ESTIMATION WITH NEAREST NEIGHBOUR IMPUTATION AT STATISTICS CANADA

Eric Rancourt, Statistics Canada
Household Survey Methods Division, 16-B, RHC Building, Ottawa, ON, Canada, K1A 0T6, rancour@statcan.ca

Key Words: Distance function, GEIS, GES, jackknife, Model.

1. INTRODUCTION

In conducting surveys, people have always had to face problems of missing data. For a few decades, survey statisticians have made use of donor imputation techniques to treat nonresponse. One only has to think of the time of punch cards when a missing record was replaced by the card of a respondent. That is how hot-deck (a form of donor) imputation was born. Over the years, donor imputation went through several refinements, the major one being making use of auxiliary variables available for both respondents and nonrespondents. In this case, in an attempt to find a donor most similar to the missing record, the closest record (according to some distance measure) is used. This closest matching hot-deck has been called nearest neighbour imputation since Sande (1981).

Already in Sande (1979), a description of the nearest neighbour imputation in the context of numerical imputation for business surveys could be found. Then, Kovar (1982) carried out an empirical study which showed the superiority of nearest neighbour imputation over hot-deck. With the advancement of technology, nearest neighbour imputation has also become easier to program. At Statistics Canada, the Generalized Edit and Imputation System (GEIS) was developed in the mid and late 80's and it certainly explains, at least in part, why in the past ten years nearest neighbour imputation has become so popular. At Statistics Canada, it is now used in a number of agriculture, business and household surveys.

Until recently, the papers of Rancourt, Särndal and Lee (1994) and Chen and Shao (1997), the properties of nearest neighbour imputation were not thoroughly known. However, it seems that the few known characteristics of nearest neighbour imputation were so attractive, that they overruled any possible undesired ones (such as bias and large variance) not very well known. Among these characteristics first comes the fact that nearest neighbour yields a "real" value for nonrespondents. Indeed, since the donated value is provided by a respondent, one can be sure that it is a possible outcome for the variables imputed (as opposed to prediction imputation methods such as mean, ratio or regression). All donor imputation methods share this feature however, but nearest neighbour imputation has the advantage of using auxiliary information. It also has an intuitive appeal. These reasons, plus the good formal properties discussed in Chen and Shao (1997) and in Section 4 now make nearest neighbour imputation a prime candidate for any imputation strategy.

In Section 2, there is a brief description of Statistics Canada's GEIS. As the uses of nearest neighbour imputation have preceded the complete understanding of its properties, Section 3 presents three important applications of nearest neighbour imputation at Statistics Canada before Section 4 which describes nearest neighbour imputation using a model approach. In Section 5, variance estimation is discussed for the model assisted approach and for the jackknife technique as well as in terms of its implementation at Statistics Canada. Finally, a number of issues are presented in Section 6, followed by concluding remarks in Section 7.

2. THE GENERALIZED EDIT AND IMPUTATION SYSTEM (GEIS)

GEIS has been developed and used at Statistics Canada since the mid 80's. It was then designed as part of the Business Survey Redesign Project, an initiative aimed at standardizing survey processes by identifying common steps of surveys and developing generalized systems to process data. On this topic, Outrata and Chinnappa, (1989) present a very interesting discussion of the issues surrounding generalized survey functions.

GEIS is a system which is primarily aimed at satisfying the edit and imputation needs of economic surveys since it is designed for continuous variables. The system is based on Oracle and can run on both Unix and mainframe platforms. The systems has three main features:

i) Editing
ii) Error localization
iii) Imputation.

They are described in a detailed documentation of the system in Cotton (1991, revised 1993).

The first module, for editing is used to define and analyze edit rules. In GEIS, edit rules must be linear equalities or inequalities. It is possible to define edit classes at any level using functions of specified variables. Also, the

system produces a number of diagnostics such as extreme points defined by the edits; implied edits, redundant edits and outliers. It also produces statistics on failure rates of the edits.

