WP. 28
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Neuchâtel, Switzerland, 5-7 October 2009)

Topic (vi): New and emerging methods

A TRIPLE-GOAL IMPUTATION METHOD FOR STATISTICAL REGISTERS

Invited Paper

Prepared by Li-Chun Zhang, Statistics Norway

Work Session on Statistical Data Editing
(Neuchatel, Switzerland, 5-7 October 2009)

A triple-goal imputation method for statistical registers

Li-Chun Zhang[1]

Abstract
Restricted neighbor imputation is a triple-goal imputation method for statistical registers. We outline the method, discuss its motivations, describe the computation algorithm, and consider its properties regarding estimation consistency and variance evaluation.

I. Introduction

Construction of a *statistical* register is one mode of statistical production that is able to combine data from different sources including sample survey, administrative registers and census. Zhang and Nordbotten (2008) argued that the following three goals are desirable of any method:

(A) It should yield efficient estimates of the totals of interest.

(B) It should yield correct co-variances among *all* the variables.

(C) It should be non-stochastic.

The nearest neighbor imputation (NNI) yields consistent estimates of totals as well as distributions (Chen and Shao, 2000). Because the imputed values are realistic ones, the NNI is able to preserve the *natural* variation in data. Moreover, it is in principle a non-stochastic method, and can be made so in practice. The underlying assumptions are nonparametric and mild. While this can be advantageous because one avoids the adoption of explicit parametric models required for regression-based frequentistic or Bayesian imputation methods, a potential drawback is the lack of efficiency compared to the model-based methods.

The method of *restricted neighbor imputation (ReNI)* is a modification of the NNI method on two accounts. The aim is to improve on its efficiency w.r.t. (A) while preserving its desirable features w.r.t. (B) and (C). First, a set of population totals, possibly on different aggregation levels, are obtained and fixed as the *imputation constraints*. Next, neighbor imputation is carried out iteratively so that the imputed population data file (i.e. statistical register) sums up to the imputation constraints as close as possible. Here, it is important to notice that the imputation constraints can be obtained separately from the subsequent imputation, and by any estimation method (or methods) that is deemed to be efficient and suitable. Below we shall outline the ReNI method, describe the computation algorithm, and consider its properties regarding estimation consistency and variance evaluation.

[1]Statistics Norway, Kongensgt. 6, PB 8131 Dep, N-0033 Oslo, Norway. E-mail: lcz@ssb.no

II. Restricted neighbor imputation

Denote by $U = \{1, ..., N\}$ the population of interest. For each unit $i \in U$, Let y_i be a variable of interest, and let x_i be a scalar variable which may depend on one or several auxiliary variables known throughout the population. For any $i \neq j \in U$, the distance between then is given by $|x_i - x_j|$. Let $a_i = 1$ if y_i is observed and $a_i = 0$ otherwise. Notice that $a_i = 0$ can arise from a number of situations. Take e.g. sampling survey. We then typically have a sample inclusion indicator I_i and a response indicator r_i, where $I_i = 1$ is i is in the sample (denoted by s) and $I_i = 0$ otherwise, and $r_i = 1$ for $i \in s$ in case of response and $r_i = 0$ otherwise. Here, we have $a_i = 0$ if $I_i = 0$, or if $I_i = 1$ but $r_i = 0$. Take another situation where two registers are linked to each other at the unit level. We may then have $a_i = 0$ if the unit i belongs to the one register that defines the population, yet it fails to match to the other register containing the y-variable. In any case, we assume the following about the mechanism by which y_i remains unobserved:

Assumption A. The population U is divided into K imputation classes such that within each class, (x_i, y_i, a_i)'s are independent and identically distributed and $P(a_i = 1|x_i, y_i) = P(a_i|x_i)$.

In survey sampling, assumption A requires what is otherwise known as non-informative sampling of the gross sample as well as missing at random (MAR) of the nonrespondents. Under assumption A, the nonrespondents and the out-of-sample units can be imputed in a single step.

