2: NN imputation under unconfounded response mechanism.

The RB's of three approaches (MOD, JKNF, and BRR) are small, maintaining good performance under Case RN1. The magnitude of the RB's for other approaches except the LINJ are substantially increased. The ACI approach breaks down since it is designed for the uniform response mechanism. The approaches with a small bias have a slightly too low coverage rate, but it is still acceptable. The BRR has the largest RMSE despite its near unbiasedness.

Case RN3: NN imputation under confounded response mechanism.

Similarly as for Case RR3, all approaches are negatively and heavily biased. The overestimation tendency of the MOD approach helps in reducing the bias although it is still severe at –21% RB.

Under the Concave population

Note that this population represents a mild violation of the assumed imputation model.

Case CR1: Ratio imputation under uniform response mechanism.

The approaches (2PH, JNKF, LINJ, BOOT, BRR and POP) that are less dependent on the imputation model performed well except BOOT. BOOT did not fare well because of the problem noted above. The MOD approach has a moderate positive bias but excellent COVR. All other approaches also have an acceptable COVR with over 92%. Even the ORD has a COVR over 90%. The RMSE of BRR is again strikingly large.

Case CR2: Ratio imputation under unconfounded response mechanism.

All approaches become more biased in this case than in Case CR2 except MOD, JNKF, and BRR. The biases for LINJ, 2PH and POP are still moderate. The COVR's are pretty good for almost all approaches.

Case CR3: Ratio imputation under confounded response mechanism.

It is very interesting to observe that all approaches worked quite well under the unconfounded response mechanism, which is a striking contrast with Case RR3. In this case, two wrong assumptions (wrong response mechanism and wrong imputation model) combined together create an artificially favorable situation. The COVR for the MOD is quite close to the nominal value and for other approaches it is little bit short but still over 90%.

Case CN1: NN imputation under uniform response mechanism.

All approaches have a more visible bias, but it is still fairly moderate with the absolute RB less than 15% (of course excluding ORD). Only the MOD approach has a positive bias. All COVR's are over 90% but fall somewhat short of the nominal value of 95%, except for the MOD which has a COVR of 95%.

Case CN2: NN imputation under unconfounded response mechanism.

The bias deteriorates further in the same direction as under the uniform RM. Nonetheless, MOD still has an excellent COVR and for others it is around or slightly lower than 90%.

Case CN3: NN imputation under confounded response mechanism.

All the approaches are unacceptably biased in the negative direction except MOD, which performed surprisingly well with only 4% RB and 92% COVR.

Multiple Imputation (MI)

Given that the imputation is model-based, the simulation results for the multiple imputation variance estimator are somewhat similar to those of MOD, in terms of RB. As the multiple imputation theory suggests, COVR is good under uniform RM and unconfounded RM. Its RMSE is substantially higher than those of the single imputation estimators when $M = 5$ but becomes more or less the same when $M = 50$. Note that $M = 2$ is not sufficient as some authors suggested (Rubin and Schenker, 1986).

Table 2. Results of the Simulation Study for Two Populations (Ratio and Concave) and Two Imputation Methods (Ratio and NN)