The second module is the error localization function. This is the part of the system which determines the amount of changes to bring to a record so that it may satisfy the edits. The system is based on the minimum change principle laid out by Fellegi and Holt (1976).

Then the third part is imputation. There are three types of imputation in GEIS, namely, logical imputation, prediction imputation and donor imputation. Logical imputation is used when only one variable of an edit is missing and it can be deduced from the others. Prediction imputation methods include all methods which impute a value obtained as a function of variables available in the response set and for the sample. Examples of this are mean, ratio and previous value imputation. Finally, donor imputation is the method where values from another record are imputed for the record with missing values. Both hot-deck and nearest neighbour imputation methods are available in GEIS.

For nearest neighbour imputation, a large number of "matching fields" can be specified. These are then used to find the record nearest to the recipient needing imputation. The distance used to find the nearest neighbour is obtained through a series of transformations of the matching variables. For each variable, the transformations are the following:

1) Data are sorted in increasing order;
2) A rank is assigned;
3) Ranks are standardized to the (0-1) scale.

Then the nearest neighbour of a record with missing values is the record which has the minimum value of

$$MAX\left\{\left|Z_{D1}-Z_{R1}\right|,\left|Z_{D2}-Z_{R2}\right|,\ldots,\left|Z_{Dp}-Z_{Rp}\right|\right\}$$

over all units in the response set, where Z_{Dp} is auxiliary variable p of the donor and Z_{Rp} is auxiliary variable p of the recipient.

3. APPLICATIONS OF GEIS AND NEAREST NEIGHBOUR IMPUTATION AT STATISTICS CANADA

Nearest neighbour imputation is used in many surveys at Statistics Canada. For instance, Whitridge and Kovar (1990) present GEIS examples. In this Section, three typical examples of applications from the business, household and agriculture fields are briefly outlined. They are the Unified Enterprise Survey, the Survey of Household Spending and the Financial Farm Survey.

3.1 Unified Enterprise Survey (UES)

UES is an annual survey which is part of the Project to Improve Provincial Economic Statistics (PIPES) and which collects information on enterprises. It allows for production of high quality provincial estimates, which can then be used to redistribute provincial taxes (in the maritime provinces). The UES is an integrated survey of industrial sectors (manufacture, construction, investments, etc.) designed to coordinate efforts and processes. The survey has a two-phase sampling design and uses a mail questionnaire with telephone follow-up.

In the UES, both unit and item imputation are used to compensate for nonresponse. For total nonresponse, tax data are used to perform mass imputation through matching, while nearest neighbour imputation is used for cases of item nonresponse. In this case, GEIS is used to determine the fields to impute and to perform nearest neighbour imputation. The process follows the nearest neighbour imputation description presented in Section 2. A description of the UES imputation strategy is given in Martin, Berniquez and Bernier (1999).

At the estimation stage, Generalized REGression estimators are used to produce domain estimates. Currently, the estimation of precision (variance) is performed on the completed data sets assuming no nonresponse, but development is under way in order to use the SIMPVAR prototype system (described in Section 5) to estimate the variance due to imputation.

3.2 Survey of Household Spending (SHS)

SHS is an annual survey of households aimed at gathering information on a wide variety of categories of household expenditures. This information is used in the PIPES program, for the goods basket used in the price index and for analytical studies. SHS is a multi-stage survey which uses the Labour Force Survey (LFS) frame. That is, it uses the same frame and same sampling scheme, but with an independent sample (not a supplement). The survey data are collected using personal interviews.

Since SHS is a very detailed survey asking a fairly large number of questions, the edit and imputation strategy is divided into 24 subsets of edits (such as those for Income & Taxes, Food, Clothing, Sports, etc.) which are applied within imputation classes. Within each of the classes,

nearest neighbour imputation is performed for continuous variables using GEIS as described in Section 2. More details can be found in Vandermeer (1998). As a result of the imputation, each imputed section of the questionnaire is filled by values of a donor record that is not necessarily the same across edit classes.