Consider $i \in U_k$, where U_k denotes the kth imputation class. Let $y_i^* = y_i$ be observed if $a_i = 1$, and let $y_i^* = y_j$ be imputed if $a_i = 0$ where $j \in U_k$ and $a_j = 1$. In the case of imputation, the unit j is called the *donor*, and i the *receiver*. Under the NNI, j is chosen to minimize $|x_i - x_j|$. Multiple nearest neighbors arise if at least two donors have the minimum 'distance' to the receiver. To obtain a non-stochastic procedure, one may use a suitable additional variable z_i, and select the nearest neighbor based on $|z_i - z_j|$ in addition to $|x_i - x_j|$. For instance, z_i can be the post zip code, the identification number, *etc.*.

The nearest neighbor criterion can be extended to an m-neighbor criterion, for $m \geq 2$. Let $D_{i;m}$ contain the m nearest neighbors of unit i measured w.r.t. x (and possibly z in addition), for $i \in U_k$ and $D_{i;m} \subset U_k$. An m-neighbor imputation method selects the donor from $D_{i;m}$, i.e. $y_i^* = y_j$ if $a_i = 0$ where $j \in D_{i;m}$ and $a_j = 1$. The resulting imputed population vector, denoted by $\mathbf{y}^* = (y_1^*, ..., y_N^*)$, is referred to as an *$m$-neighbor imputation*. Let $\hat{Y}$ be a chosen estimate of the population total $Y = \sum_{i \in U} y_i$. Let $Y^* = \hat{Y}$ be an imputation constraint. Let $Y_{(m)}^*$ be the imputed total corresponding to an m-neighbor imputation. The *ReNI method of the degree m* sets out to minimize $(\hat{Y}^* - \hat{Y}_{(m)}^*)^2$ among all possible m-neighbor imputations.

In practice the number of possible imputations is typically too large for a complete search. Instead, we start with an *initial* imputation, and iteratively re-impute as long as $(\hat{Y}^* - \hat{Y}_{(m)}^*)^2$ can be reduced by replacing one current donor with another, until a tolerance margin has been reached. The algorithm is referred to as *fine-tuning (FT)*, and is given as follows:

Fine-tuning (FT) *Denote by R and D, respectively, the set of receivers and donors. Set m.*

-Initialization: *For each $i \in R$, find an initial donor $j \in D$ and set $I_i^* = j$.*

-Iteration: *For each $i \in R$, find donor $j' \in D_{i;m}$ that minimizes $(\hat{Y}^* - \hat{Y}^*_{(m)})^2$ on substituting j' for the current donor identified by I^*_i, and set $I^*_i = j'$ afterwards.*

-Stopping rule: *Repeat the round of Iteration until no exchanges of donors can be made, or if the reduction in $(\hat{Y}^* - \hat{Y}^*_{(m)})^2$ is below a specified tolerance margin.*

As the default choice we use the nearest neighbor imputation as the initial imputation. To further speed up the ReNI, a *jump-start (JS)* step can be applied before fine-tuning. The aim is to assign as many as possible nearest-neighbor imputations in a sensible way, so that many fewer receivers are left to fine-tuning afterwards. The JS-algorithm is given as follows:

Jump-start (JS) *Denote by R and D, respectively, the set of receivers and donors.*

-Initialization: *For each $j \in D$, set $d_j = 0$. For each $i \in R$, find first its nearest neighbor, denoted by $j \in D$, then set $I_i = j$ and increase d_j by 1.*

-Calibration: *Let $\mathbf{Y}^*_R = (|R|,\ Y^* - Y_D)^T$ where $|R|$ is the size of R and $Y_D = \sum_{j\in D} y_j$. Define a column vector $\mathbf{y}_j = (1,\ y_j)^T$ for $j \in D$. Put*

$$d'_j = d_j g_j \qquad where \qquad g_j = 1 + (\mathbf{Y}^*_R - \tilde{\mathbf{Y}}_R)^T \tilde{\mathbf{A}}^{-1} \mathbf{y}_j \tag{1}$$

for $\tilde{\mathbf{Y}}_R = \sum_{j\in D} d_j \mathbf{y}_j$ and $\tilde{A} = \sum_{j\in D} d_j \mathbf{y}_j \mathbf{y}_j^T$. It is easily verified that $\sum_{j\in D} d'_j y_j = \mathbf{Y}^*_R$.

