Balanced k-Nearest Neighbor Imputation

Caren Hasler and Yves Tillé
Institut de Statistique, Université de Neuchâtel, Neuchâtel, SWITZERLAND
e-mail: caren.hasler@unine.ch, yves.tille@unine.ch

Abstract

In order to overcome the problem of item nonresponse, random imputations are often used because they tend to preserve the distribution of the imputed variable. Among the methods of random imputation, the random hot-deck has the interesting property that the imputed values are observed values. We present a new random method of hot-deck imputation which enables us to select the imputed values such that some balancing equations are satisfied and such that the donors are selected in neighborhoods of the recipients.

Keywords: missing data, nonresponse, sampling

1 Introduction

Nonresponse is an important problem in survey. Indeed, the error caused by nonresponse onto the estimates can be more severe than the error caused by the sampling design. Nonresponse arises when a sampled unit does not respond to one or more items of a survey. One differentiates item nonresponse (a sampled unit does not respond to a particular question) from unit nonresponse (a sampled unit does not respond to the entire survey). Reweighting procedures are often used to deal with unit nonresponse whereas imputation methods are used to treat item nonresponse. Imputation is a technique allowing filling the hole due to a missing value.

The imputation methods can be classified into two groups: the deterministic imputation methods and the random imputation methods. The first group contains the methods yielding the same imputed value if the imputation is repeated. Among others deterministic imputation methods, one finds the ratio imputation, the regression imputation, the respondent mean imputation, and the nearest neighbor imputation (Chen and Shao, 2000). This group of methods produces good totals estimations. Nevertheless they often fail to estimate quantiles. The second group contains the methods yielding different imputed values if the imputation is repeated. Among these random methods, one finds the multiple imputation methods presented in Rubin (1987), the imputation with added residuals considered in Chauvet et al. (2010) and in Chauvet et al. (2011), and the random k-nearest neighbor imputation. Unlike the deterministic imputation methods, the random imputation methods have the advantage to tend to preserve the distribution of the imputed variable. Nevertheless such methods imply the presence of an additional term in the variance estimator due to the randomness of imputation, which is called the *imputation variance*. Many authors have been interested in minimizing the imputation variance. For instance, Kalton and Kish (1981) propose to select donors among the respondents without replacement and with a stratification of responses. Chen et al. (2000) propose for this aim to adjust the imputed values, Kim and Fuller (2004) and Fuller and Kim (2005) use the fractional hot-deck imputation, and Chauvet et al. (2010, 2011) propose a balanced random imputation method consisting in randomly select residuals.

One can alternatively classify the imputation methods into the donor imputation methods and the predicted value imputation methods. One denotes by donor imputation methods the fact that the value of a respondent is assigned to a nonrespondent. The unit providing the value is called a *donor* and the unit receiving

the value is called a *recipient*. A hot-deck method is a donor imputation method where a missing value is replaced with an observed value extracted from the same survey. The reader can for instance refer to Andridge and Little (2010) for a review of hot-deck imputation. In contrast, the predicted value imputation methods use function of the respondents values to predict the missing values.

In this paper, we propose a new method of random hot-deck imputation. This method, even though it is random, has the interesting property to reduce the imputation variance compared to other random imputation methods. We called this method the balanced k-nearest neighbor imputation method.

2 Notation and Concepts

Consider a finite population $U = \{1, 2, \ldots, i, \ldots, N\}$ and the variable of interest $\mathbf{y} = (y_1, \ldots, y_i, \ldots, y_N)'$. In a first phase, a random sample S of size n is drawn with a given sampling design $p(S)$. Let $\pi_i = \Pr(i \in S)$ denote the first order inclusion probability of unit i and let w_i denote its Horvitz-Thompson weights $1/\pi_i$ (Horvitz and Thompson, 1952). If a census is considered, the inclusion probabilities and the design weights are equal to 1. In a second phase, a subset of respondents $S_r = \{r_1, r_2, \ldots, r_{n_r}\}$ is drawn from S with an usually unknown conditional distribution $q(S_r|S)$; the values y_i of the variable of interest are known for the units of S_r only. Let $S_m = \{m_1, m_2, \ldots, m_{n_m}\}$ denote the complement of S_r in S, i.e. the subsample of S containing the units with missing data (the nonrespondents). Note the respective sizes of these subsets are n_r and n_m. It is supposed that the units respond independently from each other. Then, each unit $i \in S$ has an usually unknown response propensity $\theta_i = \Pr(i \in S_r | i \in S)$ and $q(S_r|S) = \prod_{i \in S_r} \theta_i \prod_{i \in S_m} (1 - \theta_i)$. In a third phase, nonresponse is corrected through imputation. Imputed values y_j^*, $j \in S_m$ are drawn with a conditional distribution $I\left(y_j^* | S, S_r\right)$.