In SHS, estimation is performed using the LFS weighting and estimation program. In LFS, weights are obtained using calibration to a number of auxiliary totals such as combinations of age & sex. Then variance estimation is performed with the use of the jackknife technique. The system computes variances on the completed data set assuming that data are as if they were from respondents. Estimation of the variance under imputation is currently being developed for the LFS, with the intent of incorporating the method into the estimation system and transferring it to supplement surveys and those using the same frame as LFS. The LFS uses the jackknife technique which could account for imputation as in Rao and Shao (1992). For nearest neighbour imputation, the required jackknife correction is presented in Section 5.

3.3 Financial Farm Survey (FFS)

FFS is a bi-annual survey collecting information on agriculture operations in Canada. The survey collects information on revenues, balance sheet and investments. Its results are mainly used by Agriculture and Agri-Food Canada and by the Canadian System of National Accounts. The survey is based on stratified simple random sampling and collection is performed through CATI.

The edit and imputation of FFS is carried out using GEIS. First, special attention is devoted to the top 25 records and then GEIS is used for nearest neighbour imputation, again as described in section 2. As in SHS, the data are grouped into classes within which imputation is performed. Details of the imputation approach can be found in Caron (1996), revised by Lalande (1998).

In FFS, estimation is performed using the Horvitz-Thompson estimator and the variance is obtained using the usual variance estimator modified to take into account specific weights. Currently, the estimation system assumes that the completed data are as if they were all obtained by respondents.

4. PROPERTIES OF NEAREST NEIGHBOUR IMPUTATION

The literature on the properties of nearest neighbour imputation is thin. In Rancourt, Särndal and Lee (1994), a model is used to obtain a variance estimator. Steel and Fay (1995) also used a model to adopt the jackknife technique to the replication approach by using the first two nearest neighbours. However, the actual properties of the nearest neighbour imputation method (such as bias) are only presented in Chen and Shao (1997). In this section, after briefly describing the approach of Chen and Shao (1997), we will show that using the model approach of Rancourt, Särndal and Lee (1994) leads to properties which are in agreement with the results of Chen and Shao (1997).

4.1 Bias of nearest neighbour imputation

The objective is to obtain an estimate of $Y_U = \sum_U y_k$ the population total for variable y. Using the sample s, the estimator used is $\hat{Y}_s = \sum_s w_k y_k$, where w_k is the sampling weight. In presence of nonresponse, we have a response set r and a nonresponse set o. The size of r is m and the size of o is $n-m$. In this case, the estimator (for simple random sampling) is $\hat{Y}_{\bullet s} = \frac{N}{n}(\sum_r y_k + \sum_o \hat{y}_k)$ where $\hat{y}_k$ is the imputed value and N and n are respectively the size of the population and the sample.

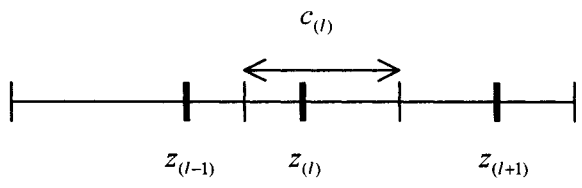
Binomial approach

In the case of only one auxiliary variable, nearest neighbour imputation is obtained through a univariate distance function. With the Euclidean distance function, the imputed value is for unit $\hat{y}_k$: $\hat{y}_k = y_{l(k)}$ where $\|z_k - z_l\|$ is minimum for l among all units in the response set. In other words, l is the donor for unit k, hence $y_{l(k)}$.

To evaluate the properties of the method, Chen and Shao (1997) assumed a uniform response probability mechanism and used the fact that the number of times (t) that a donor is selected follows a Binomial

$$t_l \rightarrow B(n-m, \pi_l),$$

where π_l can be seen as proportional to the distance c_l that there is between the midpoint to two neighbours of unit l on the ordered data set. Note that nonrespondents are assumed to be spread evenly among respondents (uniform response mechanism). Graphically, we have on the ordered set:



and in fact, $\pi_l = F\left(\frac{z_{(l+1)} + z_{(l)}}{2}\right) - F\left(\frac{z_{(l)} + z_{(l-1)}}{2}\right)$, where F is the marginal distribution of z (with $z_{(0)} = -\infty$ and $z_{(m+1)} = +\infty$). Then, the bias is found to be asymptotically 0. Using the same approach, Chen and Shao (1997) obtained a variance expression.

