A Generalized Framework to Evaluate Imputation Strategies: Recent Developments[1]

Darren Gray
Statistics Canada, 150 Tunney's Pasture Driveway, Ottawa, ON K1A 0T6

Abstract
For many National Statistical Organizations, imputation is the preferred treatment for item non-response. Consequently, the choice of imputation strategy can have a significant impact on resulting statistical estimates. Recently, a generalized framework to evaluate and improve imputation strategies at Statistics Canada was proposed and used to examine the choice of imputation strategies within the confines of the Canadian Census Edit and Imputation System (CANCEIS). The goal now is to develop a generalized, user-friendly tool for survey methodologists, managers and statisticians, allowing them to assess and compare imputation strategies on existing datasets. The focus of this paper is on the development of the tool itself, in particular the choice of simulation parameters and output measures. Finally, we explore how best to present results, including data visualization, to facilitate data-driven decision-making in survey design.

Key Words: Imputation, simulation, data visualization

1. Introduction

Missing data is an issue affecting all National Statistical Organizations (NSOs). In general, missing data for statistical products can be classified as either *unit non-response* (all data associated with a record is missing) or *item non-response* (some data associated with a record is missing). Imputation is a common method for combating missing data, in particular item non-response; for a general discussion on the topic, and statistical data editing in general, we refer the reader to De Waal, Pannekoek & Scholtus (2011).

Statistics Canada has developed two generalized tools for statistical data editing and imputation: Banff (Statistics Canada, 2017), designed primarily for economic statistics and numerical variables, and the Canadian Census Edit and Imputation System (CANCEIS), designed for the Census as well as household and social surveys. This project is motivated by the author's role in the support of Banff users, and research into imputation methods.

In a Banff support role, we frequently encounter questions about imputation methods, design, and strategies, and how these can and should be tested. Common questions fall along these lines:

- Which imputation method is most effective?

[1] The content of this paper represents the position of the author and may not necessarily represent that of Statistics Canada.

- How should parameters be chosen?
- What effect does non-response and imputation have on the quality of statistical estimates?

More recently, modernization efforts at Statistics Canada have led to new questions regarding non-response and imputation. These include an increased interest in assessing the quality of administrative datasets that have undergone editing, and curiosity about third-party open-source imputation software, including popular R packages such as *missForest* (Stekhoven, 2015) and *mice* (van Buuren & Groothuis-Oudshoorn, 2010).

To meet these demands, we envision a generalized tool allowing users to assess and compare imputation strategies, with the goal of facilitating data-driven decision-making in survey design. Importantly, such a tool should be general enough to meet the following objectives:

- A simple, intuitive, and reproducible framework for evaluating methods
- The ability to investigate and assess arbitrary imputation methods, including "black boxes"
- A suite of analysis tools suitable for a wide variety of intended data uses

2. A Framework for Assessing Imputation

As missing data has long been a thorn in the side of statisticians, there is a long history of imputation methods and various criteria for evaluating them. Chambers (2001) laid out the following five "Performance measurements for imputation":

1. **Predictive Accuracy:** The imputation procedure should maximise preservation of true values. That is, it should result in imputed values that are 'close' as possible to the true values.
2. **Ranking Accuracy:** The imputation procedure should maximise preservation of order in the imputed values. That is, it should result in ordering relationships between imputed values that are the same (or very similar) to those that hold in the true values.
3. **Distributional Accuracy:** The imputation procedure should preserve the distribution of the true data values. That is, marginal and higher order distributions of the imputed data values should be essentially the same as the corresponding distributions of the true values.
4. **Estimation Accuracy:** The imputation procedures should reproduce the lower order moments of the distributions of the true values. In particular, it should lead to unbiased and efficient inferences for parameters of the distribution of the true values (given that these true values are unavailable).
5. **Imputation Plausibility:** The imputation procedure should lead to imputed values that are plausible. In particular, they should be acceptable values as far as the editing procedure is concerned.

Chambers goes on to define "imputation performance measures" for various variable types. A key caveat of the measures are that they are designed to assess "the performance of an editing and imputation method when the 'true' values underpinning either the incorrect or missing data items are known".

One method to compare imputed values to true values, is by starting with the true values and simulating non-response.

