Working Paper

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Workshop on Statistical Data Editing
(Neuchâtel, Switzerland, 18-20 September 2018)

Balanced Imputation for Swiss Cheese Nonresponse

Prepared by Audrey-Anne Vallée and Yves Tillé, Institute of Statistics, University of Neuchâtel, Switzerland

I. INTRODUCTION

1. Nonresponse is often inevitable in large scale surveys. There are two types of nonresponse: unit nonresponse and item nonresponse. Unit nonresponse occurs when the values of every variables of the survey are missing for some sampled units. Item nonresponse occurs when some, but not all, variables are missing for a set of sampled units. The estimators of the parameters of interest may be seriously affected by the missing values, which can introduce a bias and cause an additional variability. Reweighting the respondent units and imputing the missing values allow to reduce the bias and the variance caused by nonresponse.

2. In surveys with multiple variables of interest, different types of item nonresponse can occur. In a first case, only one variable contains missing values. The other variables of the survey are completely observed and they can be used to impute the missing values. Several imputation methods have been developed in this context. Haziza [2009] presents an overview of deterministic and random imputation methods, including multiple and fractional imputation. Andridge and Little [2010] present an overview of donor imputation methods. In a second case, nonresponse can affect multiple variables monotonously. This pattern is suitable to longitudinal studies, when units stop participating in surveys over time. The third case is treated in this paper: swiss cheese nonresponse, or non-monotone nonresponse. This type of nonresponse occurs when all the variables of a survey contain missing values without a particular pattern. There are few treatments for such a multivariate nonresponse. It is difficult to preserve the distributions of the variables and the relationships between the variables with such missing values in the dataset. Judkins [1997] proposed donor imputation methods and Andridge and Little [2010] present an overview of existing methods. Some methods allow to impute iteratively; for instance Raghunathan et al. [2001] use a sequence of regression models between the variables.

3. In this paper, balanced K-nearest neighbor imputation [Hasler and Tillé, 2016] is extended to the swiss cheese nonresponse case. The properties and the advantages of this method justify the interest of developing it for the multivariate case. It is a donor imputation method, so a random imputation method which tends to preserve the distributions of the variables. Balanced imputation is used to reduce the additional variability caused by the random method. A donor imputation method also allows to impute a dataset containing continuous and categorical variables. In addition, only one donor is selected to replace all missing values of a nonrespondent. In the multivariate case, this should ensure coherence between the imputed values of a unit. Also, a nonrespondent can be imputed by donors that are close, or similar, to it. Last, with calibration methods and balanced sampling methods, the donors are selected so that if the observed values were imputed, the imputed total estimators and

the observed total estimators should be the same. The context and the requirements of the method are presented in Sections II and III. The construction of the matrix of imputation probabilities is detailed in Section IV. The selection of the donors is treated in Section V and the imputation in Section VI.

II. Swiss cheese nonresponse

1. Consider a finite population U of size N with J variables of interest. A sample s of size n is randomly selected with respect to a sampling design $p(s)$. The inclusion probability of order one of unit k is π_k and the inclusion probability of order two of units k and ℓ is $\pi_{k\ell}$. The vector of J variables of interest, $\mathbf{x}_k = (x_{k1}, \ldots, x_{kj}, \ldots, x_{kJ})^\top$, is not necessarily fully observed for all $k \in s$. It is expected that a set of sampled units is completely observed, while all the variables of the rest of the sample are subject to nonresponse. Consider $s_r \subset s$ a set of n_r units for which the J variables are completely observed. Consider $s_m = s - s_r$, a set of $n_m = n - n_r$ units such that some values, but not all, are missing. The nonresponse is non-monotone, it has no particular pattern.

2. Suppose that the missing values are treated by imputation. The imputed value of unit k for the variable j is x^*_{kj}. Then the population total of the variable j, $X_j = \sum_{k\in U} x_{kj}$, can be estimated by

$$\widehat{X}_j = \sum_{k\in s} r_{kj} d_k x_{kj} + \sum_{k\in s} (1 - r_{kj}) d_k x^*_{kj},$$

where $d_k = \pi_k^{-1}$ is the sampling weight of unit k and r_{kj} is 1 if the variable j of unit k is observed and 0 otherwise.

III. Requirements

1. The proposed method is elaborated to ensure coherence and accuracy in the imputed dataset. Four requirements are stated:

(i) The imputed values should be selected among the values of the n_r completely observed units: a donor imputation method should be used.
(ii) Only one donor should be selected per unit: all the missing values of a unit should be imputed by the same donor.
(iii) The donors should be selected among the K nearest neighbors of the unit with missing values.
(iv) If the observed values of the nonrespondents were imputed, the total estimator of each variable should remain unchanged.