The aim is to estimate the population total $t_{\mathbf{y}} = \sum_{i \in U} y_i$ of the variable of interest $\mathbf{y}$. In the presence of complete response to the variable of interest $\mathbf{y}$ the estimator $\widehat{t}_{\mathbf{y}} = \sum_{i \in S} w_i y_i$ is adequate. In the presence of nonresponse, the previous estimator is intractable and the imputed estimator $\widehat{t}_{\mathbf{y}}^I = \sum_{i \in S_r} w_i y_i + \sum_{i \in S_m} w_i y_i^*$ is used. A vector $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iq})'$ of q auxiliary variables is assumed to be known for each unit i in the sample S.

The total variance of an imputed estimator (for a total or another statistic of interest) $\widehat{\theta}_I$ can be written

$$\mathrm{Var}\left(\widehat{\theta}_I\right) = \mathrm{Var}_p \mathrm{E}_q \mathrm{E}_I \left(\widehat{\theta}_I\right) + \mathrm{E}_p \mathrm{Var}_q \mathrm{E}_I \left(\widehat{\theta}_I\right) + \mathrm{E}_p \mathrm{E}_q \mathrm{Var}_I \left(\widehat{\theta}_I\right) \qquad (2.1)$$

where the subscripts p, q and I represent respectively the sampling mechanism, the nonresponse mechanism, and the imputation mechanism described above. The first term in (2.1) represents the sampling variance, the second term represents the nonresponse variance and the last term represents the imputation variance.

3 Methodology for random hot-deck donor imputation methods

Random hot-deck donor imputation consists in filling a missing value with an observed value extracted from the same survey; for each nonrespondent, a donor is randomly chosen among the respondents. Consequently, random hot-deck donor imputation can be achieved through the realization of a random matrix $\boldsymbol{\phi} = (\phi_{ij})$,

$(i,j) \in S_r \times S_m$ such that

$$\phi_{ij} = \mathbb{1}_{y_j^* = y_i}. \tag{3.1}$$

As exactly one donor is selected for each nonrespondent, $\boldsymbol{\phi}$ must satisfy

$$\sum_{i \in S_r} \phi_{ij} = 1, \quad \text{for each } j \in S_m. \tag{3.2}$$

However, no conditions are set in $\sum_{j \in S_m} \phi_{ij}$ for $i \in S_r$ as a respondent can impute several nonrespondents. Taking the conditional expectation both sides of equation (3.1) generates a matrix of imputation probabilities $\boldsymbol{\psi} = (\psi_{ij})$, $(i,j) \in S_r \times S_m$

$$\psi_{ij} = \mathrm{E}_I(\phi_{ij}) = \mathrm{E}_I\left(\mathbb{1}_{y_j^* = y_i}\right) = \Pr\left(y_j^* = y_i | S, S_r\right).$$

By definition, $\boldsymbol{\psi}$ satisfies

$$\sum_{i \in S_r} \psi_{ij} = 1, \quad \text{for each } j \in S_m, \tag{3.3}$$

$$0 \leq \psi_{ij} \leq 1, \quad \text{for each } (i,j) \in S_r \times S_m. \tag{3.4}$$

The considered methodology for random hot-deck donor imputation is therefore operated in two stages. In the first stage, the matrix of imputation probabilities ($\boldsymbol{\psi}$) is defined and in the second stage a realization of the matrix of imputation ($\boldsymbol{\phi}$) is carried out. Even though the matrix of imputation probabilities is usually explicitly defined, the matrix of imputation can be hard to generate.

3.1 Random k-nearest neighbor imputation method

We define here the *k-nearest neighbors* of a nonrespondent unit $j \in S_m$ ($\mathrm{knn}(j)$) as its k most similar respondents units $i \in S_r$, i.e. $\mathrm{knn}(j) = \{i \in S_r | \mathrm{rank}\,(d(i,j)) \leq k\}$. It is here considered that $d(\cdot,\cdot)$ is the Mahalanobis distance define through the auxiliary variables, $d(i,j) = \left\{(\mathbf{x}_i - \mathbf{x}_j)' \mathbf{S}^{-1} (\mathbf{x}_i - \mathbf{x}_j)\right\}^{1/2}$ where $\mathbf{S}$ is the variance-covariance matrix of the auxiliary variables. If the values of the auxiliary variables are known on the sample level only, $\mathbf{S}$ must be estimated.