-Imputation: *For $i = 1, ..., |R|$, set $a^*_i = 0$ if $d'_{I_i} < 1$, otherwise set $a^*_i = 1$ and decrease d'_{I_i} by 1.*

After the JS-step, each $i \in R$ with $a^*_i = 1$ receives a NN-imputation, and only the ones with $a^*_i = 0$ are left over for the FT-step. The d'_j's are calibrated w.r.t. $\mathbf{Y}^*_R$. We would manage to impute for all the receivers by the JS-step if the d'_j's are all integers, in which case the NNI automatically satisfies the imputation constraint Y^*. Otherwise, as it is usually the case, the constraint on the imputed total for the remaining receivers is given by $Y^* - \sum_{j\in D} y_j - \sum_{i\in R} a^*_i y^*_i$.

Both the x- and y-variables are scalars in the above exposition. Some extensions are worth noticing. The imputation constraint Y^* can sometimes be specified at the imputation class level, denoted by Y^*_k for $k = 1, ..., K$, in which case the ReNI method will be separately carried out inside each imputation class. The imputation constraint can also be a vector of variables, denoted by $\mathbf{Y}^*$. The squared difference $(Y^* - Y^*_{(m)})^2$ will then need to be replaced by a suitable metric between two vectors, and the Calibration (1) in the JS-step modified accordingly. In particular, there is the possibility of combining imputation constraints on different aggregation levels. For instance, in addition to $\hat{Y}$ for the whole population, one may specify $\hat{Y}'_k$ as the target total in each imputation class U_k, and use $\mathbf{Y}^* = (\hat{Y}, \hat{Y}'_1, ..., \hat{Y}'_K)^T$ as the imputation constraints. Notice that the imputation will no longer be independent between the classes due to the overall constraint $\hat{Y}$, although the donor may still be select within each class. Finally, in order to identify the neighbors, multiple characteristics can be used, denoted by $\mathbf{x}$. Sometimes these can be combined to obtain a scalar $x' = g(\mathbf{x})$, such that we are formally back in the situation above. Or a suitable metric may be available to measure the distance between two x-vectors. Or one may choose to match the x-variables one by one, always selecting the closest match on the next variable given the variables that have already been looked at. This can be a good alternative when there are both categorical

and continuous auxiliary variables, in which case the categorical ones are typically matched first. Or it may be the case that the different characteristics assume different priorities.

III. On the properties of ReNI method

a. Consistency

Any statistical imputation method assumes in some form exchangeability between a missing value and an observed one, after certain aspects of the units have been controlled for. The NNI method is thus intuitively consistent provided this is the case given equal x-values. Given simple random sampling and a single imputation class ($K = 1$), Chen and Shan (2000, Theorem 1) proved that the NNI-based sample mean, denoted by $\bar{y}^*_{NNI}$, is an asymptotically unbiased estimator of the unconditional model expectation of y, denoted by $E(y)$. Below we first restate their result and then discuss how it can be extended to the situations that we are concerned of.

Theorem 1 (Chen and Shao, 2000). Suppose that (i) assumption A holds; (ii) there exist constants $M_1 < M_2$ and C, where (M_1, M_2) may be $(-\infty, \infty)$, such that the function $\psi(x) = E(y|x)$ is a monotone function when $x < M_1$ or $x > M_2$, and $|\psi(x) - \psi(x')| \leq C|x - x'|$ when $x, x' \in [M_1, M_2]$; (iii) the marginal distribution of x has a density, $E|x|^3 < \infty$ and $E|\psi(x)|^3 < \infty$, and (iv) $\inf_{x\in\Omega} P(a = 1|x) > 0$ where Ω is the support of the marginal distribution of x. Then $E(\bar{y}^*_{NNI}|r) - E(y) = o_p(n^{1/2})$, where r is the number of observed units.

