BEA Working Paper Series, WP2022-3

Banff or Simputation? Assessing Alternative Approaches to the Imputation of Missing and Erroneous Data on BEA's Multinational Enterprise Surveys

Author Larkin Terrie, U.S. Bureau of Economic Analysis

Contact Larkin.Terrie@bea.gov

Date March 2022

Abstract The U.S. Bureau of Economic Analysis (BEA) employs automated data editing and imputation systems to process a subset of its multinational enterprise (MNE) surveys. Until recently, all of the auto-editing systems used at BEA were built around the Banff system for data editing and imputation produced by Statistics Canada, which runs in SAS. To enhance BEA's flexibility, BEA researchers have explored the feasibility of creating auto-editing systems built around software other than Banff, focusing in particular on a set of R packages created by researchers at Statistics Netherlands. Since previous research has established that Banff produces highly accurate imputations on BEA's MNE surveys, a key question is whether these R packages produce imputations that are as accurate as those produced by Banff. Among these R packages, the Simputation package is responsible for almost all of the imputation functionality. This project employs a simulation-based approach to assess the relative accuracy of the imputations produced by Banff and Simputation, using data collected by two different MNE survey instruments. The simulation results indicate that Simputation is sufficiently accurate to make the Statistics Netherlands R packages a viable alternative to Banff. However, the results differ by survey instrument, suggesting that Simputation might produce more accurate imputations when the instrument collects relatively few data items and that Banff might be more accurate for forms that are longer and more complex.

Keywords Survey editing, imputation, foreign direct investment

JEL codes C83, F23

Bureau of Economic Analysis
U.S. DEPARTMENT OF COMMERCE

The views expressed in this paper are those of the authors and do not necessarily represent the U.S. Bureau of Economic Analysis, or the U.S. Department of Commerce.

1. Introduction

Since 2016, the U.S. Bureau of Economic Analysis (BEA) has used automated systems to “edit” a portion of the forms received for its annual and benchmark surveys of multinational enterprises.[1] Survey editing is the practice of reviewing submitted survey forms for accuracy and completeness before they are incorporated into official statistics.[2] Editing at BEA (as well as at other statistical agencies in the United States and around the world) has traditionally been done by subject-matter experts who manually review individual submitted survey forms. As the number of respondents to BEA’s multinational enterprise (MNE) surveys has dramatically increased in recent years, the adoption of automated editing routines for a subset of survey forms has helped reduce the burden on expert editors and allowed them to focus their efforts on forms that have the largest impact on published statistics.

Until recently, all of BEA’s auto-editing systems were built around the Banff system for data editing and imputation developed by Statistics Canada. Banff runs as an add-on to SAS statistical software and consists of nine independent SAS procedures that can be used to identify and correct erroneous items in survey data.[3] BEA has successfully used auto-editing routines built around six of the nine Banff procedures to process a portion of the forms submitted for a variety of its MNE surveys, including: the 2014 and 2019 Benchmark Surveys of U.S. Direct Investment Abroad; the 2017 Benchmark Survey of Foreign Direct Investment in the United States; and the 2016, 2018, and 2019 Annual Surveys of Foreign Direct Investment in the United States. Moreover, research conducted at BEA has shown that auto-editing survey forms with Banff does not reduce the accuracy of published statistics, as the results of auto-editing and traditional manual editing tend to be highly similar (Xu, Kim, and Terrie 2017).

[1] These surveys are used to produce widely used statistics on U.S. direct investment abroad and foreign direct investment in the United States. As defined in these surveys, a direct investment relationship (and thus a multinational enterprise) exists when an investor resident in one country owns 10 percent or more of the voting interest of an incorporated business enterprise, or an equivalent interest in an unincorporated business enterprise, resident in another country. These surveys include both annual and benchmark surveys. The benchmark surveys are conducted every five years in place of the annual surveys, and, for the benchmark surveys, all companies that meet the reporting requirements are required to report, whereas for annual surveys, only companies notified by BEA and that meet certain thresholds are required to report.

[2] An excellent and comprehensive overview of issues related to statistical data editing can be found in de Waal et al. (2011). See also Fellegi and Holt (1976).

[3] A detailed explanation of these nine methods can be found in Banff Support Team (2017). Insightful discussions of the development of Banff and its predecessor, the Generalized Edit and Imputation System, can be found in Giles and Patrick (1986), Kovar et al. (1988), Whitridge and Kovar (1990), Kozak (2005), Mohl (2007), and Gray (2018). Applications of Banff outside of BEA are presented in Barboza and Turner (2011), Johanson (2013), and Seyb et al. (2009).

As auto-editing has become an increasingly important part of BEA's survey processing, BEA researchers have explored the feasibility of creating auto-editing systems built around software other than Banff. They have focused in particular on a set of R packages created by researchers at Statistics Netherlands.[4] These freely available packages were created with the production of official statistics in mind and provide much of the same functionality as Banff.[5] As such, these R packages provide BEA with a potential foundation for building a complete set of auto-editing routines that accomplish everything that the routines built around Banff accomplish.

An important difference, however, between these R packages and Banff is in the procedures they provide for imputing new values for missing or erroneous survey items. Among the R packages from Statistics Netherlands, almost all of the imputation functionality is provided by the Simputation package.[6] While Simputation does offer some of the same imputation methods as Banff, there are methods available in Banff that are not available in Simputation and vice versa. For example, Banff and Simputation both provide donor imputation procedures—wherein a form with invalid or missing data receives data from a form identified as being similar to it—but they offer different algorithms for determining donor-recipient matches. In addition, the model-based imputation methods offered by each are different: Banff provides a linear regression-based method and Simputation provides a variety of methods based on both regression and classification and regression trees.[7]

Given these differences, a key question for BEA in regard to the desirability of building auto-editing routines around the R packages from Statistics Netherlands is how accurately Simputation imputes data for missing or erroneous survey items compared to Banff. This paper addresses this question by attempting to measure the relative accuracy of the imputations produced by Banff and Simputation.[8] To be sure, measuring the accuracy of an imputation (or of many imputations) can be difficult, since the true value for which an imputation acts as substitute—and against which its accuracy should be

[4] Mark Van Der Loo and Edwin de Jonge have played especially important roles in developing these R packages.

[5] These packages include Deductive, DcModify, EditRules, ExtremeValues, ErrorLocate, Lumberjack, RSPA, Simputation, Validate, ValidateDb, and ValidateTools. All of these packages can be downloaded for free from the Comprehensive R Archive Network (CRAN), https://cran.r-project.org/.

[6] The exception is the deductive imputation method, the equivalent of deterministic imputation in Banff, which is provided by the Deductive package. Since the deductive and deterministic methods are identical, they are not a source of variation in the imputations produced by Banff and Simputation and are not addressed in this paper's analysis. However, see section II for a brief explanation of how they work.

[7] For information on all of the functionality provided by Simputation, see https://CRAN.R-project.org/package=simputation.

[8] This paper thus belongs to a broader literature that assesses the application of automated editing and imputation procedures to surveys conducted by National Statistical Institutes. For example, see Bianchi et al. (2020), Dorinski (1998), Lange (2020), Salvucci et al. (2012), Scholtus and Daalmans (2020), and Scholtus et al. (2017).

measured—is generally not known. To surmount this problem, this paper adopts a simulation-based approach. In a nutshell, the approach is to simulate the presence of missing and erroneous data on survey forms without missing or invalid data, generate separate sets of imputations with Banff and Simputation for the items simulated as being missing/erroneous, and compare the results by measuring which set of imputed values is closer to the original responses.[9]

The analysis assesses the relative accuracy of imputations produced by Banff and Simputation for two different BEA MNE survey forms: the BE-10D and the BE-15B. Four different auto-editing systems are thus analyzed: a Banff-based system for the BE-10D, a separate Banff-based system for the BE-15B, as well as Simputation-based systems for the 10D and 15B. The 10D and 15B were chosen for this analysis because (1) they are among the forms for which BEA has developed Banff-based auto-editing systems and (2) they differ significantly in terms of their complexity and in the number of data items collected, making them representative of the variety of form types auto-edited at BEA.

The BE-15B is the more complex of the two forms and is collected as part of BEA's Annual Survey of Foreign Direct Investment in the United States (the BE-15). This survey collects financial and operating data on foreign-owned U.S. business enterprises, or "U.S. affiliates." The affiliates are responsible for completing the survey, and they submit a BE-15A, BE-15B, or BE-15C form depending on their size—operationalized as the maximum of the absolute values of assets, net income, and sales—and whether they are majority foreign owned. As indicated in table 1, the 15B is the form for mid-sized affiliates that are majority foreign owned as well as for mid- and large-sized affiliates that are minority foreign owned.[10] In terms of its complexity and length, it is longer and more complex than the 15C but shorter and less complex than the 15A.

Table 1. Criteria for Filing Each of the BE-15 Survey Forms

Size	Level of Foreign Ownership	Form
Greater than $40 million but no greater than $120 million	Minority or Majority	C
Greater than $120 million but no greater than $300 million	Minority or Majority	B
Greater than $300 million	Minority	B
	Majority	A

[9] In using a simulation-based approach, this paper follows a tradition in the statistical literature of using simulations to assess the quality of imputations. See, for example, Beaumont and Bocci (2007), Di Zio et al. (2006), Dorinski et al. (1996), and Gray (2020).