2. Requirement (i) ensures that the imputed values are realistic and observed, for both categorical and continuous variables. Also, a random imputation method tends to preserve the distributions of the variables. The aim of requirement (ii) is to preserve the relationships between the variables. Requirement (iii) allows the imputation of a nonrespondent by a similar unit and ensures a coherence between the imputed values and the observed values of the nonrespondent. For instance, if the gender and the height of people are measured, a missing height of a man should be imputed by the height of a man. The idea behind requirement (iv) is that the observed information is unchanged if the units with missing values were completely imputed. The estimators based on known values would not be affected.

3. To implement a donor imputation method, each fully observed unit receives a probability to donate its values to each nonrespondent. Then, one donor per nonrespondent can be selected with

respect to those imputation probabilities. The imputation probabilities respecting requirements (i)-(iv) are detailed in Section IV. The selection of the donors is detailed in Section V.

IV. Matrix of imputation probabilities

1. The first step of a donor imputation method is to assign imputation probabilities to complete respondents. Consider $\boldsymbol{\psi} = (\psi_{ik})$, where $(i,k) \in s_r \times s_m$, the matrix of imputation probabilities. The element ψ_{ik} is the probability that respondent i gives its values to nonrespondent k and

$$\psi_{ik} \geq 0. \tag{1}$$

Only one donor is randomly selected for each unit $k \in s_m$ with respect to the imputation probabilities. The sum of the imputation probabilities associated with a nonrespondent should be 1,

$$\sum_{i \in s_r} \psi_{ik} = 1, \tag{2}$$

for all $k \in s_m$. All the missing values of unit k are imputed by the corresponding values of the donor. Requirements (i) and (ii) are then fulfilled. Requirement (iii) limits the set of possible donors of a unit to its K nearest neighbors. In this case, the probability that unit $i \in s_r$ gives its values to the nonrespondent $k \in s_m$ is non-zero only if i is one of the K nearest neighbors of k;

$$\psi_{ik} = 0 \text{ if } i \notin \text{kpp}(k), \tag{3}$$

where $\text{kpp}(\ell) = \{j \in s_r | \, \text{rank}(d(j,\ell)) \leq K\}$ and $d(.,.)$ is a distance function.

2. Requirement (iv) suggests that if the observed values of unit $k \in s_m$ were imputed by the corresponding values of the donor, the total estimator of each variable would remain the same as the total estimator calculated with the observed values only. The imputation probabilities are then chosen so that if the known values of the units in s_m were imputed by the expectation of their imputed value, the total estimators would correspond to the estimators based on the observed values. So the imputation probabilities respect

$$\sum_{k \in s_m} d_k r_{kj} \sum_{i \in s_r} \psi_{ik} x_{ij} = \sum_{k \in s_m} d_k r_{kj} x_{kj}, \tag{4}$$

for $j = 1, \ldots, J$.

3. Equation (4) can be rewritten as

$$\sum_{i \in s_r} \left(\sum_{k \in s_m} d_k r_{kj} \psi_{ik} \right) r_{ij} x_{ij} = \sum_{k \in s_m} d_k r_{kj} x_{kj}, \tag{5}$$

for $j = 1, \cdots, J$. The imputation probabilities respecting (5) can be found by calibration [Deville and Särndal, 1992]. Consider the initial imputation probabilities

$$\psi_{ik}^0 = \begin{cases} \dfrac{1}{K} & \text{if } i \in \text{kpp}(k), \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

Final imputation probabilities ψ_{ik} close to ψ_{ik}^0 respecting (5) are sought. Consider the distance function

$$G(\psi_{ik}, \psi_{ik}^0) = \psi_{ik} \log(\frac{\psi_{ik}}{\psi_{ik}^0}) + \psi_{ik}^0 - \psi_{ik},$$

then the elements of $\boldsymbol{\psi}$ are obtained by minimizing

$$\mathcal{L} = \sum_{k\in s_m}\sum_{i\in s_r} G(\psi_{ik},\psi_{ik}^0) - \sum_{j=i}^{J}\lambda_j\left[\sum_{i\in s_r}\sum_{k\in s_m} d_k r_{kj}\psi_{ik}r_{ij}x_{ij} - \sum_{k\in s_m} d_k r_{kj}x_{kj}\right].$$

Using

$$\frac{\partial\mathcal{L}}{\partial\psi_{ik}} = \log\frac{\psi_{ik}}{\psi_{ik}^0} - \sum_{j=1}^{J}\lambda_j d_k r_{kj}r_{ij}x_{ij} = 0,$$

the imputation probabilities are

$$\psi_{ik} = \psi_{ik}^0 \exp\left[\sum_{j=1}^{J}\lambda_j d_k r_{kj}r_{ij}x_{ij}\right]. \tag{7}$$

Calibration techniques are used to find $\lambda = (\lambda_1,\ldots,\lambda_J)^\top$ respecting Equations (1)-(4).