When the NN-imputation is carried out for the units outside of the sample as well as the non-respondents in the sample, we are concerned about the imputed population mean, denoted by $\bar{Y}^*_{NNI}$. The conditions (ii) and (iii) above are the same model assumptions about x and y. When the missing data arise from linking different registers, the plausibility of conditions (i) and (iv) need to be checked for the actual data sources. Whereas in the case of simple random sampling, one may regard the sample selection as an extra phase of response such that, for each $i \in U$, we have $P(a_i = 1|x_i, y_i) = \pi P(r_i = 1|x_i)$ where π is the constant inclusion probability, and both the conditions (i) and (iv) remain valid. It follows that $E(\bar{Y}^*_{NNI}|r) - E(y) = o_p(n^{1/2})$, just like $\bar{y}^*_{NNI}$ based on imputation for sample nonrespondents only.

As pointed out by Chen and Shao (2000), $\bar{y}^*_{NNI}$ remains asymptotically unbiased for $E(y)$ under stratified simple random sampling. For $\bar{Y}^*_{NNI}$ we notice that $P(a_i = 1|x_i, y_i) = P(I_i = 1)P(r_i = 1|x_i)$, where $P(I_i = 1)$ is the unconditional inclusion probability of a randomly chosen unit i from the population U. Under stratified simple random sampling, we have $P(I_i = 1) = \sum_h P(i \in U_h)P(I_i = 1|i \in U_h) = \sum_h (N_h/N)(n_h/N_h) = n/N$, where U_h denotes the h-th stratum, and (N_h, n_h) the corresponding stratum population and sample sizes, and $n = \sum_h n_h$ and $N = \sum_h N_h$. Thus, the assumption A remains valid, and so is the asymptotic unbiasedness of $\bar{Y}^*_{NNI}$.

Let $\mathbf{Y}^*_{NNI}$ be a vector of NNI-based population means, and let g be a differentiable function, then $g(\mathbf{Y}^*_{NNI})$ is asymptotically unbiased for $g(\mathbf{Y})$. This follows from the result for arbitrary scalar $\bar{Y}^*_{NNI}$ on application of Taylor's expansion. In particular, it follows that the imputed population covariance, denoted by $\sigma^*_{12} = \bar{Y}^*_{3,NNI} - \bar{Y}^*_{2,NNI}\bar{Y}^*_{1,NNI}$ where $y_{3i} = y_{1i}y_{2i}$ for $i \in U$, is asymptotically unbiased for the covariance $\sigma_{12} = E(y_{1i}y_{2i}) - E(y_{1i})E(y_{2i})$.

The asymptotic unbiasedness of the estimators remains under the same conditions (i) - (iv) when NN-imputation is replaced by m-neighbor imputation for fixed m. The key is to realize that, asymptotically as $n \to \infty$, condition (iv) implies there will be m-th neighbor that is arbitrary close to any given receiver. Condition (ii) is thus equally effective in the case of m-neighbor imputation. Provided the variance of an imputed estimator is of the order $O(n^{-1})$, neighbor imputation yields consistent estimation of both the population totals and the population covariances. Additional imputation constraint, hence the ReNI method, does not distort the consistency property provided the constraints are asymptotically unbiased themselves.

b. Variance

Assume that each unit $i \in U$ is associated with a vector of interest variables $\mathbf{y}_i$. The population means of interest based on donor imputation are then given by

$$\bar{\mathbf{Y}}^* = \sum_{j \in D} \mathbf{y}_i / N + \sum_{i \in R} \mathbf{y}_i^* / N = \sum_{j \in D} (1 + d_j) \mathbf{y}_j / N$$

where d_j is the number of times $i \in D$ is used as a donor for the units in R. Assume bi-partition of the vector of interest $\mathbf{y}_i^T = (\mathbf{y}_{1i}^T, \mathbf{y}_{2i}^T)$, and the corresponding population means $(\bar{\mathbf{Y}}_1, \bar{\mathbf{Y}}_2)$. Suppose that the imputation is subjected to the constraint $\bar{\mathbf{Y}}_1^* = \hat{\bar{\mathbf{Y}}}_{\mathbf{1}}$, for some chosen estimator of $\bar{Y}_1$. The imputed population mean $\bar{\mathbf{Y}}_2^*$ is not directly subjected to constraint, but it depends on $\bar{\mathbf{Y}}_1^*$ through the donor imputation. This is the setting for the ReNI method.