[10] In other words, The BE-15B collects information on majority foreign-owned U.S. affiliates with assets, net income, or sales of over $120 million (positive or negative) but for which none of these is greater than $300 million and on minority foreign-owned affiliates with assets, net income, or sales of over $120 million (positive or negative).

The BE-10D is collected as part of BEA's Benchmark Survey of U.S. Direct Investment Abroad (the BE-10). This survey, conducted every 5 years in place of the Annual Survey of U.S. Direct Investment Abroad, collects financial and operating data on U.S. multinational enterprises, which includes data on the U.S. parent and its foreign affiliates. The BE-10D is the simplest of the forms that make up the BE-10. It collects data on the smallest category of foreign affiliates, those whose assets, net income, and sales are all less than $25 million (positive or negative). Due to the small size of these affiliates, their data are only collected every 5 years as part of this benchmark survey.

For each of these forms, the analysis is divided into two parts. In the first part, the simulation framework is used to compare all of the imputation methods available in Simputation against one another to determine which tend to produce the most accurate imputations for each of the two forms. Based on the results of this part of the analysis, the optimal imputation procedures for the 10D and 15B, respectively, are identified and used to construct complete Simputation-based auto-editing systems for the two forms. The second part of the analysis again uses the simulation framework, this time to compare the complete Simputation-based auto-editing systems against the Banff-based systems previously developed. To avoid bias that could arise from using the same data to select imputation procedures (part one of the analysis) and test these procedures (part two), different datasets are used in each part of the analysis for each form. For the 10D, the most recent two years of survey data, 2014 and 2019, are used: 2019 for the selection of imputation procedures and 2014 to compare the complete Simputation and Banff auto-editing systems. For the 15B, the most recent two survey years before the adoption of auto-editing, 2014 and 2015, are used: 2015 for selecting imputation procedures from the Simputation package and 2014 for the Simputation-Banff comparison.[11]

The rest of the paper proceeds as follows. Section 2 provides an overview of the imputation methods available in Banff and Simputation. Section 3 explains the simulation framework that is used for the analysis in subsequent sections and the statistics that will be used to measure the accuracy of imputations. Section 4 presents the results of comparing the different imputation methods available in Simputation against one another for both the 10D and 15B and identifies the optimal methods for each form. Section 5 presents the results of comparing the complete Simputation and Banff-based auto-editing systems against one another for both the 10D and 15B. Section 6 concludes.

[11] In the 2016, 2018, and 2019 survey years, a portion of BE-15B forms were auto-edited and the rest were edited manually.

2. Overview of Imputation Methods in Banff and Simputation

Of the six Banff procedures used by BEA, three are imputation procedures: Proc Deterministic, Proc DonorImpute, and Proc Estimator.[12] While Proc Estimator is based entirely on statistical modeling, Proc Deterministic and Proc DonorImpute rely, to different degrees, on validity checks (or "edits") to make imputations. Edits are logical and mathematical rules that define the range of allowable values for survey items as well as the relationships between different items. Those in use at BEA were developed for manual survey editing and adapted to auto-editing. For example, there is an edit that requires employment to be non-negative and another that requires the ratio of total employee compensation to the number of employees to be within a certain range. Each survey form has its own set of edits, though some edits are shared by multiple forms. At an earlier stage of auto-editing, these rules are used by the error localization procedure to identify data items that need to be replaced by imputations (FTIs or "fields to impute" as referred to within Banff), but they also play a role in the imputations themselves.

In brief, the three Banff imputation procedures used by BEA work as follows:

- **Proc Deterministic** produces imputations for FTIs for which there is one and only one value that satisfies the edits. For example, an edit for the BE-15B specifies that assets must equal liabilities plus owners' equity. If a form reports a value of 30 for liabilities and 70 for owners' equity and requires an imputation for assets (i.e., assets is an FTI), deterministic imputation will produce a value of 100 for assets.
- With **Proc DonorImpute**, each record having one or more FTIs is matched to a donor record with valid data, and the FTIs in the recipient record are populated with the corresponding data from the donor. The edits, specifically those that specify interrelationships among different fields, are used by Proc DonorImpute to determine which fields to use in matching donor records to recipient records. For example, since, as mentioned above, there is an edit specifying a relationship between employment and employee compensation, a record with an FTI for employment would likely be matched to a donor that has a similar value for employee compensation.
- **Proc Estimator** performs imputation based on linear regression models or predetermined mathematical formulae that are specified by the user.

[12] There is a fourth imputation procedure available in Banff, MassImputation. It is not used by BEA because it is a specialized procedure for two-phase sampling that is not applicable to BEA's MNE surveys. See Banff Support Team (2017).

Of these three imputation procedures, only Proc DonorImpute and Proc Estimator are examined in detail in this paper. There are two main reasons for not focusing on deterministic imputation. First, in general, only a relatively small proportion of survey items requiring an imputation can be imputed using deterministic imputation.[13] Second, an identical procedure is available in one of the R packages from Statistics Netherlands, the Deductive package, which means that the performance of deterministic (or deductive) imputation is not relevant to a comparison of the accuracy of the imputations produced by Banff and the Statistics Netherlands R packages.

An important feature of Banff's donor imputation procedure that makes it different from all of the donor imputation methods available in Simputation is its use of the edits to identify matching fields. A key consequence of this feature is that, unlike the donor imputation methods in Simputation, Proc DonorImpute does not require the user to tell it which fields to use in matching donors and recipients.[14] The advantage of this feature is that it takes advantage of the expert knowledge already codified in the edits regarding the relationships among fields, thereby freeing the user, who may not be a subject-matter expert, from identifying appropriate matching fields. On the other hand, a potential disadvantage of this approach to matching, at least for BEA, stems from the fact that Banff does not accept if-then edits. Many of BEA's edits take the form of if-then statements (e.g., if employment is zero, then total employee compensation must also be zero), and using these edits with Proc DonorImpute requires partitioning the data into subsets, each of which contains records that satisfy the same set of if conditions, and then processing each subset separately (see Xu, Kim, and Terrie 2017). This solution makes it possible to use Banff's donor imputation procedure but with the drawback that the pool of potential donors for each recipient is limited to forms that satisfy the same set of if conditions.

The two forms under analysis in this paper, the BE-10D and BE-15B, differ significantly in terms of the number of edits that pertain to each. As a more complex form with a larger number of fields subject to auto-editing, the BE-15B has many more edits that specify interrelationships among its fields than does the relatively simple BE-10D. Since Banff incorporates these edits into its algorithm for matching donors and recipients for donor imputation, a question for the analysis in section 5 will be whether this difference provides Banff with an advantage in producing accurate imputations for the 15B but not the 10D.

[13] See, for example, Giles and Patrick (1986, 53) and Kozak (2005, 5).

[14] While Banff does not require user input to determine matching fields, it does allow the user to specify additional matching fields with the MUSTMATCH option.

To optimize the imputations produced by its BE-10D and BE-15B auto-editing systems, BEA has fine-tuned, or customized, its use of Proc DonorImpute and Proc Estimator in a number of ways. For example, when running Proc DonorImpute, donor and recipient forms are only allowed to match if they have reported the same four-digit NAICS (North American Industry Classification System) code. For estimator imputation, BEA researchers have developed an approach that involves using stepwise regression and other statistical techniques to ensure the selection of predictor variables that will generate high quality imputations.[15] Finally, although some research suggests donor imputation should be run before estimator imputation—to ensure donor imputation has the opportunity to impute all FTIs before any have been replaced with values generated by estimator imputation—BEA has found that in certain circumstances the quality of imputations can be improved by running estimator imputation before donor imputation.[16] That is, altering the order of estimator and donor imputation has played a role in the optimization of BEA's auto-editing systems.

Compared to Banff, Simputation offers a wider array of imputation methods. Table 2 lists 11 key imputation methods available in Simputation that were examined by BEA. Given these extensive options, the issue of which of these methods to use and in what order to run them (as well as how to fine-tune the methods chosen) assumes crucial importance in designing auto-editing systems based around Simputation. This issue will be addressed in detail in section 4 but for now it bears emphasizing that, in designing auto-editing systems for the BE-10D and 15B, an initial selection phase of analysis will be necessary in order to identify the methods most likely to produce accurate imputations for each form. Only after Simputation-based auto-editing systems have been created based on this analysis will it be possible to compare the accuracy of the imputations produced by Simputation and Banff.

[15] To balance the goals of ensuring imputations are as accurate (on average) as possible and generating imputations for as many forms as possible, BEA uses an approach that involves multiple runs of estimator imputation with successively simpler models used in each run. That is, the first run involves making imputations with a model based on stepwise entry and retention parameters that are relatively permissive, allowing a large number of predictor variables into the model. Each successive model is then based on entry and retention parameters that are more restrictive than the preceding model. In this way, the model most likely to create high quality imputations – but also least likely to generate a large number of imputations since it requires a form have valid data for a large number of fields – is run first. Each subsequent model, though somewhat less likely to produce accurate imputations, is also more likely to produce a large number of imputations since it requires less valid data to make an imputation on a given form. This same approach is used with the regression-based imputation methods available from Simputation in sections 4 and 5 below.