Model approach

Imputation methods can be represented by a model. For example, for ratio imputation, the model is straightforward and its form is $\xi : y_k = \beta z_k + \varepsilon_k$; $E_\xi(\varepsilon_k) = 0$; $E_\xi(\varepsilon_k^2) = \sigma^2 z_k$; and $E_\xi(\varepsilon_k \varepsilon_{k'}) = 0$ for $k \neq k'$. For nearest neighbour imputation, the choice of a model is not obvious. However, upon noting that the matching variable(s) used to find the nearest record need to be correlated to the variable of interest (to preserve joint distributions), it can be seen that a model such as the ratio one has to be at least close to the one for univariate nearest neighbour imputation. In fact, it is the same model, but the values imputed for nonrespondents are different from those imputed by ratio. Comparing the imputed values from three imputation methods, we can see the link to the common model above. Note that Greek letters stand for model parameters, and roman letters for finite population parameters.

Ratio: $\hat{y}_k = \frac{\sum_r y_k}{\sum_r z_k} z_k = \hat{B} z_k$

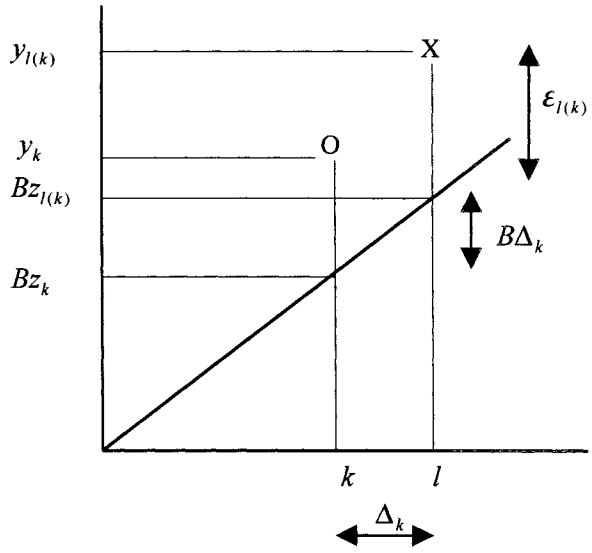
Ratio + residuals: $\hat{y}_k = \frac{\sum_r y_k}{\sum_r z_k} z_k + e_k^* = \hat{B} z_k + e_k^*$,

where e_k^* is a residual randomly selected from the response set.

Nearest neighbour: $\hat{y}_k = y_{l(k)} = B z_{l(k)} + E_{l(k)}$.

For ratio imputation, the model is obviously appropriate. For ratio with added residuals, the model is also appropriate, but a value is generated "around the expectation line". For nearest neighbour imputation, the imputed value is the expected value plus an *actual* residual, not centered on unit k but rather on $l(k)$, the donor.

A look at a graphical representation shows well how the ratio model applies to nearest neighbour imputation.



If we want to evaluate the bias (conditional on s and r), we have:

$$\hat{Y}_{\bullet s} = \frac{N}{n} \sum_s y_{\bullet k} = \frac{N}{n} \left(\sum_r y_k + \sum_o \hat{y}_k \right).$$

It can be written as

$$\hat{Y}_{\bullet s} = \frac{N}{n} \left[\sum_r (B z_k + E_k) + \sum_o (B z_{l(k)} + E_{l(k)}) \right]$$

or since $z_{l(k)} = z_k + \Delta_k$,

$$\hat{Y}_{\bullet s} = \frac{N}{n} \left[B \sum_s z_k + \sum_r E_k + B \sum_o \Delta_k + \sum_o E_{l(k)} \right].$$