Variance estimation approach	Ratio population									Concave population								
	Uniform			Unconfounded			Confounded			Uniform			Unconfounded			Confounded		
	RB	RMSE[1]	COVR	RB	RMSE[1]	COVR	RB	RMSE[1]	COVR	RB	RMSE[1]	COVR	RB	RMSE[1]	COVR	RB	RMSE[1]	COVR
Ratio Imputation																		
ORD	-30.9	25.2	88.6	-39.8	35.4	86.7	-56.7	54.7	76.4	-22.0	32.2	90.7	-33.7	54.6	89.5	-30.0	41.5	87.7
2PH	2.0	22.5	94.2	-8.6	25.7	93.4	-33.7	38.1	84.5	-1.4	30.3	93.8	-15.2	43.2	93.5	-9.9	34.5	91.2
MOD	7.0	23.5	94.9	4.9	29.4	95.0	-24.3	33.8	86.6	12.8	37.7	95.2	1.9	47.3	95.8	8.9	42.1	93.6
JKNF	4.6	25.9	94.4	8.8	35.5	95.1	-22.7	36.0	86.6	1.0	36.9	93.7	-2.0	53.6	95.1	-0.1	44.4	92.0
LINJ	2.4	23.2	94.2	-0.6	28.9	94.3	-28.3	36.2	85.7	-1.2	31.2	93.8	-11.6	44.8	94.0	-7.1	36.4	91.5
BRR	5.2	31.7	93.5	9.4	41.7	94.3	-22.2	39.9	86.0	1.2	46.6	92.8	-1.3	63.8	94.2	0.4	53.2	91.4
BOOT	-6.2	25.3	92.5	-8.1	31.5	92.9	-33.9	41.1	83.8	-8.7	36.8	92.2	-17.4	52.2	92.4	-13.7	42.2	90.0
POP	1.3	22.7	94.1	-1.8	28.3	94.2	-29.2	36.5	85.5	-1.7	30.8	93.7	-12.2	44.5	94.0	-7.7	36.0	91.5
MI (2)	10.4	67.1	93.2	11.7	98.7	92.9	-12.7	81.9	85.4	10.7	88.7	94.1	-2.9	116.8	93.7	11.0	104.9	92.6
MI (5)	12.1	37.6	94.8	12.4	53.2	94.8	-13.0	47.1	87.7	11.3	52.0	95.1	-2.8	70.4	95.2	12.4	62.3	93.9
MI (50)	12.3	26.9	95.4	13.4	36.2	95.9	-13.9	32.7	88.5	11.7	39.5	95.6	-3.0	50.8	95.8	12.9	45.7	94.4
NN imputation																		
ORD	-35.0	34.3	87.3	-44.8	48.8	84.3	-59.5	68.5	73.8	-31.9	44.0	87.7	-40.9	60.7	85.1	-47.6	69.8	81.1
2PH	-13.0	24.8	92.3	-21.4	32.0	91.2	-43.4	52.4	81.8	-8.4	31.8	92.7	-8.8	39.2	92.5	-23.7	45.9	88.5
MOD	6.6	33.4	94.7	12.3	53.6	95.2	-20.6	47.0	87.3	13.15	47.5	95.1	28.8	95.0	95.7	4.3	63.9	92.2
JKNF	-2.5	33.4	93.4	-3.7	44.8	93.3	-30.1	49.4	84.7	-12.9	39.5	91.4	-14.3	50.4	90.9	-26.5	55.2	86.7
LINJ	-12.7	25.4	92.3	-14.6	32.0	92.3	-38.9	49.4	83.3	-8.2	32.6	92.7	-5.0	42.2	93.0	-21.4	46.0	88.9
BRR	-2.4	38.9	92.5	-3.3	49.8	92.7	-29.7	52.7	84.0	-12.8	47.8	90.5	-14.1	57.5	90.1	-26.4	60.7	85.8
BOOT	-14.3	32.4	91.3	-19.2	41.6	90.6	-40.8	54.8	81.5	-19.9	42.2	89.9	-24.9	53.1	88.5	-34.6	60.4	84.6
ACI	-4.4	29.5	93.2	-21.4	37.6	90.6	-42.4	53.7	81.3	-8.6	33.8	92.4	-22.8	45.9	89.6	-31.3	53.4	86.0
POP	-7.6	28.9	92.9	-12.2	38.0	92.4	-36.1	49.8	83.5	-13.0	35.2	91.8	-17.8	45.5	90.7	-28.7	52.6	86.7
MI (2)	-22.1	31.7	89.9	-30.1	43.0	87.4	-52.4	63.9	76.6	-23.2	40.3	89.4	-31.2	53.7	86.8	-42.2	67.3	82.3

[1] The RMSE's are in hundreds.

4. DISCUSSION AND RECOMMENDATIONS

In this section we discuss the different approaches and make some recommendations. These recommendations are based on theoretical properties as well as limited but illuminating empirical results. Because of its large bias, the ordinary variance estimator (ORD) must be ruled out when the survey data set contains imputed values, unless the nonresponse rate is very small, say, less than 10% (the nonresponse rate is not the only important factor in this consideration, though). Therefore, in the following discussion we refer only to the other variance estimators discussed in the preceding sections.