Previous work of this type at Statistics Canada includes simulation tools developed by Haziza (2003) and Stelmack (2018). Our contribution builds on the work of Haziza and Stelmack but aims to further generalize the work by increasing the options available for generating non-response, and a wider array of analysis options. In particular, whereas Chambers, Haziza and Stelmack analyze imputation results using numerical measures, we focus on the benefits of data visualization analysis.

Section 3 gives an overview of the non-response and imputation process. In Section 4 we discuss the analysis modules available, with examples.

3. Non-Response and Imputation

Let the original data consist of a numerical variable of interest, $y = \{y_1, \dots, y_n\}$, along with any auxiliary data that may be required for either imputation or analysis. For the purposes of assessing imputation performance, the y variable must consist only of 'true' values, i.e., non-missing and without error.

Let $j = \{1, \dots, m\}$ be the index of simulations we intend to run. Within each simulation, we select a sub-sample of units and set the corresponding y-values to missing. We use the binary indicator δ_{ij} to denote missingness: $\delta_{ij} = 1$ if y_i is missing in simulation j and zero otherwise.

The choice of non-response matrix, $[\delta_{ij}]$, and the method in which it is generated, will impact any conclusions that can be made at the analysis phase. In the tools developed by Haziza and Stelmack, the set of missing data in each individual simulation is generated via a Poisson sampling process with each unit having an independent probability of non-response, p_i, set by the user, or constructed by the simulation tool. Various non-response mechanisms can be modelled by the simulation, depending on how the p_i values are assigned:

- Missing Completely at Random (MCAR): The probability of non-response is constant for each unit.
- Missing at Random (MAR): The probability of non-response depends on observed data.
- Missing Not at Random (MNAR): The probability of non-response depends on unobserved data.

A common approach for simulation studies is to define p_i as a function of auxiliary data (for MAR) or the variable of interest itself (for MNAR).

Under Poisson sampling, the number of missing values can vary in each simulation. Additionally, as each simulation independently selects a Poisson sample, there is no control over the number of times each unit is imputed over all simulations. Rosen (1997a, 1997b) introduced Pareto sampling as a method of variable-inclusion probability sampling with a fixed sample size; in this respect it is similar to Conditional Poisson sampling but more efficient for our purposes. The current tool also provides a k-fold cross-validation module, which fixes the number of times each unit is imputed over all

simulations. In total, the current tool includes the following five sampling methods for generating the non-response indicators δ_{ij}:

- Poisson sampling
- Bernoulli sampling
- Pareto sampling
- Simple random sampling without replacement (SRSWOR)
- k-fold cross-validation

Bernoulli sampling and SRSWOR are simply special cases of Poisson and Pareto sampling when the non-response probabilities are constant; they are included for user convenience.

After generating non-response, users are responsible for imputing the resulting dataset. The resulting values are denoted $\hat{y}_{ij}$; we note that by design $\hat{y}_{ij} = y_i$ when $\delta_{ij} = 0$. The only imputation requirements are as follows:

- For inferential purposes, imputation must be performed independently on each simulation.
- Each missing value must be imputed with a single numerical value.

In particular, users must make sure that all values requiring imputation are imputed; this is a requirement for the analysis portion of the simulation.

The tool includes only one built-in imputation method: random hot deck donor imputation. This simple imputation may or may not be appropriate in practice, but is included in the tool as an optional baseline test for comparison purposes.

4. Analysis

Before analysis, it is important to determine an evaluation criteria for assessing imputation methods. According to the Generic Statistical Data Editing Model (United Nations Economic Commission for Europe, 2019), the primary goal of statistical data editing is the "treatment of the data to achieve fitness for use". Chambers, when discussing the five performance measures for imputation referenced in Section 2, argues a similar point:

"Nor are the properties themselves mutually exclusive. In fact, in most uses of imputation within NSIs the aim is to produce aggregates estimates from a data set, and criteria (1) and (2) below will be irrelevant. On the other hand, if the data set is to be publicly released or used for development of prediction models, then (1) and (2) become rather more relevant."

Along those lines, the current tool includes three analysis modules:

1) Univariate Distribution Analysis
2) Estimator Analysis
3) Predictive Analysis

These three analysis modules are suitable for a variety of data needs. We note that they correspond to three of the five imputation performance measurements proposed by Chambers.