The random k-nearest neighbor imputation method (kNN) consists in filling the missing value of an unit $j \in S_m$ with the value of one of its k-nearest neighbors selected randomly with equal probabilities. It results in the matrix of imputation probabilities $\boldsymbol{\psi}^{(k)} = \left(\psi_{ij}^{(k)}\right)$, $(i,j) \in S_r \times S_m$ with $\psi_{ij}^{(k)} = 1/k \mathbb{1}_{i \in \mathrm{knn}(j)}$. Therefore, the matrix of imputation probabilities related to the kNN is a matrix containing exactly k non-null coefficients in each column and all these non-null coefficients are equal to $1/k$. This particular matrix of imputation probabilities is the starting point of the method proposed in this paper, the balanced k-nearest neighbor imputation method (bkNN).

4 Balanced k-nearest neighbor imputation method

The new method we propose is a method of random hot-deck imputation. This method relies on two main ideas. The first idea is that the donors are chosen in neighborhoods of the recipients; for each nonrespondent, a donor is randomly selected among its k nearest neighbors. The second idea is that the imputation process conserves the estimator of the total of the auxiliary variables; if the auxiliary variables suffered from nonresponse, their imputed estimator of the total would match their total estimator under complete response. Hence, the bkNN involves

a matrix of imputation probabilities $\boldsymbol{\psi}^{(bk)} = \left(\psi_{ij}^{(bk)}\right)$, $(i,j) \in S_r \times S_m$ such that $\psi_{ij}^{(bk)} \neq 0$ only if $i \in \text{knn}(j)$ and such the underlying imputation mechanism implies that conditionally on the sampling mechanism and on the nonresponse mechanism $\widehat{t}_{\mathbf{x}}^{I} = \widehat{t}_{\mathbf{x}}$. It implies that the method we propose involves a matrix of imputation probabilities $\boldsymbol{\psi}^{(bk)} = \left(\psi_{ij}^{(bk)}\right)$ and a matrix of imputation $\boldsymbol{\phi}^{(bk)} = \left(\phi_{ij}^{(bk)}\right)$, $(i,j) \in S_r \times S_m$, such that $\psi_{ij}^{(bk)} \neq 0$ only if $i \in \text{knn}(j)$ and such that

$$\sum_{j \in S_m} w_j \sum_{i \in S_r} \psi_{ij}^{(bk)} \mathbf{x}_i = \sum_{j \in S_m} w_j \mathbf{x}_j \tag{4.1}$$

$$\sum_{j \in S_m} w_j \sum_{i \in S_r} \phi_{ij}^{(bk)} \mathbf{x}_i = \sum_{j \in S_m} w_j \mathbf{x}_j \tag{4.2}$$

$$\mathrm{E}_I\left(\phi_{ij}^{(bk)}\right) = \psi_{ij}^{(bk)} \quad (i,j) \in S_r \times S_m \tag{4.3}$$

are satisfied or almost.

It can be shown that if a strict linear relation between the variable of interest and the auxiliary variables holds, the bkNN provides imputed estimators of the total of the variable of interest with a nearly null imputation variance, i.e. $\mathrm{Var}_I\left(\widehat{t}_{\mathbf{y}}^{I}\right) \approx 0$. Therefore, if the relation between the variable of interest and the auxiliary variables is close to a linear relation, the bkNN is particularly effective in the sense that it is a random imputation method with an imputation variance of the total estimator negligible.

Algorithms 1 and 2 present how the matrix of imputation probabilities $\boldsymbol{\psi}^{(bk)}$ and the matrix of imputation $\boldsymbol{\phi}^{(bk)}$ can be obtained. In order to simplify notation, the indices $i = 1, \ldots, n_r$ and $j = 1, \ldots, n_m$ stand in the Algorithms for the respondent units and for the nonrespondent units respectively.

The main idea of Algorithm 1 is to find a matrix of imputation probabilities $\boldsymbol{\psi}^{(bk)}$ close to the matrix of imputation probabilities relative to the kNN, $\boldsymbol{\psi}^{(k)}$, and satisfying (4.1). The starting point of this Algorithm is therefore the matrix $\boldsymbol{\psi}^{(k)}$. Throughout the process, a null coefficient remains null. The bkNN therefore provides, for each particular nonrespondent, a donor randomly chosen among its k nearest neighbors. Starting with $\boldsymbol{\psi}^{(k)}$ in step 1, calibration and normalization are then alternated in step 2. The calibrations provide matrices $\boldsymbol{\psi}(2l)$ for $l \geq 1$, with nonnegative coefficients and satisfying equation (4.1). However, these matrices $\boldsymbol{\psi}(2l)$ do not necessarily satisfy (3.3) and (3.4). The normalizations provide matrices $\boldsymbol{\psi}(2l+1)$ for $l \geq 1$ satisfying (3.3) and (3.4) but not necessarily satisfying equation (4.1). Taking, in step 3, the limit of the sequence $\boldsymbol{\psi}(2l+1)$ for $l \geq 1$ provides, if the limit exists, the matrix $\boldsymbol{\psi}^{(bk)}$ with the required properties, i.e. satisfying equations (3.3), (3.4), and (4.1) simultaneously.