Evaluating this latter expression in the case that nonresponse does not depend on y with respect to the model stated above, we obtain:

$$\begin{aligned} \hat{Y}_{\bullet s} &= \frac{N}{n} \left[B \sum_s z_k + B \sum_o \Delta_k \right] \\ &= \frac{N}{n} \left[\frac{\sum_s y_k}{\sum_s z_k} \sum_s z_k + B \sum_o \Delta_k \right] \\ &= \frac{N}{n} \left[\sum_s y_k + B \sum_o \Delta_k \right] \end{aligned}$$

Therefore, the conditional bias of $\hat{Y}_{\bullet s}$ is $\frac{N}{n} B \sum_o \Delta_k$.

If follows that:

1) If there are no nonrespondents, then the bias is 0;
2) As the size m of the response set increases $(m \to \infty)$, then $\Delta_k \to 0$
   a) for continuous distributions;
   b) for distance measures D which have basic properties such as $(y_{l(k)} \to y_k) \Rightarrow (D \to 0)$.

It is also interesting to know that in Sande (1979), it was noted and seen experimentally that norms are locally alike; which seems to dwarf the importance of the latter condition. So under these conditions (which are fairly general) the bias of $\hat{Y}_{\bullet s}$ is asymptotically 0, and this is consistent with the results of Chen and Shao (1997).

4.2 Variance

Following the decomposition in Särndal (1992), we can see that

$$(\hat{Y}_{\bullet s} - Y_U) = (\hat{Y}_s - Y_U) + (\hat{Y}_{\bullet s} - \hat{Y}_s),$$

where $\hat{Y}_s$ is the estimator which would be used in the case of complete response. If the bias is negligible, then we have

$$V_{\text{TOT}} = E_p E_q (\hat{Y}_s - Y_U)^2 + E_p E_q (\hat{Y}_{\bullet s} - Y_s)^2 + \text{Mix T.}$$

which corresponds to

$$V_{\text{TOT}} = V_{\text{SAM}} + V_{\text{IMP}} + V_{\text{MIX}}.$$

The sampling variance is simply the variance of $\hat{Y}_{\bullet s}$ (and $\hat{Y}_s$) with respect to the sampling design. Therefore, it is simply:

$$V_{\text{SAM}} = N^2 \frac{1-f}{n} S_{yU}^2.$$

The imputation variance component was obtained in Forget (1999) by using the model in Section 4.1. The expression is

$$V_{\text{IMP}} = \frac{N^2}{n^2} \left[(\sum_r t_l^2 z_l + \sum_o z_k) \sigma^2 + B^2 \sum_o \Delta_k \right],$$

where t_l is the number of times that donor l is used.

For the mix term, which is not crucial for the discussion in this paper, the reader is referred to Forget (1999) where it is also obtained with the help of the model in Section 4.1.

5. ESTIMATION OF THE VARIANCE

5.1 Model assisted approach

In the model assisted approach presented in Särndal (1992), the goal is to obtain an estimator of the total variance by estimating each of the terms in

$$V_{\text{TOT}} = V_{\text{SAM}} + V_{\text{IMP}} + V_{\text{MIX}}.$$

Estimation of the sampling component

As given in Rancourt, Lee and Särndal (1994), and Forget (1999), an estimator of V_{SAM} can be the ordinary formula:

$$\hat{V}_{\text{ORD}} = N^2 \frac{1-f}{n} S_{y\bullet s}^2.$$

This estimator is good for V_{SAM} provided that $S_{y\bullet s}^2 \approx S_{ys}^2$; which happens if the response set and the imputation classes are large enough. Also, the response mechanism must not depend on the variable of interest or on auxiliary variables correlated with y not used in the imputation process.