While Chambers, Haziza, and Stelmack all propose numerical assessment measures, we have instead decided to focus on data visualization for our analysis. We believe that the concepts these aim to measure can be sufficiently captured by data visualization techniques. Additionally, data visualization offers some of the following benefits over numerical outputs:

- **A well designed visual output is intuitive.** We want users to make appropriate inferences and ultimately, data-driven decisions based on the simulation studies. When done appropriately, visuals can highlight and convey information to the user much more efficiently than tables of numerical outputs.
- **Visualization is comprehensive.** Numerical outputs have the benefit of brevity, but come with a degree of information loss. For certain types of information, this trade-off is unnecessary.
- **Visualization allows for exploratory analysis.** In particular, visualization may reveal patterns in the data that summary statistics do not – one of the reasons statisticians are expected to plot residuals when fitting a model.

To demonstrate these benefits, we include three examples in this paper. For the purpose of these examples, we've selected a publicly available data set from the University of California, Irvine (UCI) Machine Learning Repository (Dua & Graff, 2019). The Residential Building Data Set (Rafiei & Adeli, 2015) consists of 372 records and 105 variables. From this data set we chose a single variable of interest, and a selection of auxiliary variables for imputation purposes.

4.1 Univariate Distribution Analysis

The objective of this module is to determine the effectiveness of an imputation method at preserving the univariate distribution of the variable of interest. Visual comparison is an intuitive choice for comparing univariate distributions.

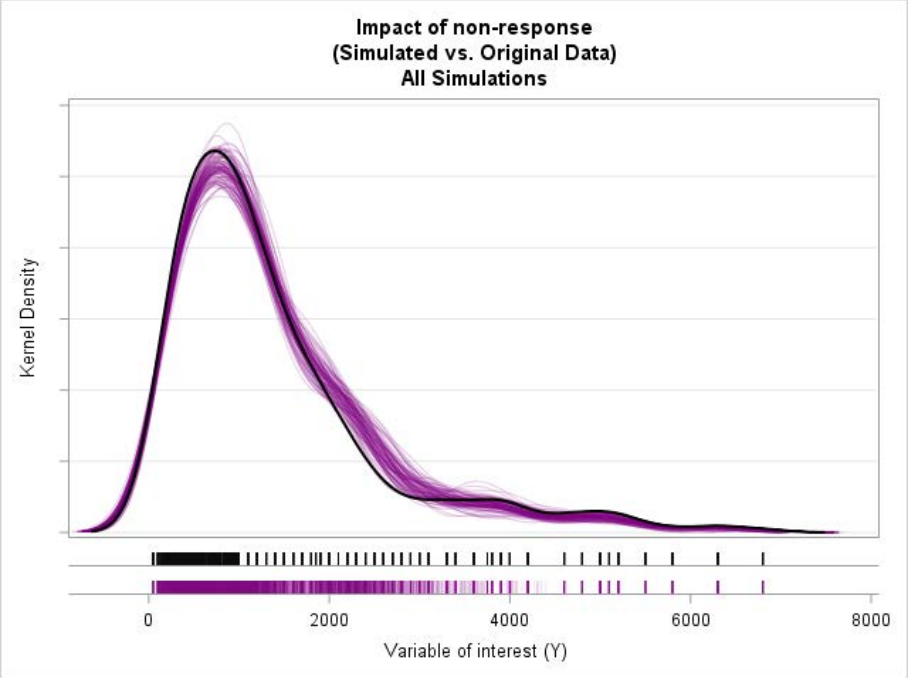
For the accompanying example we've run 100 simulations, generating non-response using Bernoulli sampling with unit non-response probabilities of 25%. Imputation was performed using the missForest R package on default settings.

Chart 1 includes the kernel density curve of the original data (in black) and additional curves for each simulation. Visual inspection shows that the general shape of the distribution is preserved, although there does seem to be some shifting of data towards the middle of the distribution. The chart also includes a fringe plot (also referred to as a rug plot) with individual data points plotted as semi-transparent vertical line segments. The fringe plot does not convey as much information about the overall distribution, but does highlight where the imputation method is introducing new values.