Once matrix $\boldsymbol{\psi}^{(bk)}$ has been obtained, a realization of matrix $\boldsymbol{\phi}^{(bk)}$, and thus a random selection of donors, has to be carried out. As matrix $\boldsymbol{\psi}^{(bk)}$ satisfies equation (4.1), a necessary and sufficient condition for matrix $\boldsymbol{\phi}^{(bk)}$ to satisfy equation (4.2) is that it satisfies $\sum_{j \in S_m} w_j \sum_{i \in S_r} \phi_{ij}^{(bk)} \mathbf{x}_i = \sum_{j \in S_m} w_j \sum_{i \in S_r} \psi_{ij}^{(bk)} \mathbf{x}_i$ which can be rewritten

$$\sum_{j \in S_m} \sum_{i \in S_r} \frac{w_j \psi_{ij}^{(bk)} \mathbf{x}_i}{\psi_{ij}^{(bk)}} \phi_{ij}^{(bk)} = \sum_{j \in S_m} \sum_{i \in S_r} w_j \psi_{ij}^{(bk)} \mathbf{x}_i. \tag{4.4}$$

This equation is a typical equation of balancing. However, as a donor can be used to impute several nonrespondents, it is a matter of a selection with replacement. The

Algorithm 1 Procedure to obtain the matrix of imputation probabilities $\boldsymbol{\psi}^{(bk)}$

Step 1:
- Set $\boldsymbol{\psi}(1) = \boldsymbol{\psi}^{(k)}$, the matrix of imputation probabilities relative to the kNN.
- Add a constant auxiliary variable if none of the auxiliary variables is a constant.
- Let $\mathbf{X}_r$ be the matrix of dimension $n_r \times q$ whose i-th row is $\mathbf{x}_i$, for $i = 1, \ldots, n_r$.
- Let $\mathbf{t} \in \mathbb{R}^q$ be the vector $\sum_j w_j \mathbf{x}_j$.

Step 2: For $l \geq 1$,
- Let $\mathbf{d} \in \mathbb{R}^{n_r}$ be the vector $d_i = \sum_j \psi(2l-1)_{ij} w_j$ for $i = 1, \ldots, n_r$.
- Obtain $\mathbf{g} \in \mathbb{R}^{n_r}$ through calibration on $\mathbf{X}_r$, with initial weights $\mathbf{d}$, with total $\mathbf{t}$, and with the raking method.
- Let $\boldsymbol{\psi}(2l)$ be the matrix defined as $\psi(2l)_{ij} = \psi(2l-1)_{ij} g_i$.
- Let $\boldsymbol{\psi}(2l+1)$ be the matrix defined as $\psi(2l+1)_{ij} = \frac{\psi(2l)_{ij}}{\sum_i \psi(2l)_{ij}}$.

Step 3: Set $\boldsymbol{\psi}^{(bk)} = \lim_{l \to +\infty} \boldsymbol{\psi}(2l+1)$.

Cube Method proposed by Deville and Tillé (2004) permits to select balanced samples, but it is adequate for the purpose of balanced sampling without replacement. As a solution, Chauvet et al. (2011) propose to achieve balanced sampling with replacement through balanced sampling without replacement within a population of cells. Indeed, it is a matter of selection in a population of cells $(i, j) \in S_r \times S_m$ without replacement as a nonrespondent unit $j \in S_m$ is imputed exactly once. The procedure to obtain matrix $\boldsymbol{\phi}^{(bk)}$ use this idea and is presented in Algorithm 2. In step 1, a vector of strata is created. To each nonrespondent $j \in S_m$, a stratum consisting of the cells (i, j) for $i \in S_r$ is attached. Selection with replacement of one donor for each nonrepondent can thus be achieved through selection of exactly one unit in each stratum. In step 2, balancing variables are defined and in step 3 a sample is selected through stratified balanced sampling. A suitable stratified balanced sampling algorithm for this purpose is proposed in Hasler and Tillé (2013). Indeed, it permits to select exactly one donor in each stratum while satisfying as good as possible the balancing equations. The selected sample is a vector and this one provides $\boldsymbol{\phi}^{(bk)}$ in step 4 by filling a column matrix of the same dimension than initial matrix $\boldsymbol{\psi}^{(bk)}$. Algorithm 2 produces a matrix of imputation $\boldsymbol{\phi}^{(bk)}$ satisfying equations (3.2) and (4.3), and satisfying (or almost) equation (4.4). As a matter of fact, matrix $\boldsymbol{\phi}^{(bk)}$ satisfies or almost