Estimation of the imputation component

For the imputation component, all that is necessary is to estimate the B and σ^2 parameters of the model used to obtain the V_{IMP} formula. The second term of the imputation variance, $B^2 \sum_o \Delta_k$, is small and of lower magnitude than the first and can be left out. Then using $\hat{\sigma}^2 = \frac{\sum_r e_k^2}{\sum_r z_k}$ as an estimator for σ^2 and $e_k = y_k - \hat{B} z_k$, we obtain the following estimator for V_{IMP}:

$$\hat{V}_{\text{IMP}} = \frac{N^2}{n^2} (\sum_r t_l^2 z_l + \sum_o z_k) \hat{\sigma}^2.$$

Estimation of the mix component

The estimate for the mix term can be found in Rancourt, Särndal and Lee (1994) or Forget (1999). As it is often zero or close to zero under a variety of conditions, it is not discussed here.

5.2 Jackknife technique

The principle of the jackknife technique is to recalculate the estimator after deleting a unit from the sample and use the variance between the recomputed estimates to obtain an estimate of the variance. After the deletion of unit j, the estimator of the population total is

$$\hat{Y}_{\bullet s}^{(j)} = \frac{N}{n-1}\sum_{k\neq j\in s} y_{\bullet k}$$

where (j) denotes that unit j was deleted. Ignoring the finite population correction, the jackknife variance estimator is

$$\hat{V} = \frac{n-1}{n}\sum_{j\in s}(\hat{Y}_{\bullet s}^{(j)} - \hat{Y}_{\bullet s})^2.$$

For data sets containing imputed values, the jackknife must be corrected. Rao and Shao (1992) proposed a method to correct the estimator by adjusting the imputed values when the j^{th} deleted unit is in the response set. The data set after adjustment of the imputed values is

$$y_{\bullet k}^{(aj)} = \begin{cases} y_k & \text{if } k\in r \\ \hat{y}_k + a_k^{(j)} & \text{if } k\in o \text{ and } j\in r \\ \hat{y}_k & \text{if } k\in o \text{ and } j\in o \end{cases}$$

where $a_k^{(j)}$ is the adjustment. The jackknife variance estimator is then given by

$$\hat{V}_{\text{JACK}} = \frac{n-1}{n}\sum_{j\in s}(\hat{Y}_{\bullet s}^{(aj)} - \hat{Y}_{\bullet s}^{(a)})^2$$

where $\hat{Y}_{\bullet s}^{(aj)} = \frac{N}{n-1}\sum_{k\neq j\in s} y_{\bullet k}^{(aj)}$ and $\hat{Y}_{\bullet s}^{(a)} = \frac{1}{n}\sum_{j\in s}\hat{Y}_{\bullet s}^{(aj)}$.

In Steel and Fay (1995), they proposed corrected variance formulae for nearest neighbour imputation using the ratio model and the second nearest neighbour:

$$\hat{V}_{J,1}^{*} = \frac{N-n}{N}\hat{V}_{JACK} + \frac{n-1}{n}\frac{1}{Nn}\sum_r\sum_{s-r}\left(\frac{\bar{y}_r^{(j)}}{\bar{x}_r^{j}}x_k - \frac{\bar{y}_r}{\bar{x}_r}x_k\right)^2 + \frac{1}{2}\frac{1}{Nn}\sum_{s-r}\left(\frac{y_{l(k)}}{x_{l(k)}}x_k - \frac{y_{l2(k)}}{x_{l2(k)}}x_k\right)^2$$

and

$$\hat{V}_{J,2}^{*} = \frac{N-n}{N}\hat{V}_{JACK} + \frac{n-1}{n}\frac{1}{Nn}\sum_r\sum_{s-r}\left(\frac{\bar{y}_r^{(j)}}{\bar{x}_r^{j}}x_k - \frac{\bar{y}_r}{\bar{x}_r}x_k\right)^2 + \frac{1}{Nn}\sum_{s-r}\left(\frac{y_{l(k)}}{x_{l(k)}}x_k - \frac{\bar{y}_r}{\bar{x}_r}x_k\right)^2.$$

Two versions of the jackknife technique are also considered for nearest neighbour imputation in Chen and Shao (1999), where imputed values are i) partially re-imputed or ii) partially adjusted, based on a probability p_j which depends on the nearest neighbour for partial adjustment; and on the two nearest neighbours for partial re-imputation.