[16] For example, Kovar et al. (1988, 629) note that donor imputation is generally to be preferred over estimator imputation because all imputations for a given record are taken from the same donor record, thereby preserving the relationships among the imputed fields. Preserving the relationships between variables in the micro data is not, however, of primary importance to BEA since it only publishes aggregate-level data and does not publish variances or covariances. Also, see Terrie (2018) for a discussion of evidence that imputation quality can sometimes be improved by running estimator imputation before donor imputation.

Table 2. Key Imputation Functions in the Simputation Package

Function	Description	Type
impute_shd	Sequential hot deck imputation	Donor
impute_rhd	Random hot deck imputation	
impute_knn	*K*-nearest neighbor imputation	
impute_pmm	Predictive mean matching	
impute_lm	Linear regression model-based imputation	Regression
impute_rlm	Robust linear regression through *M*-estimation	
impute_en	Elastic net/lasso/ridge regression imputation	
impute_cart	Classification and regression tree imputation	Decision Tree
impute_rf	Random forest imputation	
impute_mf	Multivariate imputation based on iterative random forest estimates	
impute_em	Multivariate imputation based on Expectation Maximization-estimation of multivariate normal parameters	EM estimation

As indicated in table 2, there are four main categories of imputation methods provided by Simputation: donor, regression, decision tree, and an EM estimation method.[17] In regard to the donor-based methods, they vary significantly in the sophistication of the techniques they use to pair donors with recipients, but they have in common that they rely on the user to specify which fields should play a role in determining matches.[18] Of the three regression-based methods, impute_lm stands out in that it offers the same functionality as Banff's Proc Estimator (except for Proc Estimator's ability to make imputations based on any user-specified mathematical formula). The other two Simputation functions in this family, impute_rlm and impute_en, allow for imputation using regression-based methods that are not available in Banff. The function impute_rlm offers robust linear regression through *M*-estimators—a class of estimator explicitly designed to be robust to violations of assumptions about the probability distribution from which the data are drawn (Huber 1981). The function impute_en performs imputation using the "elastic net" method, which is a modification of OLS regression that incorporates a penalization technique designed to improve the quality of predictions generated by the resulting model (Zou and Hastie 2005).

[17] As alluded to above, the methods provided by Banff only cover two of these categories: regression (Proc Estimator) and donor (Proc DonorImpute).

[18] For a detailed discussion of each of these donor-based methods, see Scholtus (2014).

Unlike Banff, Simputation also provides model-based imputation methods that do not involve linear regression. Three of these functions are based on decision trees: impute_cart, impute_rf, and impute_mf. The first of these involves constructing a single classification and regression tree, the second is based on random forests, and the third is a multivariate imputation technique that involves iterative random forests (Stekhoven and Buehlmann 2012). Finally, Simputation also offers impute_em, which is a multivariate imputation technique involving Expectation Maximization (Dempster, Laird, and Rubin 1977).

3. Methodology

The analyses in sections 4 and 5 use a simulation-based framework, developed in Terrie (2018), to assess the accuracy of imputations on the BE-10D and 15B survey forms. In section 4, the framework is used to evaluate the different imputation methods available in Simputation in order to identify the methods most likely to produce accurate imputations for the 10D and 15B, respectively. Once identified these methods are combined into complete auto-editing systems for imputing missing and erroneous data for the 10D and 15B, which are then, in section 5, compared to the Banff-based systems previously developed for these forms. As explained above, four datasets are used in the analysis: 2019 BE-10D and 2015 BE-15B data in section 4 and 2014 BE-10D and 2014 BE-15B data in section 5. By using different datasets for the selection (section 4) of imputation methods and the testing (section 5) of these methods for each form, the analysis seeks to avoid the possibility of results being biased due to overfitting.

The simulation framework involves selecting non-missing, non-erroneous survey responses to be treated as if they were missing or erroneous—i.e., as if they were fields to impute (FTIs). A precondition for conducting tests with this framework is the creation of datasets from which all records (i.e., individual submitted forms) with missing or erroneous data have been excluded. Error localization was run on each of the four datasets involved in the analysis, and any record identified as having missing or erroneous data was dropped. Then, for each simulation run conducted on each of these “clean” datasets, a new set of responses is selected to be treated as FTIs and replaced by imputations generated by Simputation or Banff. By conducting a large number of these simulation runs and comparing the differences between the resulting imputed values and the corresponding reported values, conclusions can be drawn about the average quality of the imputations for each of the survey items, or fields, subject to imputation.

An important feature of this framework is that the fields designated as simulated FTIs in each run of the simulation are not chosen at random. They are chosen based on an algorithm that is designed to mimic the actual distribution of missing and erroneous data on reported 10D and 15B forms.[19] This approach was adopted because (1) certain patterns of missing/erroneous data are more common than others and (2) it is more difficult to accurately impute data for certain patterns of missingness/erroneousness than others. For example, it is relatively common for employment and total employee compensation both to have invalid or missing data for a given record. Employment (i.e., number of employees) and employee compensation also tend to be highly correlated with one another, meaning that it is easier to impute one of them if valid data has been reported for the other. The results of the simulation will thus be more realistic and useful if the frequency with which employment and employee compensation are simultaneously simulated as FTIs mirrors the actual frequency with which they are simultaneously FTIs, and the same argument applies for many other patterns of missing/erroneous data.

The paper's results are presented in terms of two measures of the distance between the values imputed for simulated FTIs and their original (i.e., reported) values: percent total error and percent total absolute error. The percent total error (percent error for short) for a given field, or survey item, measures the total difference between imputed values and reported values over all runs of the simulation as a proportion of the reported values' total. Its mathematical representation for field *i* (assets, sales, etc.) is the following:

$$\frac{\sum_{k=1}^{m}\sum_{j=1}^{n} s_{ijk} - o_{ij}}{\sum_{k=1}^{m}\sum_{j=1}^{n} o_{ij}} \times 100$$

where s_{ijk} is the imputed value for field *i* in record *j* = 1, …, n in simulation run *k* = 1, …, *m* and o_{ij} is the corresponding reported, or original, value for the field and record in question.[20] In contrast, the percent total absolute error (or percent absolute error) for a given field measures the total absolute value of the

[19] This algorithm involves using a series of logistic regression models to assign a probability, *p*, of being an FTI to each of the items on every individual 10D and 15B form in the "clean" dataset. In each iteration of the simulation, whether each field will be a simulated FTI is determined by performing a random draw from a Bernoulli based on *p*. In order to capture the likelihood of different combinations of data items simultaneously being FTIs on a given form, the probabilities assigned to a given form in a given iteration of the simulation are not independent of one another. The dependence of these probabilities is ensured by assigning the probabilities to the data items on a given form in a sequential manner such that the probabilities assigned later in the sequence depend on those assigned earlier.

[20] A form is only included in the calculation for a given field and simulation run if it is one of the forms on which the field is a simulated FTI.

differences between imputed values and reported values over all runs of the simulation as a proportion of the total absolute value of the reported values. For field *i*, it can be represented as:

$$\frac{\sum_{k=1}^{m}\sum_{j=1}^{n}\left|s_{ijk}-o_{ij}\right|}{\sum_{k=1}^{m}\sum_{j=1}^{n}\left|o_{ij}\right|}\times 100$$

The key difference between these two measures lies in whether positive and negative differences between imputations and reported values (i.e., over and underestimates) are allowed to cancel one another out. Since, in the calculation of percent error, positives and negatives cancel out, percent error is essentially a measure of the proximity of the aggregate of the imputed values to the aggregate of the corresponding reported values. On the other hand, since positive and negative differences do not cancel out in the calculation of percent absolute error, it can be seen as measuring how close each individual imputation is, on average, to its corresponding reported value. Both sets of results are of interest to BEA—the non-absolute value differences because BEA publishes aggregate values based on its survey results and the absolute value differences because it publishes subtotals for certain industries and countries.

The accuracy of each imputation method in section 4 and of each auto-editing system in section 5 is measured based on 100 simulation runs (i.e., *m* = 100). In other words, in section 4 each of Simputation's 11 imputation methods is assessed using 100 simulation runs with the 2015 BE-15B data and a separate 100 simulation runs with the 2019 BE-10D data, where each run for each form is characterized by a distinct set of simulated FTIs. Likewise, in section 5 the complete Banff and Simputation-based auto-editing systems are each tested based on 100 simulation runs with the 2014 BE-15B data and 100 simulation runs with the 2014 BE-10D data, each run with each form having its own distinct set of simulated FTIs. Finally, to ensure the comparability of results between imputation methods in section 4 and between auto-editing systems in section 5, a common set of simulated FTIs is used for each form in each section.[21]

[21] For example, for the simulation runs conducted with the 2015 BE-15B data in section 4, the first run for each of the 11 methods consists of the exact same set of simulated FTIs; the second run then consists of another (distinct from those used in the first run) set of simulated FTIs that are also the same across the 11 imputation methods; and so on for all 100 runs. Similarly, for the simulation runs using the 2019 BE-10D data in section 4, each imputation method is confronted with an identical set of simulated FTIs in each of its respective simulation runs from 1 to 100. In section 5, an analogous approach is used. For the 100 simulation runs using the 2014 BE-10D data, the Banff and Simputation-based auto-editing systems had to make imputations for the same sets of simulated FTIs in each of the 100 runs, and the same thing is true of the 100 runs conducted with the 2014 BE-15B data on the 15B auto-editing systems.