Bias of the variance estimators. The variance estimators that we have studied have been shown to perform well. As theory would lead us to expect, they only have limited bias, under the conditions for which they are designed. However, the approaches react differently when (some of) these conditions are violated.

All approaches seem to be robust to a mild violation of the model assumptions. The MOD approach shows some sensitivity to the imputation model assumption. Nonetheless, the results for MOD under the Concave population are acceptable. However, when the assumed response mechanism is violated, all the approaches can totally fail, especially under the confounded mechanism. The variance formulae, particularly the 2PH and ACI approaches, developed assuming the uniform response mechanism show some sensitivity to the assumption. Therefore, when these approaches are used, the user should be vigilant in choosing the response mechanism. The use of (2.1.5) would be helpful for estimation of the imputation variance component by the 2PH approach.

If the imputation method is deterministic, the ACI approach may need some modification in order to estimate the sampling variance correctly.

If computational burden is of concern, then the JKNF, BRR, and BOOT approaches are not the most appropriate. When the sampling fraction is appreciable, the JKNF and BRR overestimate the variance, and an fpc-correction should be applied if possible or some other approach (2PH, MOD, LINJ, ACI and POP) should be used. If the prototype estimator is a non-smooth function of estimated totals, then the BRR or BOOT approaches are prime candidates.

When the imputation method is composite in the sense that imputed values are used for imputation of other variables, the POP approach may be suitable, since it was developed with this scenario in mind. However, the BOOT and BRR approaches should also be able to handle this case (at least in principle), unless the sampling fraction is appreciable, in which case an fpc-correction must be worked out for the BRR.

For NN imputation, all approaches seem to be working well except the 2PH, LINJ, and BOOT approaches, for which the formula for ratio imputation (2PH and LINJ) or an adjustment for ratio imputation (BOOT) was used. Since it is donor imputation, the imputation error is larger than ratio imputation and thus this increased variability has to be captured. The adjustment used for JKNF and BRR worked very well.

The form of the BOOT approach that we have examined underestimates the variance because the fpc is overcorrected.

Coverage rate. All approaches yielded a fairly good coverage rate, as long as the bias is not severe (say, less than 20% absolute RB). However, the coverage rate is on the low side in most cases. Therefore, improvement might be realized by the use of Student *t*-values instead of standard normal values in the construction of the confidence interval.

Required information. All approaches require that imputed units be flagged in the data file and need information on the imputation method, imputation auxiliary variables, and imputation classes. The donor imputation methods sometimes require identification of the donor. The JKNF, BRR, BOOT, and POP approaches do not need this information in the case of hot-deck imputation, but they (probably excluding POP) do need it in the case of NN imputation, so that the adjustment in Subsection 2.3 can be used. It may sound bizarre at first to say that the ACI approach, which is designed for donor imputation, does not require this information; this is so because the approach computes the imputation variance using imputed values for respondents. In any case, it would be a good practice to include, as far as possible, the donor information along with other necessary information. When this information is required but not available, it outs a limitation on the choice of approach.

Software availability. Another important consideration is the kind of software currently available. If the available software uses the jackknife technique, then it would be natural to use JKNF with an appropriate fpc adjustment. If, however, the available software is based on the Taylor method, then the replication or resampling approaches (JKNF, BRR, and BOOT) are automatically ruled out. The calculation of the sampling variance component can be done using the already available standard package if the imputation method is stochastic or NN, provided that the response rate is not extremely low. For deterministic and non-donor imputation, such as ratio or regression imputation, the choice would be between (i) adding random residuals to the imputed values for variance estimation, or (ii) developing a new software component for estimation of the sampling variance. To estimate the imputation variance component, a new software module is needed. A point to consider in this case is the type of survey for which the package will be used. If the package is intended mainly for business surveys, it is reasonable in many situations to assume an unconfounded (non-UWIC) response mechanism with one or more continuous imputation auxiliary variables, and thus, a approach (e.g., MOD and JKNF) designed for this mechanism would be more appropriate. The ACI and 2PH approaches are particularly sensitive to a violation of the assumption of a uniform response mechanism within imputation classes. If the package is intended mainly for social surveys, it is often possible to create imputation classes that come reasonably close to satisfying the UWIC condition, and then all the approaches are applicable. But since sample designs of social surveys tend to be complex, replication approaches may be more appropriate.