Additional analysis features in the tool include the option to plot histograms instead of density curves, and jitter-and-box plot combinations instead of fringe plots. Additionally, users can choose to overlay asymptotic information (distributions derived from the complete set of simulations) or inspect individual simulations. Finally, users can choose

to filter the data so as to focus only on the distribution of missing data, instead of the complete data set.

Chart 1: Univariate Distribution Analysis (Example)



4.2 Estimator Analysis

Let θ be an estimator derived from the original data set (as a function of the variable y and possibly auxiliary variables). Let θ_j be that same estimator derived from simulation j. Previous work in this field has focused on numerical measures such as relative bias (RB) and relative root mean square error (RRMSE) to assess imputation strategies:

$$RB(\hat{\theta}) = \frac{1}{m}\sum_{j=1}^{m}\frac{(\hat{\theta}_j - \theta)}{\theta}$$

$$RRMSE(\hat{\theta}) = \frac{\sqrt{\frac{1}{m}\sum_{j=1}^{m}(\hat{\theta}_j - \theta)^2}}{\theta}$$

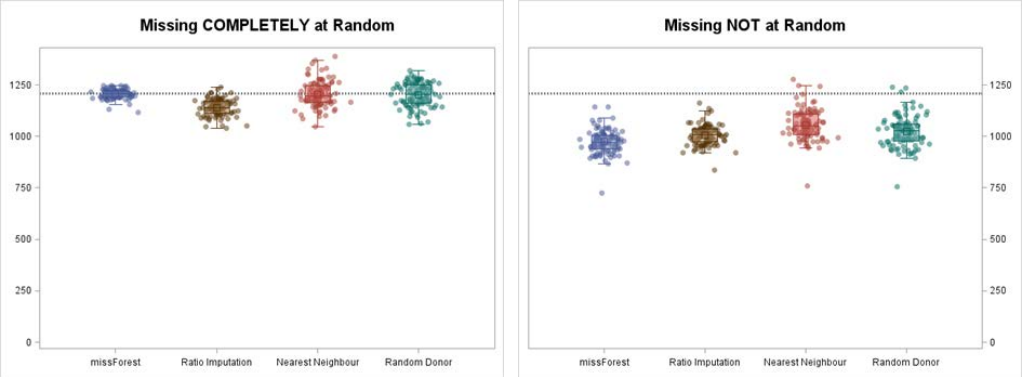
The underlying concepts of these measures can be suitably captured simply by plotting the set of individual simulation estimators $\{\theta_j\}$ against the true value θ.

The following example compares the effectiveness of four imputation methods across two non-response mechanisms: MCAR and MNAR. We again generated 100 simulations under each mechanism. We used SRSWOR and Pareto sampling so as to fix the overall non-response rate in each simulation at 25%. The results are given in Chart 2. This chart

includes a jitter plot of all simulations, in addition to summary boxplots. The true value θ of the estimator is plotted as a horizontal dotted line.

Under the MCAR simulation, it is clear that the missForest imputation method (blue) performs best – it is unbiased and produces the smallest variance across simulations. The two donor-based methods (red and green) perform similarly, while ratio imputation (brown) is more robust but introduces some bias. Under the MNAR simulation, missForest arguably performs the worst, although all four methods produce a bias in the resulting estimator. We refer to this type of test – comparing results over different hypothetical non-response mechanisms – as a *sensitivity* test. As many imputation methods rely on the assumption that data is either MCAR or MAR, this type of test highlights the importance of investigating potential outcomes when these assumptions fail.

Chart 2: Estimator Analysis (Example)



4.1 Predictive Analysis

Predictive measures assess the imputation effectiveness with respect to individual units. Ideally, an imputation method would perfectly reproduce missing values. Short of this goal, we would like to see small imputation errors, i.e., the difference between an imputed value and its original value. Numerical measures of predictive analysis include mean average error (MAE) and root mean square error (RMSE). We instead propose plotting the errors in a scatterplot. Plotting them against the original data $\{y_i\}$, the imputed data $\{y_{ij}\}$, or auxiliary data can provide insight into such issues as heteroskedascity and imputation model misspecification.

For this example, we used leave-one-out cross-validation (LOOCV), repeated ten times. LOOCV is a special case of cross-validation: in each simulation, exactly one record is set to missing and imputed. As there are 372 records in our test data, and each one is imputed exactly ten times, this generates 3720 simulations.