In assessing the simulation results, it is important to bear in mind that any given imputation method will not, in general, be able to generate imputations for all of the FTIs—whether simulated or not—in a dataset. Imputations are essentially extrapolations based on the valid data on a submitted form, and it is frequently the case that there is not enough valid data in other fields for the method to make an extrapolation for a given FTI.[22] In addition, imputation methods vary in terms of how many and which survey items must have valid data in order to generate an imputation for a particular FTI. As a result, in section 4 of the present study, although each imputation method is, for each survey form, confronted with the same set of simulated FTIs, each method will generally generate imputations for a distinctive combination of these FTIs. A degree of caution is thus required when comparing the results for any given set of methods, since only a subset of the imputations made by each method will be for the same simulated FTIs (i.e., be for the same fields in the same records in the same simulation runs).

This paper responds to this complication by comparing simulation results in two distinct (and complementary) ways. In one set of tables, methods are compared on a pairwise basis, meaning that the percent error and percent absolute error presented for each field are based only on the imputations that the two methods being compared have in common. In another set of tables, these two measures are presented for three or more methods based on all imputations made by each method (i.e., imputations do not have to be shared by all methods in the table to be included). The value of the pairwise comparison tables is that they provide an essentially “pure” comparison of the relative accuracy of the two methods since imputations are included only if both methods have made an imputation for the FTI (and simulation run) in question. The one-to-one correspondence of the imputed items on which these tables are based also makes possible the use of pairwise *t* tests—which are more powerful than unpaired *t* tests—to investigate whether any differences between the two methods’ values for percent error or percent absolute error are statistically significant.[23] The drawback of the pairwise comparison tables is that their creation involves discarding data—though the extent of discarded data depends on the degree to which the two methods’ imputations are for the same simulated FTIs. A more complete picture of an individual method’s accuracy—though not necessarily of its relative accuracy compared to other methods—can be obtained by including all of its imputations in the calculation of the error statistics, which is the rationale for including tables in which the error statistics are based on all imputations made by each method.

[22] This situation generally arises when a form has multiple FTIs.

[23] The units of analysis for the pairwise t tests for percent error and percent absolute error are, by necessity, different. For percent absolute error, the test is based on a pairwise comparison of all individual imputations for a given field across all simulation runs. For percent error, the comparison is at the level of the simulation run rather than that of the individual imputed field. In a nutshell, since percent error is a measure of the degree of error present in the aggregate of imputed values, it is only meaningful to conduct the test at a level of analysis for which an aggregate can be calculated.

4. Creating Simputation-Based Auto-Editing Systems

Given the variety of imputation methods provided by Simputation, there is a wide array of options for setting up a Simputation-based auto-editing system. To make this task more manageable, BEA focused on creating auto-editing systems for the 10D and 15B that follow the Banff model of using one model-based imputation method and one donor imputation method. As explained above, for each of the 11 Simputation methods tested, 100 simulation runs were conducted, for each form, in which only the method under analysis was used to produce imputations (2 forms × 11 methods × 100 runs = 2,200 total simulation runs). The data used for the 10D are from survey year 2019, and the data used for the 15B are from 2015. Based on the results of this analysis, the best method of each type (donor and model-based) will be selected for each survey form.

Before presenting the results of this analysis, it is worth noting that, while the results are not presented here (though see Terrie 2018), a similar approach was used to optimize the Banff-based auto-editing systems currently in use at BEA. As Banff only provides one model-based and one donor imputation method, choosing among the available methods does not present the same difficulty as with Simputation. However, as discussed above, Banff's imputation methods can still be customized in a variety of ways, such as using industry codes to restrict the allowable donor-recipient matches and using stepwise selection in the identification of predictor variables for Proc Estimator, and simulation studies played a role in determining which of these customizations would be most helpful.[24]

Tables 2 through 5 provide the results for Simputation's donor imputation methods for the 10D and 15B forms. All of the auto-edited fields on each survey are listed in the tables. There are 7 auto-edited fields on the relatively simple 10D form and 17 auto-edited fields on the more complex 15B. These forms do collect more than 7 and 17 data items, respectively, but the other fields are not auto-edited for a variety of reasons. For example, many fields take categorical rather than numeric values, and BEA does not use pre-packaged statistical imputation procedures on categorical variables, as these procedures have generally been designed for use with continuous, numeric variables.[25]

[24] The imputation methods available in Simputation can also be customized in many ways. To ensure the comparability of results generated by different methods, similar customizations were used, where possible, for similar methods. For example, the same selection algorithm was used to select matching variables for all of the donor imputation methods, and, for all of the donor methods, donors and recipients were required to have the same four-digit NAICS code. Likewise, the decision tree-based methods were all customized in the same way, as were the regression-based methods, to the extent possible.

[25] Key categorical variables on the 10D include industry and country of the affiliate and on the 15B industry of the affiliate and country and industry of the foreign parent and ultimate beneficial owner. To the extent that it is necessary to identify errors in or perform recoding on these fields, these issues are dealt with in a stand-alone "pre-editing" program designed by BEA specifically for that purpose.

Table 3 presents the complete results for all of the donor imputation methods for the 15B, and table 4 provides a pairwise comparison of the two best performing of these methods, the *k*-nearest neighbor (impute_knn) and predictive mean matching (impute_pmm) methods. Overall, the results in these two tables tend to favor impute_knn as being the best donor imputation method for use with the BE-15B. The sequential and random hot deck methods (impute_shd and imputed_rhd, respectively) are disqualified for selection as the best donor imputation method for the 15B by their high percent absolute errors. While their percent errors are generally similar to those for impute_knn and impute_pmm, their percent absolute errors are higher than those obtained by the other two methods for all but two of the fields (capital gains and manufacturing employment)—and in most cases they are considerably higher. In regard to impute_pmm, its main weakness is that it imputed a much smaller proportion of FTIs than did the other methods (see the last row of table 3)—and was able to make no imputations whatsoever for assets, net income, and owners' equity. To be sure, impute_pmm does tend to produce somewhat more accurate imputations than does impute_knn (see the pairwise comparisons in table 4), but the small proportion of FTIs imputed is a major liability for this method. An accurate method is of little use if it imputes so little data that it prevents BEA from auto-editing all of the forms that require auto-editing.

Table 3. Comparison of Donor Imputation Methods for the BE-15B

Field	**Pct. Abs. Error**			
	impute_knn	**impute_shd**	**impute_rhd**	**impute_pmm**
Assets	32.10	163.50	116.39	
Cap. Gains	119.88	306.03	205.65	343.70
Employment	44.24	83.70	138.17	19.13
Emp. Comp.	43.87	84.32	246.71	31.02
Mfg. Emp.	130.34	38.41	73.95	55.99
PP&E Exp.	131.62	247.80	202.78	51.96
Exports	85.07	102.02	134.07	40.42
Gross PP&E	45.57	197.93	151.99	21.89
Imports	80.56	101.88	116.13	65.04
Interest Paid	84.17	123.00	147.09	34.07
Interest Rec.	85.58	101.11	133.01	80.54
Liabilities	49.36	139.91	324.32	1.97
Net Income	93.97	154.12	154.32	
Own. Equity	48.48	149.81	162.84	
R&D	49.02	135.95	189.38	66.16
Sales	44.80	57.04	126.34	18.55
U.S. Inc. Tax	100.83	116.88	151.81	111.10

Field	**Pct. Error**			
	impute_knn	**impute_shd**	**impute_rhd**	**impute_pmm**
Assets	-6.24	124.31	-32.31	
Cap. Gains	-178.78	-177.23	-355.78	203.94
Employment	-13.54	30.77	13.42	-5.76
Emp. Comp.	-18.44	32.89	112.37	-0.43
Mfg. Emp.	76.67	-7.10	-7.49	25.66
PP&E Exp.	48.86	181.89	44.40	-0.15
Exports	8.98	47.09	-18.18	12.79
Gross PP&E	-14.47	170.15	17.53	1.18
Imports	2.73	54.57	-16.22	26.83
Interest Paid	-11.71	36.30	10.10	-8.85
Interest Rec.	16.43	54.00	-3.58	37.72
Liabilities	-24.95	92.92	184.52	0.28
Net Income	67.35	22.74	105.62	
Own. Equity	-23.65	87.97	-1.42	
R&D	-21.79	89.05	45.00	4.40
Sales	-24.45	33.26	-2.09	5.97
U.S. Inc. Tax	-3.56	3.53	8.75	-4.26
Pct. Imputed	**94.06**	**77.19**	**77.19**	**8.16**

Table 4. Pairwise Comparison of impute_knn and impute_pmm for the BE-15B[26]

Field	Pct. Abs. Error		Pct. Error	
	impute_knn	impute_pmm	impute_knn	impute_pmm
Cap. Gains	198.69	343.70	199.52	203.94
Employment	24.84	19.13^{+}	-4.52	-5.76
Emp. Comp.	27.17	31.02	-8.09	-0.43
Mfg. Emp.	44.88***	55.99	-2.51***	25.66
PP&E Expend.	59.20	51.96^{+}	-13.83	-0.15^{+}
Exports	62.67	40.42**	-11.34	12.79
Gross PP&E	50.11	21.89***	-17.80	1.18***
Imports	60.81	65.04	-16.81	26.83
Interest Paid	67.69	34.07^{+}	-10.96	-8.85^{+}
Interest Rec.	64.00**	80.54	-4.40**	37.72
Liabilities	38.22	1.97***	-11.68	0.28***
R&D	82.45	66.16^{+}	-20.50	4.40
Sales	27.12	18.55**	-9.16	5.97*
U.S. Income Tax	91.57*	111.10	-0.57	-4.26

Results of pairwise *t* tests are indicated by the superscript asterisks (*) and plus signs (+). If, for a given field, one of the methods has a percent absolute error or percent error that is statistically significantly closer to zero than the corresponding error for the other method, then its error percentage is marked with an asterisk or plus sign according to the following scheme: *** if the difference is significant at the α = 0.0001 level, ** if significant at α = 0.001, * if significant at α = 0.01, and + if significant at α = 0.05.