The variance is evaluated both w.r.t. a model of $\mathbf{y}_i$ given x_i, and the distribution from which the donor set D arises. Denote by E_p and V_p the expectation and variance w.r.t. the distribution of D. For instance, in the case of survey sampling, E_p and V_p depend on the sampling design as well as the response mechanism subjected to the assumption A. Next, denote by E_m and V_m the expectation and variance w.r.t. the model under the assumption A, i.e. arising from the randomness in $\mathbf{y}_j$ conditional x_j. For this we specify the first two model-based moments of $\mathbf{y}_i$ conditional on x_i where, given the bi-partition of $\mathbf{y}_i$,

$$\psi_i = E(\mathbf{y}_i | x_i) = \psi(x_i) = \begin{pmatrix} E(y_{1i}|x_i) \\ E(y_{2i}|x_i) \end{pmatrix} = \begin{pmatrix} \psi_{1i} \\ \psi_{2i} \end{pmatrix}$$

and

$$\tau_i = V(\mathbf{y}_i | x_i) = \tau(x_i) = \begin{pmatrix} V(\mathbf{y}_{1i}|x_i) & Cov(\mathbf{y}_{1i}, \mathbf{y}_{2i}|x_i) \\ Cov(\mathbf{y}_{2i}, \mathbf{y}_{1i}|x_i) & V(\mathbf{y}_{2i}|x_i) \end{pmatrix} = \begin{pmatrix} \tau_{11,i} & \tau_{12,i} \\ \tau_{21,i} & \tau_{22,i} \end{pmatrix}$$

We assume deterministic donor imputation such that there is no extra imputation variance.

Let $x_D = \{x_j; j \in D\}$. The variance of the constrained imputation mean $\bar{\mathbf{Y}}_1^*$ is given by

$$V(\bar{\mathbf{Y}}_1^*) = E_p\{V_m(\bar{\mathbf{Y}}_1^* | x_D)\} + V_p\{E_m(\bar{\mathbf{Y}}_1^* | x_D)\} = E_p\{V_m(\hat{\bar{\mathbf{Y}}}_1 | x_D)\} + V_p\{E_m(\hat{\bar{\mathbf{Y}}}_1 | x_D)\}$$

where $V_m(\hat{\bar{\mathbf{Y}}}_1 | x_D)$ and $E_m(\hat{\bar{\mathbf{Y}}}_1 | x_D)$ are the model variance and expectation of the chosen estimator

$\hat{\bar{\mathbf{Y}}}_1$. Notice that, under any unconstrained deterministic donor imputation, we have

$$\mu_D = E_m(\bar{\mathbf{Y}}^*|x_D) = N^{-1}\sum_{j\in D}(1+d_j)\psi_j \quad \text{and} \quad \Omega_D = V_m(\bar{\mathbf{Y}}^*|x_D) = N^{-2}\sum_{j\in D}(1+d_j)^2\tau_j$$

w.r.t. the model only, because d_j does not depend on $\mathbf{y}$ and is a constant conditional on x_D (and, of course, x_R). It follows that, without the imputation constraint, we would have

$$V(\bar{\mathbf{Y}}_1^*) = N^{-2}E_p\{\sum_{j\in D}(1+d_j)^2\tau_{11,j}\} + N^{-2}V_p\{\sum_{j\in D}(1+d_j)\psi_{1,j}\}$$