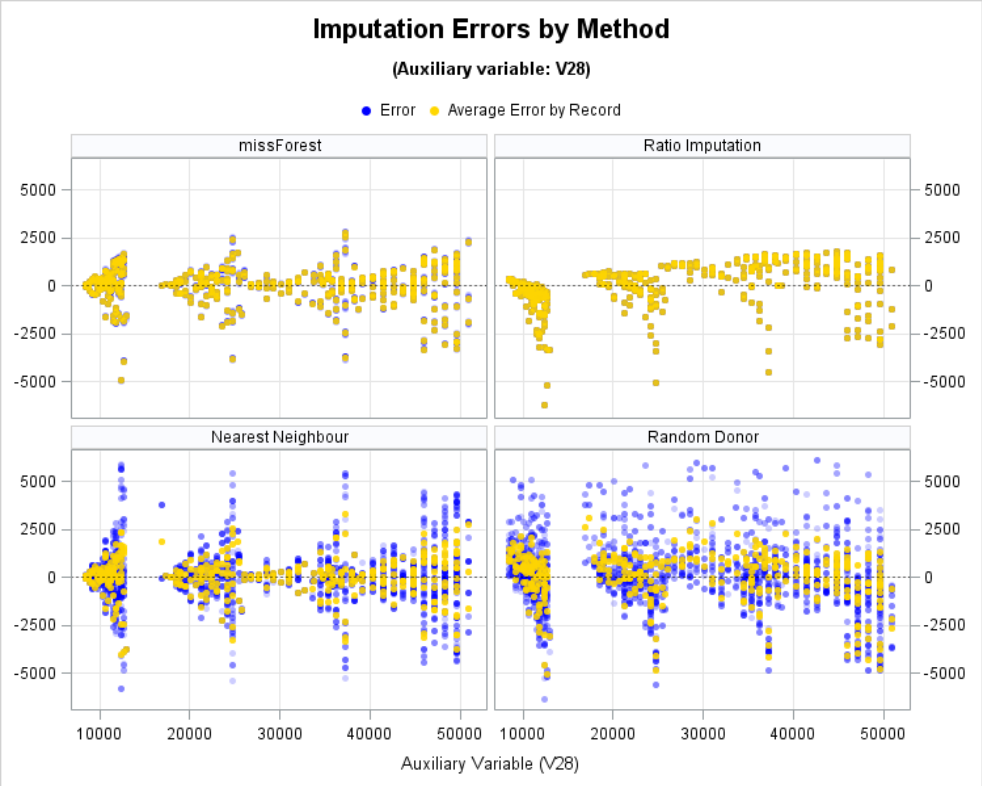
LOOCV is a good place to start when investigating the predictive performance of an imputation method because it tests the method's performance on each record under a best-case scenario: when all information (except for the record in question's value) is available. Additionally, because each unit is imputed the same number of times, there is no need to encode non-response rates associated with each record into the visual outputs.

(It might otherwise be difficult to visually distinguish between records with poor predictive imputation results and records with high non-response rates.)

Repeating each simulation multiple times is an additional feature offered by the tool, and is designed to help users estimate the variance due to imputation conditioned on a non-response pattern. Given a non-response pattern, some imputation methods will always produce the same imputed values. For example, a nearest-neighbour donor imputation will behave this way, as long as there are no distance ties between donors. Other imputation methods such as stochastic regression imputation will introduce a random term into the imputation process. Repeating the same non-response pattern multiple times is a way of examining this property.

We tested the same four imputation methods (missForest, ratio imputation, nearest-neighbour donor imputation, and random hot-deck donor imputation) as in the previous example. For the first three, we used auxiliary variable V28 as part of the imputation process.

Chart 3: Predictive Analysis (Example)



In Chart 3 we plot the imputations errors $\{e_{ij} = \hat{y}_{ij} - y_i | \delta_{ij} = 1\}$ against auxiliary variable V28 in the horizontal axis. As each of 372 records were imputed ten times, this represents 3720 data points; they are plotted as semitransparent blue dots. Overlaid on top of the errors, we plot the average error

$$\bar{e}_i = \frac{\sum_j e_{(i)j}}{\sum_j \delta_{(i)j}}$$

in yellow (372 data points). The average can be viewed as an approximation of the asymptotic error of the imputation method, albeit over only ten iterations.