Tables 5 and 6 present, respectively, complete results for donor imputation methods for the 10D and a pairwise comparison of impute_knn and impute_pmm. In partial contrast to the preceding results, these two tables indicate that predictive mean matching (impute_pmm) is the best donor imputation method for the 10D. As seen in table 5, the sequential and random hot deck methods tend to have much higher percent errors for most fields, in both absolute and non-absolute terms, than do impute_knn and impute_pmm. The pairwise comparison of impute_knn and impute_pmm in table 6 indicates that the imputations generated by impute_pmm tend to be more accurate than those generated by impute_knn. As was the case with the 15B, impute_knn tends to impute a larger proportion of FTIs than does impute_pmm. However, this disadvantage in terms of the number of imputations generated is of less concern than it was with the 15B because impute_pmm generates imputations for nearly half of all FTIs, and, as will be seen below, impute_pmm is paired with a model-based imputation method that is able to generate imputations for a high proportion of FTIs on the 10D.

[26] For each field, two pairwise t tests were conducted: one for percent absolute error and one for percent error. The tests for percent absolute error are based on calculating, for both methods, the absolute difference between each individual imputation and its corresponding reported value and testing whether the mean of these differences across all simulation runs is greater for one method than the other. For percent error, the tests involve calculating, for both methods, the difference between the sum of the imputed values and the sum of their corresponding reported values for each simulation run and testing whether the mean of the absolute value of these differences is greater for one method than the other. In other words, the tests for percent error are of whether the differences for one method are closer to zero than those of the other method rather than of whether the raw difference between the two methods' errors is itself significant.

Table 5. Comparison of Donor Imputation Methods for the BE-10D

Field	**Pct. Abs. Error**			
	impute_knn	**impute_shd**	**impute_rhd**	**impute_pmm**
Assets	85.84	700.27	135.44	61.81
Debts Payable	118.46	114.28	107.84	130.49
Debts Receivable	146.53	174.59	125.49	157.18
Employment	116.45	159.71	154.58	101.81
Liabilities	83.98	260.11	146.69	66.19
Net Income	134.44	197.30	174.07	110.95
Sales	81.47	121.34	136.77	70.45

Field	**Pct. Error**			
	impute_knn	**impute_shd**	**impute_rhd**	**impute_pmm**
Assets	-6.33	699.49	9.61	0.73
Debts Payable	-44.51	-66.94	-81.98	6.12
Debts Receivable	-24.56	-2.09	-66.04	6.59
Employment	-7.16	30.40	2.18	4.66
Liabilities	-14.16	175.47	-1.11	-0.36
Net Income	20.84	-120.61	-17.01	3.56
Sales	-12.99	27.80	-1.06	1.12
Pct. Imputed	**99.08**	**98.62**	**97.77**	**47.94**

Table 6. Pairwise Comparison of impute_knn and impute_pmm for the BE-10D

Field	**Pct. Abs. Error**		**Pct. Error**	
	impute_knn	**impute_pmm**	**impute_knn**	**impute_pmm**
Assets	69.01	62.14***	-10.32	0.50***
Debts Payable	116.31***	131.15	-44.82	5.51***
Debts Receivable	144.05***	159.15	-28.87	6.48**
Employment	115.72	101.81***	-7.14	4.66^{+}
Liabilities	75.54	66.19***	-14.80	-0.36***
Net Income	125.40	110.95***	22.87	3.56^{+}
Sales	76.53	70.45***	-14.82	1.12***

The next step is to choose model-based imputation methods. The results for the 15B are in tables 7 and 8a to 8f. These tables present, respectively, the complete results for all model-based methods and pairwise comparisons of the robust linear models (impute_rlm) method with each other method. The results in these tables tend to support the conclusion that impute_rlm is the best model-based imputation method for use with the 15B. To be sure, when the results are based on all imputations made by each method, as in table 6, other methods, notably impute_en and impute_rf, have somewhat lower percent errors and percent absolute errors for many of the fields on the 15B. However, impute_rlm imputed a much larger proportion of simulated FTIs than the methods that appear to outperform it in table 7 (see the last row of the table), suggesting that its higher average errors may be due to its having imputed more items that are relatively hard to impute and for which the other methods generated no imputations at all.[27]

Evidence that impute_rlm does in fact tend to produce imputations that are as accurate or more accurate than the other methods is provided in tables 8a to 8f, in which it is compared one-to-one against each of the other model-based methods using only the fields that both methods imputed. To be sure, the comparison of impute_rlm and impute_en in table 8b suggests that a case might be made for selecting impute_en rather than impute_rlm. For nine of the seventeen fields, impute_en produces imputations that are more accurate than those produced by impute_rlm as measured by one or both error measures, whereas impute_rlm only performs better than impute_en on six of the seventeen fields. Despite this slight advantage for impute_en in the pairwise comparison, impute_rlm was selected for use with the 15B based on two considerations. The first consideration is based on the fact that, since different sets of imputations are used in each pairwise comparison, transitivity does not necessarily hold between pairwise comparison results. That is, even though impute_en slightly outperforms impute_rlm in their pairwise comparison and impute_rlm outperforms all other methods, impute_en does not actually outperform all of the other methods in pairwise comparisons.[28] The second consideration is that impute_rlm imputes a larger proportion of FTIs than impute_en or any other method that performed reasonably well on the 15B, such as impute_rf and impute_mf.

One final point regarding the results in tables 7 and 8a to 8f concerns the performance of impute_lm.[29] Recall that impute_lm is equivalent to the regression-based imputation procedure provided by Banff with Proc Estimator. Tables 7 and 8a indicate that impute_lm is clearly outperformed by a number of the other methods available in Simputation, especially impute_rlm. The results in these tables thus show that the availability of these additional model-based imputation methods is an advantage of Simputation over Banff.

[27] An item on a given form will be more difficult to impute to the extent to which fields with which it is correlated do not have valid data on which to base the imputation. For example, since employment and employee compensation tend to be highly correlated, it is more difficult to impute employment when employee compensation also needs to be imputed.

[28] In particular, impute_rf outperforms impute_en in their pairwise comparison.

[29] The extraordinarily high percent error for the imputations of capital gains made by impute_lm is primarily a function of two factors: (1) reported values of capital gains, which can be positive or negative, tend to be relatively symmetrically distributed around zero, and (2) impute_lm’s imputations systematically underestimate the reported values of capital gains, usually by a wide margin. As a result of the first factor, the denominator in the calculation of percent error tends to have a low (absolute) value, as positive and negative reported values cancel one another out. The result of the second factor, on the other hand, is a numerator that is the sum of almost entirely negative errors that, naturally, do not cancel one another out.

Table 7. Comparison of Model-Based Imputation Methods for the BE-15B

Field				Pct. Abs. Error			
	impute_lm	impute_rlm	impute_en	impute_cart	impute_rf	impute_mf	impute_em
Assets	28.38	22.68	30.79	61.95	31.81	37.41	26.55
Capital Gains	625.86	104.64	121.47	111.57	119.47	121.51	207.76
Employment	21.68	31.08	31.58	59.63	23.43	48.05	48.66
Emp. Comp.	23.86	30.74	16.51	53.96	22.62	39.46	28.07
Mfg. Emp.	84.52	90.88	83.52	72.82	74.11	71.36	68.52
PP&E Exp.	139.80	93.51	78.67	98.98	85.26	93.04	105.27
Exports	82.20	54.08	29.51	67.34	34.52	69.94	89.86
Gross PP&E	39.02	25.85	13.00	76.72	25.27	44.52	41.46
Imports	109.93	61.74	37.78	63.37	44.09	92.92	134.85
Interest Paid	244.36	67.45	66.91	58.68	51.24	57.56	169.42
Interest Rec.	442.25	68.09	68.36	103.59	99.90	103.48	392.74
Liabilities	35.87	34.52	20.34	50.37	28.33	37.24	26.04
Net Income	102.60	96.71	84.93	97.17	91.57	87.34	80.79
Owners' Equity	43.33	43.86	12.70	62.56	26.10	50.61	37.67
R&D	97.26	27.41	22.62	93.29	56.63	73.64	62.31
Sales	27.85	31.62	23.32	79.22	26.86	38.64	33.89
U.S. Inc. Tax	237.56	106.45	96.90	83.48	78.64	78.51	135.33