Now, we assume that asymptotically as $N\to\infty$, the central limiting theorem applies to the donor imputation, such that $\bar{Y}^*$ attains the multivariate normal distribution $\bar{Y}^* \simeq N(\mu_D,\ \Omega_D)$. It then follows that

$$E_m(\bar{\mathbf{Y}}_2^*|\bar{\mathbf{Y}}_1^*, x_D) = \mu_{D,2} + A_D(\bar{\mathbf{Y}}_1^* - \mu_{D,1}) \qquad \text{for} \quad A_D = \Omega_{D,21}(\Omega_{D,11})^{-1}$$
$$V_m(\bar{\mathbf{Y}}_2^*|\bar{\mathbf{Y}}_1^*, x_D) = \Omega_{D,22.1} = \Omega_{D,22} - \Omega_{D,21}(\Omega_{D,11})^{-1}\Omega_{D,12}$$

where the expectations and covariances correspond to the bi-partition of $\mathbf{y}$. Thus, we have

$$\begin{aligned}
V(\bar{\mathbf{Y}}_2^*) &= V_p\{E_m(\bar{\mathbf{Y}}_2^*|x_D)\} + E_p\{V_m(\bar{\mathbf{Y}}_2^*|x_D)\} \\
&= V_p\{E_m(\bar{\mathbf{Y}}_2^*|x_D)\} + E_p\{V_m(E_m[\bar{\mathbf{Y}}_2^*|\bar{Y}_1^*]|x_D) + E_m(V_m[\bar{\mathbf{Y}}_2^*|\bar{Y}_1^*]|x_D)\} \\
&= V_p(\mu_{D,2}) + E_p\{A_D V_m(\bar{\mathbf{Y}}_1^*|x_D)A_D^T + \Omega_{D,22.1}\} \\
&= E_p(\Omega_{D,22}) + V_p(\mu_{D,2}) - E_p\{\Omega_{D,21}(\Omega_{D,11})^{-1}\Omega_{D,12} - A_D V_m(\bar{\mathbf{Y}}_1^*|x_D)A_D^T\} \qquad (2)
\end{aligned}$$

Under the ReNI method, we have $V_m(\bar{\mathbf{Y}}_1^*|x_D) = V_m(\hat{\bar{\mathbf{Y}}}_1|x_D)$. Whereas, without the imputation constraint, we have $V_m(\bar{\mathbf{Y}}_1^*|x_D) = \Omega_{D,11}$, such that $\Omega_{D,21}(\Omega_{D,11})^{-1}\Omega_{D,12} = A_D V_m(\bar{\mathbf{Y}}_1^*|x_D)A_D^T$ and variance (2) reduces to the unconstrained one, i.e. $V(\bar{\mathbf{Y}}_2^*) = E_p(\Omega_{D,22}) + V_p(\mu_{D,2})$. It follows that the imputation constraint can improve the efficiency of pure donor imputation provided $V_m(\hat{\bar{\mathbf{Y}}}_1|x_D)$ is 'smaller' than $\Omega_{D,11}$.

This can be illustrated using the special case of bi-variate $\mathbf{y}_i^T = (y_{1i}, y_{2i})$ and $\tau_i = \tau$. We have, then, $\Omega_D \to \tau/n = \bar{\tau}$ as $N\to\infty$, where n is the number of donors available. Thus,

$$V_{ReNI}(\bar{Y}_2^*) = V_{DI}(\bar{Y}_2^*) - E_p\{\bar{\tau}_{21}\bar{\tau}_{11}^{-1}(1 - V_m(\hat{\bar{Y}}_1|x_s)/\bar{\tau}_{11})\bar{\tau}_{12}\}$$

where V_{DI} denotes the overall variance in the case of unconstrained donor imputation, and $\bar{\tau}_{11}$ is the variance of $\bar{Y}_1^*$ based on donor-imputation. The ReNI method can reduce the variance of unconstrained donor imputation provided $V_m(\hat{\bar{Y}}_1|x_D) < \bar{\tau}_{11}$. However, in the case of independent y_{1i} and y_{2i}, i.e. $\tau_{12} = \tau_{21} = 0$, we have $V_{ReNI}(\bar{Y}_2^*) = V_{DI}(\bar{Y}_2^*)$, i.e. imposing imputation constratint on $\bar{Y}_1^*$ has no effect on the variance of $\bar{Y}_2^*$, no matter however efficient $\hat{\bar{Y}}_1$ might be.