By looking at the yellow scatterplot, we can examine the relationship between the average imputation errors and the auxiliary variable V28. As V28 was the auxiliary variable used for imputation purposes in three of the four methods (random donor does not use any auxiliary information), the average error can be viewed as an approximation of the imputation model residuals. In this case we are looking for a random distribution of the residuals around V28. While this appears to hold true for missForest and Nearest Neighbour methods, there is a clear pattern to the residuals generated by the Ratio Imputation method. Using this evidence, a user might choose to remove this methods from consideration, or add additional explanatory variables that could improve the imputation model.

Additionally, from the presence of blue data points, we can see that both donor methods (nearest neighbour and random hot-deck) introduce a degree of variance within the imputation method itself, while the other two methods do not.

5. Conclusion

In this paper we've presented a framework for assessing and comparing imputation methods on a known data set. We introduced some new methods for generating non-response and analysing results using data visualization, alongside examples. These examples are intended only to demonstrate the potential application of the assessment and comparison tool, not as a commentary on the imputation methods themselves.

This tool is designed for users to assess and compare one or more imputation methods in a controlled simulation environment. From an inferential standpoint, the accuracy of any conclusions depends on how well the simulation mimics the targeted, real-world process. This depends on two factors:

- How well does the training data – in this case the original data $\{y_i\}$ – represent the true target data?
- How well does the simulated non-response mechanism mimic the true mechanism?

In general, the true non-response mechanism is unknown. By investigating various non-response mechanisms, as demonstrated in our second example, users may at least test an imputation method's sensitivity to non-response mechanism type.

On the other hand, it is generally stated that for simulation scenarios, and other data-driven learning experiments, that conclusions are only as good as the training data. In other words, if the training data is not representative, it is difficult to make any resulting inferences. One issue for this particular type of experiment is that in many cases, we expect users to construct the training data from data that has already undergone non-response. If the pre-simulation non-response mechanism is MAR or MNAR, then the training data will not be representative. Nonetheless, much can still be learned about the behaviour of an imputation method under simulation, making it a worthwhile undertaking.

While the current version of the tool only assesses imputation performance on a univariate numerical variable, future work is planned for categorical and multivariate data as well.

Acknowledgements

The author would like to thank Keren Li, University of Ottawa, and Helen Fu, Simon Fraser University, for the work they did in support of this project.

References

Buuren, S. V., & Groothuis-Oudshoorn, K. (2010). mice: Multivariate imputation by chained equations in R. *Journal of statistical software*, 1-68.

Chambers, R. (2001). Evaluation criteria for statistical editing and imputation, national statistics methodological series no. 28. *University of Southampton*.

De Waal, T., Pannekoek, J., & Scholtus, S. (2011). *Handbook of statistical data editing and imputation* (Vol. 563). John Wiley & Sons.

Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Haziza, D. (2003). The Generalized Simulation System (GENESIS): A Pedagogical and Methodological Tool. *2003 Joint Statistical Meetings – Section on Survey Research Methods.*

Rafiei, M. H., & Adeli, H. (2015). A novel machine learning model for estimation of sale prices of real estate units. *Journal of Construction Engineering and Management, 142*(2), 04015066.

Rosén, B. (1997a). Asymptotic theory for order sampling. *Journal of Statistical Planning and Inference*, *62*(2), 135-158.

Rosén, B. (1997b). On sampling with probability proportional to size. *Journal of Statistical Planning and Inference*, *62*(2), 159-191.

Statistics Canada (2017). Functional Description of the Banff System for Edit and Imputation, *Technical Report*.

Stekhoven, D. J. (2015). missForest: Nonparametric missing value imputation using random forest. *Astrophysics Source Code Library*.

Stelmack, A. (2018). On the Development of a Generalized Framework to Evaluate and Improve Imputation Strategies at Statistics Canada, *United Nations Statistical Commission and Economic Commission for Europe – Workshop on Statistical Data Editing.*

United Nations Economic Commission for Europe (UNECE) (2019). Generic Statistical Data Editing Model (GSDEM). [https://statswiki.unece.org/display/sde/GSDEM].