Field				Pct. Error			
	impute_lm	impute_rlm	impute_en	impute_cart	impute_rf	impute_mf	impute_em
Assets	-0.91	-6.98	-18.42	-10.60	2.37	-2.13	-3.64
Capital Gains	99,059.04	-92.92	-113.66	-96.86	-126.16	-185.12	429.28
Employment	5.72	-3.96	-6.37	11.92	-0.15	17.95	23.72
Emp. Comp.	4.61	-2.42	-2.84	-8.11	-5.74	2.47	6.40
Mfg. Emp.	26.12	49.83	23.58	-28.10	-29.01	-24.83	55.74
PP&E Exp.	88.28	17.89	-13.86	-21.59	-10.54	-19.65	58.81
Exports	54.02	6.15	9.03	-17.25	-11.63	-16.82	64.41
Gross PP&E	17.26	-4.13	-3.81	2.73	-5.19	7.64	26.10
Imports	87.53	20.68	2.54	-13.59	-17.01	-7.21	115.36
Interest Paid	205.08	14.09	-25.37	-19.91	-8.91	-14.32	103.03
Interest Rec.	406.96	0.73	-19.34	-47.06	-44.04	-39.83	367.06
Liabilities	15.26	14.44	-12.40	-13.26	-6.76	-2.82	7.51
Net Income	-39.51	-82.97	-49.88	-57.06	-36.66	-37.06	8.32
Owners' Equity	-25.44	-34.83	-8.03	-2.96	3.92	-14.50	-10.01
R&D	82.20	-2.80	7.31	-13.85	-33.12	-17.60	40.22
Sales	6.76	10.30	-1.91	21.19	-1.23	-1.55	0.18
U.S. Inc. Tax	321.30	72.33	21.53	-4.81	4.35	4.95	138.77
Pct Imputed	**65.21**	**80.05**	**55.16**	**83.21**	**55.38**	**35.94**	**66.61**

Tables 8a-8f. Pairwise Comparisons of Model-Based Imputation Methods for the BE-15B

Table 8a. Comparison of impute_rlm and impute_lm

Field	Pct. Abs. Error		Pct. Error	
	rlm	lm	rlm	lm
Assets	23.41***	28.38	-8.19	-0.91
Capital Gains	105.19***	623.94	-170.38***	108,959.09
Employment	20.74***	22.29	-1.33**	6.02
Emp. Comp.	19.10***	24.07	-1.06*	4.74
Mfg. Emp.	88.09	83.14***	47.15	24.91***
PP&E Exp.	94.32***	138.57	19.09***	87.04
Exports	52.34***	82.20	8.66***	54.02
Gross PP&E	25.72***	37.59	-4.09***	15.92
Imports	54.85***	109.38	23.88***	87.62
Interest Paid	62.64***	240.88	13.52***	201.51
Interest Rec.	46.23***	380.78	-7.48***	344.92
Liabilities	36.10	35.87	17.99	15.26***
Net Income	96.76***	102.60	-84.50	-39.52***
Owners' Equity	43.09	43.33	-36.85	-25.44^{+}
R&D	25.96***	97.26	-2.19***	82.19
Sales	25.57***	27.45	5.09**	7.65
U.S. Inc. Tax	98.39***	237.58	58.74***	321.40

Table 8b. Comparison of impute_rlm and impute_en

Field	Pct. Abs. Error		Pct. Error	
	rlm	en	rlm	en
Assets	25.41*	31.03	-9.32*	-18.75
Capital Gains	104.47**	120.72	-95.00	-112.62
Employment	17.54***	30.79	-0.27***	-5.11
Emp. Comp.	19.50	15.76***	-1.89	-2.35
Mfg. Emp.	84.95	79.49*	45.00	19.60***
PP&E Exp.	88.66	77.76*	22.48	-14.84
Exports	28.70	30.86	11.83	10.18
Gross PP&E	15.52	12.85*	-5.45	-3.96^{+}
Imports	30.85**	37.21	1.54^{+}	2.37
Interest Paid	45.15***	65.33	-4.47***	-27.10
Interest Rec.	57.83^{+}	66.44	-7.88*	-22.02
Liabilities	29.12	20.73**	6.18	-13.00
Net Income	96.00	85.65***	-79.78	-49.40*
Owners' Equity	21.73	12.35^{+}	-12.39	-8.44
R&D	20.23	22.54	2.59	7.79
Sales	25.53	22.99	5.30	-1.60*
U.S. Inc. Tax	99.14	96.67	61.04	21.76***

Table 8c. Comparison of impute_rlm and impute_cart

Field	Pct. Abs. Error		Pct. Error	
	rlm	cart	rlm	cart
Assets	19.67***	61.17	-2.44***	-12.18
Capital Gains	107.41	112.24	-97.94	-89.74
Employment	30.96***	58.81	-2.55***	10.97
Emp. Comp.	31.84***	52.83	-3.26*	-8.67
Mfg. Emp.	93.22	68.42***	53.34	-32.24***
PP&E Exp.	93.22	95.64	20.50	-19.95
Exports	55.13***	67.42	6.33	-17.46
Gross PP&E	26.55***	72.92	-3.77**	-0.56
Imports	62.12	60.78	20.91	-15.40
Interest Paid	66.92	55.05***	11.02	-20.19
Interest Rec.	67.61***	100.96	0.33***	-51.79
Liabilities	34.74^{+}	49.88	14.54	-15.19
Net Income	96.69	96.88	-83.36	-58.95*
Owners' Equity	43.86*	61.68	-34.83	-4.42
R&D	30.00***	94.62	-3.11***	-13.29
Sales	33.90***	78.43	11.88***	24.50
U.S. Inc. Tax	106.66	82.61***	72.57	-3.63***

Table 8d. Comparison of impute_rlm and impute_rf

Field	Pct. Abs. Error		Pct. Error	
	rlm	rf	rlm	rf
Assets	22.72**	32.08	-2.57	2.15
Capital Gains	106.91	119.25	-110.50	-121.58
Employment	20.12*	23.29	0.98	0.04
Emp. Comp.	19.49*	21.90	-1.80^{+}	-4.88
Mfg. Emp.	87.56	69.55***	47.34	-33.70**
PP&E Exp.	86.90	84.54	22.01	-10.75
Exports	29.12***	35.84	11.04	-11.34
Gross PP&E	18.69***	25.09	-8.51	-5.65
Imports	31.19***	43.80	0.88***	-16.77
Interest Paid	48.77	48.98	-3.38	-10.63
Interest Rec.	64.53***	99.96	-1.92***	-48.85
Liabilities	32.91	28.18	8.36	-8.03
Net Income	96.66	92.30^{+}	-84.25	-37.09**
Owners' Equity	0.00**	26.10	0.00*	3.92
R&D	20.87***	57.07	2.12***	-33.32
Sales	26.34	26.85	5.93	-1.37^{+}
U.S. Inc. Tax	97.75	78.34***	57.76	5.01***

Table 8e. Comparison of impute_rlm and impute_mf

Field	Pct. Abs. Error		Pct. Error	
	rlm	mf	rlm	mf
Assets	19.41^{*}	37.08	-11.14	-2.54
Capital Gains	113.22	122.84	-288.33	-161.34
Employment	28.06^{***}	46.96	-2.09^{***}	17.85
Emp. Comp.	25.82^{***}	37.95	0.57^{+}	2.82
Mfg. Emp.	91.72	66.14^{***}	50.80	-30.05^{***}
PP&E Exp.	98.06	89.80	17.35	-17.6
Exports	58.44	66.37	7.78	-11.11
Gross PP&E	26.52^{***}	42.11	-4.88^{*}	5.76
Imports	68.98	88.81	26.90	-11.34
Interest Paid	65.58	53.76^{***}	9.26	-15.75
Interest Rec.	71.03^{***}	102.59	4.25^{**}	-46.55
Liabilities	29.89	34.48	9.44	-5.94
Net Income	96.75	88.15^{**}	-82.87	-37.20^{**}
Owners' Equity	43.60	48.58	-33.78	-16.73
R&D	35.45^{***}	71.81	-10.42	-19.16
Sales	34.43	33.8	6.89	0.21^{+}
U.S. Inc. Tax	111.19	77.66^{***}	84.98	6.59^{***}

Table 8f. Comparison of impute_rlm and impute_em

Field	Pct. Abs. Error		Pct. Error	
	rlm	em	rlm	em
Assets	20.66^{***}	25.30	-9.56	-4.81
Capital Gains	104.44^{***}	204.14	-94.48^{**}	401.76
Employment	26.04^{***}	42.59	-2.63^{***}	21.10
Emp. Comp.	23.76^{*}	26.39	-0.52^{+}	6.07
Mfg. Emp.	75.27	57.32^{***}	43.03	44.46
PP&E Exp.	95.32	101.80	19.72^{***}	62.63
Exports	51.35^{***}	90.00	4.57^{***}	65.05
Gross PP&E	24.22^{*}	38.99	-5.51^{**}	23.72
Imports	57.82^{***}	131.19	17.80^{***}	112.20
Interest Paid	61.59^{***}	140.13	10.22^{***}	72.23
Interest Rec.	86.87^{***}	321.49	8.25^{***}	296.75
Liabilities	33.78	24.38^{***}	13.67	5.76^{***}
Net Income	96.64	80.85^{***}	-84.39	5.12^{***}
Owners' Equity	39.23	37.35	-22.44	-10.37
R&D	23.75^{***}	62.31	0.24^{***}	40.91
Sales	29.79	29.85	9.91	7.22^{+}
U.S. Inc. Tax	103.47^{***}	133.41	68.33^{***}	139.12

The results for model-based imputation methods for the 10D are presented in tables 9 and 10a to 10f. These results provide strong support for the conclusion that imputation based on iterative random forest estimates (impute_mf) is the most accurate method for the 10D. While impute_rlm does perform better than impute_mf—and all other methods—in regard to percent absolute error (see table 9), it does considerably worse than every method on percent error, clearly disqualifying it as the model-based method for use with the 10D.[30] Moreover, tables 10a to 10f provide pairwise comparisons of impute_mf with each other model-based method, and impute_mf clearly outperforms every other method in these comparisons.