Using a correction approach which avoids the need of finding a second nearest neighbour, Kovar and Chen (1994) used the following (ratio) correction:

$$\hat{y}_k + a_k^{*(j)} = y_{l(k)} + \left(\frac{\bar{y}_r^{(j)}}{\bar{z}_r^{(j)}} - \frac{\bar{y}_r}{\bar{z}_r}\right)z_k.$$

However, according to the model presented in Section 4, the correction should be

$$\hat{y}_k + a_k^{NN(j)} = y_{l(k)} + \left(\frac{\bar{y}_r^{(j)}}{\bar{z}_r^{(j)}} - \frac{\bar{y}_r}{\bar{z}_r}\right)z_{l(k)},$$

since the imputed value is $y_{l(k)}$ and its expectation is $Bz_{l(k)}$ and not Bz_k.

Simulations conducted for a range of situations have shown that the jackknife technique works well with the correction based on $z_{l(k)}$. In fact this correction should work better than the ratio correction as the donors get further away from the recipients.

5.3 Implementation at Statistics Canada

At Statistics Canada the Generalized Estimation System (GES) has been developed to compute estimates and their variance. GES can do domain estimation, and variance estimation is performed using the Taylor approach or the

jackknife technique. The current version of the system assumes that data sets (samples) used as input for estimation are complete. Further, GES treats the data as if they had been obtained from respondents.

The current development schedule and future implementation plans of GES include incorporation of methods which take imputation into account. To fulfill the needs of surveys requiring calculation of the variance due to imputation, a prototype system has been created in the meantime. It serves as a vehicle which will allow for estimation of imputation variance in some applications, and for programming and testing the variance estimation methods to be implemented in GES. The preliminary version of the system is based on the model assisted approach and handles nearest neighbour imputation. Future plans include the implementation of the jackknife technique into GES. Tests are also underway for its use in household surveys.

6. ISSUES

This section addresses two types of issues with respect to nearest neighbour imputation: i) strategy issues related to actual implementation of imputation; and ii) development issues which have yet to be addressed and solved.

Implementation issues

A very important issue to address when imputation is being carried out is that flags (identifiers) must be set up in order to be able, at the estimation stage, to calculate the variance due to imputation. In other words, there has to be interaction between the imputation and the estimation systems, as pointed out in Rancourt (1996).

The flags which are needed are respondent / nonrespondent identifiers; a flag indicating the imputation method that was used and indicating which auxiliary variable(s) was used to perform imputation.

Since nearest neighbour imputation can be represented with the help of the ratio model, under a model framework, it is essential that one of the first steps of the implementation of imputation be the assessment of the goodness of fit of the model within each imputation class. If the model is wrong, then the model assisted approach will suffer accordingly.

As well, since the bias depends on the fact that the size of imputation classes needs to be large (for asymptotic properties to hold), then provision should be made to avoid small imputation classes.

Development issues

This paper has dealt with simple situations such as simple random sampling without replacement. More complex designs need to be studied, but beyond the sampling design, there is a number of issues which remain open.

The bias properties in Section 4.1 remain to be thoroughly validated for complex distance functions. This includes cases where not only other distance measures than the usual Euclidean norm are used, but also when there is more than one matching field. That is, multivariate distance measures.

In many applications, it is rare to see a sole imputation method used for all units. Rather, there is a hierarchy of methods usually ordered according to the availability of the auxiliary information required. Even within the nearest neighbour imputation method, it happens that the auxiliary variables used for matching are not the same for all records, thereby adding complexity to the imputation method. In Shao and Steel (1999), this case is called composite imputation and they are the first to provide elements of solution since Rancourt, Lee and Särndal (1993).