Separate variance components. It is of a great interest to survey managers, for periodic surveys in particular, to obtain separate variance components, one for sampling variance and one for imputation variance. This will facilitate an efficient allocation of resources among sampling and data processing activities. If this is the case, the approaches (2PH, MOD, and ACI) are recommended. When there are strong reasons to use another approach, a breakdown into variance components may be obtainable in some cases by estimating the sampling variance component separately (e.g., using the standard estimator for stochastic imputation) and then subtracting it from the total variance estimated by the chosen approach. If this is done, an fpc-correction can also be incorporated, as in Lee, Rancourt, and Särndal (1995).

5. CONCLUDING REMARKS

In this paper, we have reviewed and compared, theoretically and empirically, various approaches and techniques for estimating the variance for survey data with single imputation. Based on this review, we offered some recommendations.

The problem dealt with in this paper is an old problem, going back to the early days of survey taking. But concrete solutions to the problem have only started to appear since 1990. It is thus a relatively new field of research and development. Further significant developments have been presented and published since we started writing this paper. Unfortunately, those very recent ones were not covered here and more are expected in the near future.

To facilitate correct variance estimation and making correct inferences for data containing imputations, not only theoretical solutions to the problem but also development of suitable software is important. From the survey management point of view, it is of considerable interest to have the two separate variance components (sampling and imputation), and thus, this aspect should be taken into consideration in future software development.

ACKNOWLEDGMENTS

We are very much indebted to Jean-François Beaumont for his active participation in producing all simulation results and for his comments. We also thank Thomas Belin, John Eltinge, Mike Hidiroglou and Rod Little for their useful comments.

REFERENCES

Anderson, C. and Nordberg, L. (1994). A method for variance estimation of non-linear function of totals in surveys – Theory and a software implementation. *Journal of Official Statistics*, 10, 395-405.

Bickel, P.J. and Freedman, D.A. (1984). Asymptotic normality and the bootstrap in stratified sampling. *The Annals of Statistics*, 12, 470-482.

Burns, E.M. (1990). Multiple and replicate item imputation in a complex sample survey. *Proceedings of the Sixth Annual Research Conference* , U.S. Bureau of the Census, 655-665.

Chen, Y. (1993). Balanced repeated replication variance estimators for survey data under imputation, unpublished Ph.D. thesis, University of Ottawa, Dept. of Mathematics and Statistics.

Dalenius, T. (1983). Some reflections on the problem of missing data. In: W.G. Madow and I. Olkin (eds.), *Incomplete Data in Sample Surveys*, Vol. 3. New York: Academic Press, 411-413.

Deville, J.-C. and Särndal, C.-E. (1991). Estimation de la variance en présence de données imputées. *Proceedings of Invited Papers for the 48th Session of the International Statistical Institute*, Book 2, Subject 17, 3e17.

Deville, J.-C. and Särndal, C.-E. (1994). Variance estimation for the regression imputed Horvitz-Thompson estimator. *Journal of Official Statistics*, 10, 381-394.

Efron, B. (1979). Bootstrap Methods: "Another Look at the Jackknife." *The Annals of Statistics*, 7, 1-26.

Estevao, V., Hidiroglou, M.A. and Särndal, C.-E. (1995). Methodological principles for a generalized estimation system at Statistics Canada. *Journal of Official Statistics*, 11, 181-204.

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

Fay, R.E. (1996). Alternative paradigms for the analysis of imputed survey data. *Journal of the American Statistical Association*, 91, 490-498.