[30] While the high percent errors for impute_rlm on all seven fields are undeniably problematic, it bears mentioning that, for debts payable and debts receivable, table 9 may overstate the degree of error in impute_rlm's imputations. In the "clean" dataset, the vast majority of reported values for debts payable and debts receivable are zero, and, based on this fact, impute_rlm has imputed zero for almost every single simulated FTI for these two fields. In other words, s_{ijk} is almost always zero, and, disregarding for a moment the small number of cases where it is not zero, the equation for percent error simplifies to $\frac{\sum_{k=1}^{m}\sum_{j=1}^{n}-o_{ij}}{\sum_{k=1}^{m}\sum_{j=1}^{n}o_{ij}} \times 100$. The resulting value is -100 or very close to it, depending on how frequently non-zero values were imputed. Moreover, well over 50 percent of the imputations produced by impute_rlm for debts payable and debts receivable are exactly equal to their reported values (i.e., zero), a much higher percentage than that achieved by any of the other imputation methods for these two fields.

Table 9. Comparison of Model-Based Imputation Methods for the BE-10D

Field				**Pct. Abs. Error**			
	impute_lm	**impute_rlm**	**impute_en**	**impute_cart**	**impute_rf**	**impute_mf**	**impute_em**
Assets	71.32	56.40	70.78	76.78	55.13	69.47	81.79
Debts Payable	140.69	100.00	140.43	145.71	133.54	135.19	147.06
Debts Receivable	164.48	100.00	167.39	166.13	162.78	166.51	166.63
Employment	97.63	87.22	97.48	100.79	95.09	99.61	98.45
Liabilities	72.99	69.47	70.31	81.02	64.66	69.41	75.80
Net Income	103.62	94.43	102.84	101.20	97.09	104.63	101.86
Sales	91.54	79.29	87.45	82.54	61.03	62.76	94.80

Field				**Pct. Error**			
	impute_lm	**impute_rlm**	**impute_en**	**impute_cart**	**impute_rf**	**impute_mf**	**impute_em**
Assets	2.37	-21.05	2.44	-17.35	2.89	2.16	1.30
Debts Payable	0.31	-99.97	-0.21	-6.02	-3.93	7.77	-6.00
Debts Receivable	-5.45	-99.99	-1.63	-7.54	-0.72	13.55	-4.56
Employment	-15.67	-62.92	-15.87	-12.75	-11.45	5.41	-15.03
Liabilities	-0.46	-24.03	-1.34	-5.43	-1.98	3.92	-1.46
Net Income	-17.83	61.71	15.85	-69.11	-12.75	23.10	-15.01
Sales	4.45	-31.11	3.35	0.32	-0.54	0.78	8.34
Pct. Imputed	**72.87**	**76.44**	**69.38**	**100.00**	**52.68**	**97.01**	**93.08**

Tables 10a-10f. Pairwise Comparisons of Model-Based Imputation Methods for the BE-10D

Table 10a. Comparison of impute_mf and impute_lm

Field	Pct. Abs. Error		Pct. Error	
	mf	lm	mf	lm
Assets	54.93***	71.32	1.33*	2.37
Debts Payable	133.70***	140.69	6.65	0.31^{+}
Debts Rec.	165.55	164.48	12.54	-5.45
Employment	99.29	97.63	5.24***	-15.67
Liabilities	63.30***	72.99	2.10	-0.46
Net Income	98.08***	103.62	-6.61	-17.83
Sales	59.25***	91.54	-1.89**	4.45

Table 10b. Comparison of impute_mf and impute_rlm

Field	Pct. Abs. Error		Pct. Error	
	mf	rlm	mf	rlm
Assets	54.77***	56.40	1.44***	-21.05
Debts Payable	134.62	100.00***	7.65***	-99.97
Debts Rec.	165.10	100.00***	12.15***	-100.00
Employment	98.99	87.22***	5.13***	-62.92
Liabilities	63.32***	69.47	2.08***	-24.03
Net Income	98.08	94.43***	-6.61*	61.71
Sales	59.03***	79.29	-1.86***	-31.11

Table 10c. Comparison of impute_mf and impute_en

Field	Pct. Abs. Error		Pct. Error	
	mf	en	mf	en
Assets	54.66***	70.78	1.40**	2.44
Debts Payable	133.21***	140.43	5.99	-0.21^{+}
Debts Rec.	160.22***	167.39	6.51	-1.63
Employment	99.38	97.48**	5.23***	-15.87
Liabilities	63.26***	70.31	2.09	-1.34
Net Income	97.59***	102.84	-2.91	15.85
Sales	59.23***	87.45	-1.89*	3.35

Table 10d. Comparison of impute_mf and impute_cart

Field	Pct. Abs. Error		Pct. Error	
	mf	cart	mf	cart
Assets	69.47***	76.83	2.16***	-16.86
Debts Payable	135.19***	142.32	7.77	-9.48
Debts Rec.	166.51	164.52	13.55	-9.17
Employment	99.61	100.46	5.41*	-13.09
Liabilities	69.41***	81.02	3.92^{+}	-5.43
Net Income	104.63	101.20***	23.10*	-69.11
Sales	62.76***	79.92	0.78	-2.29

Table 10e. Comparison of impute_mf and impute_rf

Field	Pct. Abs. Error		Pct. Error	
	mf	rf	mf	rf
Assets	53.64***	55.26	2.11	2.93
Debts Payable	133.11	133.41	5.29	-4.08
Debts Rec.	159.86^{+}	162.46	6.18	-1.05
Employment	98.44	94.99***	4.87^{+}	-11.56
Liabilities	58.73***	64.66	-0.77	-1.98
Net Income	97.45	97.09	-5.76	-12.75
Sales	53.90***	61.03	-2.11	-0.54

Table 10f. Comparison of impute_mf and impute_em

Field	Pct. Abs. Error		Pct. Error	
	mf	em	mf	em
Assets	69.47***	82.31	2.16	1.40
Debts Payable	135.19***	144.16	7.77	-4.70
Debts Rec.	166.51	165.42	13.55	-5.24
Employment	99.61	97.85*	5.41**	-15.61
Liabilities	69.40***	75.80	3.77	-1.46
Net Income	104.63	101.86***	23.10	-15.01
Sales	62.71***	90.67	1.10***	4.32

5. Comparison of Banff and Simputation-Based Auto-Editing Systems

Based on the results in the preceding section, complete Simputation-based auto-editing systems were created for both the BE-10D and 15B. The former combines iterative random forest estimates (impute_mf) and predictive mean matching (impute_pmm), and the latter is a combination of the *k*-nearest neighbor (impute_knn) and robust linear model (impute_rlm) imputation methods. In this section, these Simputation-based auto-editing systems are compared using the simulation framework to the previously developed Banff auto-editing systems for the 10D and 15B. Recall that to avoid an overfitting bias, the datasets used for the simulation runs in this section (2014 data for both forms) are different from the datasets used to select imputation methods in the previous section (2015 data for the 15B and 2019 data for the 10D).

Tables 11 and 12 present pairwise comparisons of the complete Banff and Simputation-based auto-editing systems for the 15B and 10D, respectively, based only on the results for simulated FTIs imputed by both systems.[31] The results for the 15B are mixed, but overall the Banff-based system performed somewhat better than the Simputation-based system. Banff produced more accurate imputations for seven of the seventeen fields according to one or both of the error measures, while Simputation produced more accurate imputations for six of the fields.

Table 11. Pairwise Comparison of Banff and Simputation for the BE-15B

Field	Pct. Abs. Error		Pct. Error	
	Simputation	Banff	Simputation	Banff
Assets	31.00	14.26^{***}	-16.81	-6.67^{***}
Capital Gains	119.35^{***}	292.93	-119.38	15.88
Employment	93.36	38.78	67.42	-3.52^{+}
Emp. Comp.	24.78^{***}	39.70	-7.16^{***}	-17.11
Mfg. Emp.	52.87	42.30^{***}	-17.52	-19.47
PP&E Exp.	79.72	74.76	-7.17	-13.79
Exports	54.10^{**}	58.28	-12.63	-1.05^{**}
Gross PP&E	44.50	36.30	25.96	-16.85
Imports	71.27^{***}	89.34	13.14^{***}	47.90
Interest Paid	61.13^{***}	83.88	13.65^{***}	36.37
Interest Rec.	56.08	50.07^{*}	2.03	4.57
Liabilities	24.58	17.19	4.00	8.03
Net Income	94.98	81.12^{+}	-87.20	-50.59^{**}
Owners' Equity	38.14	35.86	-7.18^{*}	-33.28
R&D	36.60	26.90^{+}	-23.09	5.97^{**}
Sales	28.69^{***}	40.28	7.41	-6.56
U.S. Inc. Tax	100.95	100.13	57.00	-27.47^{***}

[31] Separate tables with the "complete" results are not presented in this section. By design, the auto-editing systems impute a very high proportion of FTIs. As a result, the differences between the "complete" results and the results using only items imputed by both auto-editing systems are minimal, and little additional analytical leverage can be gained by also including an examination of the "complete" results.