c. Variance estimation

Assume that an estimator of the model-based variance of $V_m(\bar{\mathbf{Y}}_1^*|x_D) = V_m(\hat{\bar{\mathbf{Y}}}_\mathbf{1}|\mathbf{x_D})$ is given by the method of estimation. Provided model-unbiased $\hat{\bar{\mathbf{Y}}}_1$, we have $E_m(\hat{\bar{\mathbf{Y}}}_1|x_D) = E_m(\bar{\mathbf{Y}}_1|x_D) = \sum_{i\in U}\psi_i$, which is a constant of sampling. Thus, we obtain

$$\hat{V}(\bar{\mathbf{Y}}_1^*) = \hat{V}_m(\hat{\bar{\mathbf{Y}}}_1|x_D)$$

For $V(\bar{\mathbf{Y}}_2^*)$ we need to estimate Ω_D and $V_p(\mu_{D,2})$, in addition to $\hat{V}_m(\hat{\bar{\mathbf{Y}}}_1|x_D)$. Given the variance function $\tau_i = \tau(x_i)$ and its estimator, we may use a direct plug-in estimator

$$\hat{\Omega}_D = N^{-2}\{\sum_{j\in D}(1+d_j)^2\hat{\tau}_j\}$$

However, it can be difficult to specify the variance function $\tau(x_i)$ in practice. An estimator that only depends on the mean assumption of $\psi(x_i)$, but not the variance function, is given by

$$\hat{\Omega}_D = N^{-2}\{\sum_{j\in D}(1+d_j)^2(\mathbf{y}_j - \hat{\psi}_j)(\mathbf{y}_j - \hat{\psi}_j)^T\}$$

Next, consider $V_p(\mu_{D,2}) = V_p\{\sum_{j\in D}(1+d_j)\psi_{2,j}/N\}$. Since $\psi_i = \psi(x_i)$ depends on x_i only, and the assumption A implies that the donors are confined to each imputation class, we have

$$V_p(\mu_{D,2}) = \sum_{k=1}^{K} V_p\{\sum_{j\in D_k}(1+d_j)\psi_{2,j}/N\}$$

where $D_k = D\cap U_k$. However, it seems difficult to fix a particular setting that is general enough for the various situations of statistical registers. Chen and Shao (2000) showed for the NNI method that, under simple random sampling and one imputation class, $V_p(\sum_{j\in s;r_j=1}(1+d_j)\psi_j/n) \to V_m\{\psi(x)\}/n$, where the model-variance is evaluated w.r.t. the marginal distribution of x. We conjecture that the same holds asymptotically under the ReNI method, such that

$$\hat{V}_p(\mu_{D,2}) \to n^{-2}\sum_{k=1}^{K}\{n_k\hat{V}_k(\psi_2)\}$$

where $\hat{V}_k(\psi_2) = \sum_{i\in U_k}(\hat{\psi}_{2i} - \hat{\bar{\psi}}_{2,k})^2/(N-1)$ and $\hat{\bar{\psi}}_{2,k} = \sum_{i\in U_k}\hat{\psi}_{2i}/N$. An estimator of $V(\bar{\mathbf{Y}}_2^*)$ is now obtained by substituting $\hat{V}_p(\mu_{D,2})$, $\hat{\Omega}_D$ and $\hat{V}_m(\hat{\bar{\mathbf{Y}}}_1|x_D)$ into the expression (2).

References

Chen, J.H. and Shao, J. (2000). Nearest neighbor imputation for survey data. *Journal of Official Statistics*, vol. **16**, pp. 113-131.

Zhang, L.-C. and Nordbotten, S. (2008). Prediction and imputation in ISEE: Tools for more efficient use of combined data sources. Paper presented at *Work Session on Statistical Data Editing, Vienna, Austria.*