V. Imputation matrix

1. Once the matrix of imputation probabilities $\boldsymbol{\psi}$ is completed, the donors can be randomly selected. Consider $\boldsymbol{\phi} = (\phi_{ik})$, where $(i,k) \in s_r \times s_m$, the imputation matrix. The element ϕ_{ik} is 1 if unit i is selected to donate its values to unit k, 0 otherwise. Only one donor is selected per nonrespondent, so

$$\sum_{i\in s_r}\phi_{ik} = 1.$$

To respect requirement (iv), the donors should be selected so that

$$\sum_{k\in s_m}\sum_{i\in s_r}\frac{\phi_{ik}}{\psi_{ik}} d_k r_{kj}\psi_{ik}x_{ij} = \sum_{k\in s_m}\sum_{i\in s_r} d_k r_{kj}\psi_{ik}x_{ij}. \tag{8}$$

Balanced sampling [Deville and Tillé, 2004] is used to respect the balancing constraints (8). To ensure that only one donor is selected per nonrespondent and that the balancing constraints are respected once the donors are selected, the matrix $\boldsymbol{\phi}$ is generated with stratified balanced sampling [Chauvet, 2009, Hasler and Tillé, 2014]. A total of n_m strata are created and one donor is selected per stratum. A stratum corresponds to a nonrespondent. The inclusion probability used in the stratified balanced sampling is ψ_{ik} and the associated balancing variable is $d_k r_{kj}\psi_{ik}x_{ij}$, for $(i,k) \in s_r \times s_m$.

VI. Imputation and discussion

1. The imputation of the dataset is based on the matrix $\boldsymbol{\phi}$. The missing value of unit k to variable j such that $r_{kj} = 0$ is imputed by

$$x_{kj}^* = \sum_{i\in s_r}\phi_{ik}x_{ij}.$$

Requirements (i)-(iii) are perfectly respected. Requirement (iv) is either perfectly of approximately respected, according to the limits of the balancing techniques.

2. It is also possible to use a deterministic version of the proposed imputation method. The expectation of ϕ_{ik} is used for $(i,k) \in s_r \times s_m$, this is ψ_{ik}. Then the missing value of unit $k \in s_m$ for j such that $r_{kj} = 0$ is imputed by

$$x_{kj}^* = \sum_{i\in s_r}\psi_{ik}x_{ij}.$$

It is not a donor imputation method anymore, but requirement (iv) is perfectly respected.

3. Random imputation methods generally cause an additional variability of the imputed total estimator. Balanced imputation allows to minimize this variability. In further works, a variance estimator should be developed and a simulation study should illustrate the properties of imputed estimators and variance estimators.

References

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

Guillaume Chauvet. Stratified balanced sampling. *Survey Methodology*, 35:115–119, 2009.

Jean-Claude Deville and C.-E. Särndal. Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87:376–382, 1992.

Jean-Claude Deville and Yves Tillé. Efficient balanced sampling: The cube method. *Biometrika*, 91: 893–912, 2004.

Caren Hasler and Yves Tillé. Fast balanced sampling for highly stratified population. *Computational Statistics and Data Analysis*, 74:81–94, 2014.

Caren Hasler and Yves Tillé. Balanced k-nearest neighbor imputation. *Statistics*, 105:11–23, 2016.

David Haziza. Imputation and inference in the presence of missing data. In Danny Pfeffermann and C. R. Rao, editors, *Sample surveys: Design, methods and applications*, pages 215–246. Elsevier/North-Holland [Elsevier Science Publishing Co., New York; North-Holland Publishing Co., Amsterdam], 2009.

David R. Judkins. Imputing for Swiss cheese patterns of missing data. In *Proceedings of Statistics Canada Symposium*, page 97, Statistics Canada, 1997.

Trivellore E. Raghunathan, James M. Lepkowski, John van Hoewyk, and Peter W. Solenberger. A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, 27(1):85–95, 2001.