In donor imputation implementations, there is almost always a hierarchy of levels of imputation. First, a donor is searched in a given imputation class. When no record satisfies the edit rules, then a higher level (predefined and usually consisting of groups of first level classes) of imputation classes is used. This process is repeated for records which did not get an imputed value at the first level. More often than not, there are more than two levels of imputation.

7. CONCLUSION

In this paper, we have seen that nearest neighbour imputation has been used for a number years and has found many applications. We have now started to understand all the properties of nearest neighbour imputation, and theory is "catching up" with practice.

Nearest neighbour imputation is widely used at Statistics Canada for its numerous qualities: it yields possible values; it uses auxiliary variables; it is asymptotically unbiased; a nearest neighbour imputation system is available; and methods exist to account for it in variance estimation. And soon, the methods will be available in the variance estimation systems.

Finally, the use of nearest neighbour imputation has proved to be successful thus far, and its increasing use is a trend that will and should continue.

ACKNOWLEDGEMENTS

I am indebted to Jean-François Beaumont for his invaluable comments and insights on this paper.

REFERENCES

Caron, P. (1996), revised by Lalande, D. (1999). Système d'imputation par donneur pour l'Enquête financière sur les fermes de 1998. Business Survey Methods Division, Statistics Canada.

Chen, J. and Shao, J. (1997). Biases and Variances of Survey Estimators Based on Nearest Neighbor Imputation. In *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 365-369.

Chen, J. and Shao, J. (1999). Jackknife variance estimation for nearest neighbor imputation. To appear in *Proceedings of the Section on Survey Research Methods*, American Statistical Association

Cotton, C. (1991, revised 1993). Functional Description of the Generalized Edit and Imputation System. Business Survey Methods Division, Statistics Canada.

Fellegi, I.P. and Holt, D. (1976). A Systematic Approach to Automatic Edit and Imputation. *Journal of the American Statistical Association*, 71, 17-35.

Forget, N. (1999). Estimation de la variance dans les sondages utilisant l'imputation par le plus proche voisin. Master's Thesis, Université de Montréal.

Kovar, J. (1982). A Closer Look at the Nearest Neighbour/Hot Deck Imputation Methods: An Empirical Study. Working Paper, Institution and Agriculture Survey Methods Division, Statistics Canada.

Kovar, J., and Chen, E. (1994). Jackknife Variance Estimation of Imputed Survey Data. *Survey Methodology*, 20, 45-52.

Martin, C., Berniquez, G. and Bernier, J. (1999). UES Edit & Imputation System: Detailed Statement of Requirements. Statistics Canada.

Outrata, E. and Chinnappa, B.N. (1989). General Survey Functions Design at Statistics Canada. *Bulletin of the International Statistical Institute*, 53: 2, 219-238.

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

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., Särndal, C.-E., and Lee, H. (1994). Estimation of the Variance in presence of Nearest Neighbour Imputation. *Proceedings of the Section on Survey Research Methods,* American Statistical Association, 888-893.

Rao, J.N.K. and Shao, J. (1992). Jackknife Variance Estimation with Survey Data under Hot-deck Imputation. *Biometrika*, 82, 453-460.

Sande, G. (1979). Numerical Edit and Imputation. Proceedings of the 42nd Session of the International Statistical Institute, 455-463.

Sande, I. G. (1979). A personal View of Hot Deck Imputation Procedures. *Survey Methodology*, 5, 238-258.

Sande, I. G. (1981). Imputation in Surveys: Coping with reality. *Survey Methodology*, 7, 21-43.

Särndal, C.-E. (1992). Methods for Estimating the Precision of Survey Estimates When Imputation Has Been Used. *Survey Methodology*, 18, 241-252.

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.

Vandermeer, B. (1998). Details of the GEIS Imputation for the 1997 Survey of Household Spending. Household Survey Methods Division, Statistics Canada.

Whitridge, P. and Kovar, J. (1990). Applications of the Generalized Edit and Imputation System at Statistics Canada. *Proceedings of the Section on Survey Research Methods,* American Statistical Association, 105-110.