Ford, B.M. (1983). An overview of hot-deck procedures. In: W.G. Madow, I. Olkin, and D.B. Rubin (eds.), *Incomplete Data in Sample Surveys*, Vol. 2. New York: Academic Press, 185-207.

Gagnon, F., Lee, H., Provost, M., Rancourt, E. and Särndal, C.-E. (1997). Estimation of variance in presence of imputation. *Proceedings of Statistics Canada Symposium 97: New Directions in Surveys and Census*, Statistics Canada, 273-277.

Gagnon, F., Lee, H., Rancourt, E. and Särndal, C.-E. (1996). Estimating the variance of the Generalized regression estimator in the presence of imputation for the Generalized Estimation System. *Proceedings of the Survey Methods Section*, Statistical Society of Canada, 151-156.

Gross, S. (1980). Median estimation in sample survey. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 181-184.

Gurney, M. and Jewett, R.S. (1975). Constructing orthogonal replications for standard errors. *Journal of the American Statistical Association*, 70, 819-821.

Gupta, V.K. and Nigam, A.K. (1987). Mixed orthogonal arrays for variance estimation with unequal numbers of primary selections per stratum. *Biometrika*, 74, 735-742.

Hidiroglou, M.A. (1989). Notes on variance computations for data sets imputed by mean or ratio imputations. Unpublished manuscript, Statistics Canada.

Kalton, G. and Kasprzyk, D. (1986). The treatment of missing survey data. *Survey Methodology*, 12, 1-16.

Kott, P.S. (1995). A paradox of multiple imputation. *Proceedings of the Section on Survey Research Methods,* American Statistical Association, 380-383.

Kovar, J.G. and Chen, E. (1994). Jackknife variance estimation of imputed survey data. *Survey Methodology*, 20, 45-52.

Kovar, J.G. and Whitridge, P.J. (1995). Imputation of business survey data. *Business Survey Methods*, Cox, B.G., Binder, D.A., Chinnappa, B.N., Christianson, A., Colledge, M.J. and Kott, P.S. (editors), 403-423, New York: John Wiley and Sons.

Krenzke, T., Mohadjer, L. and Montaquila, J. (1998). Generalizing the imputation error variance in the Alcohol and Drug Services Study, *Proceedings of the Biometrics Section*, American Statistical Association, 118-123.

Lee, H., Rancourt, E. and Särndal, C.-E. (1994). Experiment with variance estimation from survey data with imputed values. *Journal of Official Statistics*, 10, 231-243.

Lee, H., Rancourt, E. and Särndal, C.-E. (1995). Jackknife variance estimation for data with imputed values. *Proceedings of the Survey Methods Section*, Statistical Society of Canada, 111-115.

Little, R.J.A. and Rubin, D.B. (1987). *Statistical Analysis with Missing Data*. New York: Wiley.

McCarthy, P.J. (1969). Pseudo-replication: half-samples. *Review of the International Statistical Institute*, 37, 239-264.

McCarthy, P.J. and Snowden, C.B. (1985). The bootstrap and finite population sampling. *Vital and Health Statistics*, Series 2, No. 95. DHHS Pub. No. (PHS) 85-1369.

Montaquila, J.M. and Jernigan, R. W. (1997). Variance estimation in the presence of imputed data. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 273-278.

Rancourt, E. (1996). Issues in the combined use of Statistics Canada's Generalized Edit and Imputation System and Generalized Estimation System. *Survey and Statistical Computing : Proceedings of The Second ASC International Conference*, Association for Survey Computing, 185-194.

Rancourt, E. (1999). Estimation with nearest neighbour imputation at Statistics Canada. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 131-138.

Rancourt, E., Lee, H. and Särndal, C.-E. (1993). Variance estimation under more than one imputation method. *Proceedings of the International Conference on Establishment Surveys*, American Statistical Association, 374-379.

Rancourt, E., Lee, H. and Särndal, C.-E. (1994). Bias corrections for survey estimates from data with ratio imputed values for confounded nonresponse. *Survey Methodology*, 20, 137-147.

Rancourt, E., Särndal, C.-E. and Lee, H. (1994). Estimation of the variance in the presence of nearest neighbour imputation. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 888-893.