$$\sum_{i \in S_r} \frac{w_j \psi_{ij}^{(bk)} \mathbf{x}_i}{\psi_{ij}^{(bk)}} \phi_{ij}^{(bk)} = \sum_{i \in S_r} w_j \psi_{ij}^{(bk)} \mathbf{x}_i$$

for each $j \in S_m$, which is stronger than (4.4).

5 Conclusion

In this article, we have proposed a new method of random hot-deck imputation which we call the balanced k-nearest neighbor imputation method. This method has the interesting properties to be a donor imputation and to select the donors in neighborhoods of the recipients. As it is random, this method is adequate to total as well as to quantiles estimation. Moreover, this method cancels the imputation variance of the total of the auxiliary variable. It implies that the imputation variance

Algorithm 2 Procedure to obtain the matrix of imputation $\boldsymbol{\psi}^{(bk)}$

Step 1: Let $\mathbf{h}$ be the vector of dimension $n_r n_m \times 1$, whose first n_r coordinates equal 1, whose next n_r coordinates equal 2, and so on up to n_m.

Step 2: (a) Let $\mathbf{a}_{ij} = w_j \psi_{ij}^{(bk)} \mathbf{x}_i$ for $i = 1, \ldots, n_r; j = 1, \ldots n_m$.

(b) Construct the matrix $\mathbf{A}$ of dimension $n_r n_m \times q$
$\mathbf{A} = \begin{pmatrix} \mathbf{a}_{11} & \mathbf{a}_{21} & \cdots & \mathbf{a}_{n_r 1} & \mathbf{a}_{12} & \mathbf{a}_{22} & \cdots & \mathbf{a}_{n_r n_m} \end{pmatrix}'$.

(c) Let $\dot{\boldsymbol{\pi}}$ be the vector composed of the columns of matrix $\boldsymbol{\psi}^{(bk)}$.

Step 3: Select a stratified sample balanced on $\mathbf{A}$ with inclusion probabilities $\dot{\boldsymbol{\pi}}$ with the method proposed by Hasler and Tillé (2013). The vector of integers that specifies the stratification is vector $\mathbf{h}$.

Step 4: Obtain the matrix $\boldsymbol{\phi}^{(bk)}$ by filling by column a matrix of dimension $n_r \times n_m$ with the selected sample.

of the target estimator $\widehat{t}_{\mathbf{y}}^{I}$ is reduced; it can even cancel if the relation between the variable of interest and the auxiliary variables is strictly linear.

References

Andridge, R. R. and Little, R. J. A. (2010). A review of dot deck imputation for survey non-response. *International Statistical Review*, 78:40–64.

Chauvet, G., Deville, J.-C., and Haziza, D. (2010). Adapting the cube algorithm for balanced random imputation in surveys. Technical report, ENSAI, Rennes.

Chauvet, G., Deville, J.-C., and Haziza, D. (2011). On balanced random imputation in surveys. *Biometrika*, 98:459–471.

Chen, H. L., Rao, J. N. K., and Sitter, R. R. (2000). Efficient random imputation for missing survey data in complex survey. *Statistica Sinica*, 10:1153–1169.

Chen, H. L. and Shao, J. (2000). Nearest-neighbour imputation for survey data. *Journal of Official Statistics*, 16:113–131.

Deville, J.-C. and Tillé, Y. (2004). Efficient balanced sampling: The cube method. *Biometrika*, 91:893–912.

Fuller, W. A. and Kim, J. K. (2005). Hot-deck imputation for the response model. *Survey Methodology*, 31:139–149.

Hasler, C. and Tillé, Y. (2013). Fast balanced sampling for highly stratified population.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47:663–685.

Kalton, G. and Kish, L. (1981). Two efficient random imputation procedures. In *ASA Proceedings of the Section on Survey Research Methods*, pages 146–153. American Statistical Association.

Kim, J. K. and Fuller, W. A. (2004). Fractional hot-deck imputation. *Biometrika*, 91:559–578.

Rubin, D. B. (1987). *Multiple Imputation for Nonreponse in Surveys*. Wiley, New York.