Rao, J.N.K. (1990). Variance estimation under imputation for missing data. Technical report, Statistics Canada, Ottawa.

Rao, J.N.K. (1992). Jackknife variance estimation under imputation for missing data. Technical Report, Statistics Canada, Ottawa.

Rao, J.N.K. (1993). Linearization variance estimators under imputation for missing data. Technical report, Statistics Canada, Ottawa.

Rao, J.N.K. (1996). On variance estimation with imputed survey data. *Journal of the American Statistical Association*, 91, 499-506.

Rao, J.N.K. and Shao, J. (1992). Jackknife variance estimation with survey data under hot-deck imputation. *Biometrika*, 79, 811-822.

Rao, J.N.K. and Shao, J. (1996). On balanced half-sample variance estimation in stratified sampling. *Journal of the American Statistical Association*, 91, 343-348.

Rao, J.N.K. and Sitter, R.R. (1992). Jackknife variance estimation under imputation for missing survey data. Technical report No. 214, Carleton University, Laboratory for Research in Statistics and Probability.

Rao, J.N.K. and Sitter, R.R. (1995). Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika*, 82, 453-460.

Rao, J.N.K. and Wu, C.F.J. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association*, 83, 231-241.

Rubin, D.B. (1978). Multiple imputation in sample surveys - a phenomenological Bayesian approach to nonresponse. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 20-34.

Rubin, D.B. (1987). *Multiple imputation for nonresponse in surveys*, New York: John Wiley and Sons.

Rubin, D.B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91, 473-489.

Rubin, D.B. and Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponses. *Journal of the American Statistical Association*, 81, 366-374.

Särndal, C.-E. (1990). Methods for estimating the precision of survey estimates when imputation has been used. *Proceedings of Statistics Canada Symposium 90: Measurement and Improvement of Data Quality*, Statistics Canada, 337-347.

Särndal, C.-E. (1992). Methods for estimating the precision of survey estimates when imputation has been used. *Survey Methodology*, 18, 241-252.

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

Shao, J., Chen, Y. and Chen, Y. (1998). Balanced repeated replication for stratified multistage survey data under imputation. *Journal of the American Statistical Association*, 93, 819-831.

Shao, J. and Sitter, R.R. (1996). Bootstrap for imputed survey data. *Journal of the American Statistical Association*, 91, 1278-1288.

Shao, J. and Steel, P. (1999). Variance Estimation for Survey Data With Composite Imputation and Nonnegligible Sampling Fractions. *Journal of the American Statistical Association*, 94, 254-265.

Sitter, R.R. (1992a). A resampling procedure for complex survey data. *Journal of the American Statistical Association*, 87, 755-765.

Sitter, R.R. (1992b). Comparing three bootstrap methods for survey data. *Canadian Journal of Statistics*, 20, 135-154.

Sitter, R.R. (1993). Balanced repeated replications based on orthogonal multi-arrays. *Biometrika*, 80, 211-221.

Sitter, R.R., and Rao, J.N.K. (1997). Imputation for missing values and corresponding variance estimation. *Canadian Journal of Statistics*, 25, 61-73.

Skinner, C.J. and Rao, J.N.K. (1993). Jackknife variance estimation for multivariate statistics under hot-deck imputation. *Proceedings of the International Statistical Institute*, 421-422.

Steel, P. and Fay, R.E. (1995). Variance estimation for finite populations with imputed data. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 374-379.

Tollefson, M. and Fuller, W.A. (1992). Variance estimation for samples with random imputation. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 758-763.

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

Wu, C.F.J. (1991). Balanced repeated replications based on mixed orthogonal arrays. *Biometrika*, 78, 181-188.

Yung, W. (1997). Variance estimation for public use microdata files. *Proceedings of Statistics Canada Symposium 97: New directions in Surveys and Censuses.* Statistics Canada, Ottawa, 91-95.

Zanutto, E. (1993). Jackknife variance estimation under imputation for missing data in survey samples. Master of Science Thesis, Carleton University, Ottawa.