In contrast, there is little ambiguity about the results for the 10D in table 12. Simputation clearly outperforms Banff on every field except for one, and that field, debts receivable, is essentially a tie as neither auto-editing system has an error statistic that is statistically significantly lower on either measure of error. The superior performance of the Simputation-based system on assets, liabilities, and sales is especially striking.

Table 12. Pairwise Comparison of Banff and Simputation for the BE-10D

Field	Pct. Abs. Error		Pct. Error	
	Simputation	Banff	Simputation	Banff
Assets	68.19***	117.76	3.62***	4.22
Debts Payable	112.65***	119.95	-0.04***	-13.41
Debts Receivable	164.05	164.52	11.28	-6.91
Employment	99.30***	103.94	5.46*	-9.42
Liabilities	66.49***	118.42	2.12***	5.66
Net Income	101.58***	123.86	21.69	22.40
Sales	61.85***	81.86	1.74***	-12.28

Taken together, the results in tables 11 and 12 indicate that Simputation, in combination with other R packages from Statistics Netherlands, does offer a viable alternative to Banff. Banff did perform somewhat better than Simputation on the 15B. However, Simputation did no worse or better than Banff on ten of the seventeen fields subject to auto-editing on the 15B—six fields on which at least one of its errors was statistically significantly lower plus four on which neither method clearly outperformed the other—and Simputation produced imputations for the 10D that were generally far superior to those produced by Banff. Given these results, it seems clear that both software packages should be under consideration for any additional forms that BEA decides to auto-edit.

6. Conclusion

BEA relies on auto-editing programs to help process its multinational enterprise surveys, which are used to produce widely used statistics on U.S. direct investment abroad and foreign direct investment in the United States. All auto-editing systems in use at BEA were, until recently, built around the Banff system for data editing and imputation. To enhance its flexibility, BEA has explored the feasibility of building auto-editing systems around a set of R packages created by analysts at Statistics Netherlands. This paper has examined an issue that is of central importance to whether these R packages offer a suitable alternative to Banff: can they—and in particular can the Simputation package—produce imputations that are, on average, as accurate as those produced by Banff? This question was addressed by first creating Simputation-based auto-editing systems for two of BEA's MNE survey forms, the BE-10D and the BE-15B, and then comparing the accuracy of their imputations to those produced by Banff using a simulation-based framework.

In one sense, the results of this analysis were mixed, as Simputation produced imputations for the 15B that were slightly less accurate than Banff's and generated imputations for the 10D that were significantly more accurate than Banff's. These results clearly show that neither software option is universally superior to the other. From another point of view, though, the results are a triumph for Simputation since they show that it has the potential to produce imputations that are as accurate or even more accurate than Banff. Indeed, the results unquestionably show that in the future BEA should consider both Banff and the Statistics Netherlands packages when developing new auto-editing systems.

It is likely that which of these two software options is best for a given form will depend on the characteristics of the form in question. Banff is likely to perform well relative to Simputation when the form being auto-edited has, as the 15B does, a large number of edits that identify relationships between fields on the form. Proc DonorImpute is able to use the information in these edits to identify matching fields, while Simputation's donor imputation methods do not have that capability. Moreover, recall that in section 4 impute_rlm outperformed impute_lm on the 15B and that impute_lm is effectively the same procedure as Proc Estimator. Since Banff still produced imputations that were overall somewhat more accurate than those produced by Simputation for the 15B, its advantage over Simputation likely came from Proc DonorImpute and this advantage was likely due to the information encoded in the edits regarding how to match donors and recipients.[32] However, when the edits do not specify extensive interrelationships among fields, as is the case for the 10D, Simputation may have the advantage since it offers a much wider variety of imputation methods to choose from than does Banff.

A limitation of this study is that it was not able to examine all possible ways of setting up auto-editing systems based on Banff and Simputation. The possible ways of selecting, ordering, and calibrating the imputation procedures for a given form are virtually limitless with both Banff and Simputation. These software packages provide menus of possible imputation procedures, and it is up to the analyst to decide which to use, the order in which to use them, and how to customize each one that is used. As a result, there is a necessary degree of uncertainty in any search for an optimal set of imputation procedures. This study sought to reduce this uncertainty by limiting its search to the single best model-based method and the single best donor method available in Simputation and by taking a highly systematic approach, using a simulation-based framework, to the comparison of potential model-based and donor imputation methods. The resulting Simputation-based auto-editing systems for the 15B and 10D are not guaranteed to be the best possible, but the rigorousness of their development (as well as of Banff-based systems previously developed) should ensure that the comparisons of Simputation and Banff in section 5 are meaningful and useful.

[32] An objection to this conclusion could be raised on the grounds that the superior performance of impute_rlm vis-à-vis impute_lm in section 4 might have been due to idiosyncrasies of the 2015 data and that the impute_rlm-impute_lm comparison might have turned out differently with the 2014 BE-15B data. In this scenario, the superior performance of Banff relative to Simputation in section 5 would be due, partly or entirely, to Proc Estimator since Proc Estimator and impute_lm are equivalent. To address this potential concern, a comparison of impute_rlm and impute_lm using 2014 data was conducted using the same procedures as with the 2015 data reported in section 4, and the results obtained were essentially the same as those in section 4. In other words, the superior performance of impute_rlm relative to impute_lm holds with the 2014 data, which strengthens the conclusion that Banff's superior performance relative to Simputation in section 5 is due to Proc DonorImpute, and particularly to the advantage the large number of edits for the 15B provide to Proc DonorImpute.

References

Banff Support Team. 2017. *Functional Description of the Banff System for Edit and Imputation*. Version 2.07. Statistics Canada, Ontario.

Barboza, W. and Turner, K. 2011. “Utilizing Automated Statistical Edit Changes in Significance Editing.” National Agricultural Statistics Service, Joint Statistical Meetings.

Beaumont J.F. and Bocci, C. 2009. “Variance Estimation When Donor Imputation is Used to Fill in Missing Values.” *The Canadian Journal of Statistics,* 37(3): 400-–416.

Bianchi, G., R. Filippini, R.M. Lipsi, A. Pezone, F. Scalfati. 2020. “An Overview of the Editing and Imputation Process of the 2018 Italian Permanent Census.” Paper presented at the *Conference of European Statisticians, Work Session on Statistical Data Editing*, Geneva, Switzerland, April 15–17.

Dempster, Arthur P., Nan M. Laird, and Donald B. Rubin. 1977. “Maximum Likelihood from Incomplete Data Via the EM Algorithm." Journal of the Royal Statistical Society, Series B (Methodological): 1–38.

Di Zio, Marco, Ugo Guarnera, Orietta Luzi, and Antonia Manzari. 2006. “Evaluating the Quality of Editing and Imputation: The Simulation Approach.” In *Statistical Data Editing, Vol. 3: Impact on Data Quality*, pp. 44–59. New York: United Nations.

Dorinski, Suzanne M. 1998. “Imputation Methods in the Sample Survey of Law Enforcement Agencies.” Bureau of the Census. Washington, D.C.

Dorinski, Suzanne M., Rita J. Petroni, Michael Ikeda, and Rajendra P. Singh. 1996. “Comparison and Evaluation of Alternative ICM Imputation Methods.” Washington, D.C.: Bureau of the Census.

De Waal, Ton, Jeroen Pannekoek, Sander Scholtus. 2011. *Handbook of Statistical Data Editing and Imputation*. Hoboken, NJ: John Wiley & Sons, Inc.

Fellegi, I.P. and Holt D. 1976. “A systematic approach to automatic edit and imputation.” *Journal of the American Statistical Association*, 71(353): 17–35.

Giles, Philip and Charles Patrick. 1986. “Imputation Options in a Generalized Edit and Imputation System.” *Survey Methodology* 12(1): 49–60.

Gray, Darren. 2018. “The Evolution of Banff in the Context of Modernization.” Paper presented at the *Conference of European Statisticians, Work Session on Statistical Data Editing*, Neuchatel, Switzerland, September 18–20.

Gray, Darren. 2020. “Evaluating Imputation Methods Using ImpACT: First Case Study.” Paper presented at the *Conference of European Statisticians, Work Session on Statistical Data Editing*, Geneva, Switzerland, April 15–17.

Huber, Peter J. 1981. *Robust Statistics*. Berlin: Springer, 1248–1251.

Johanson, J.M. 2012. “Banff Automated Edit and Imputation on a Hog Survey.” National Agricultural Statistics Service, American Statistical Association.

Kovar, J., P Whitridge, and J. MacMillan. 1988. “Generalized Edit and Imputation System for Economic Surveys at Statistics Canada.” Ottawa: Statistics Canada.

Kozak, Robert. 2005. “The Banff System for Automated Editing and Imputation.” SSC Annual Meeting, Proceedings of the Survey Methods Section.

Lange, Kerstin. 2020. “Automation of E&I Processes.” Paper presented at the *Conference of European Statisticians, Work Session on Statistical Data Editing*, Geneva, Switzerland, April 15–17.

Mohl, C. 2007. “The Continuing Evolution of Generalized Systems at Statistics Canada for Business Survey Processing.” In *Proceedings of the Third International Conference on Establishment Surveys* (ICESIII) (June 18–21, 2007), American Statistical Association, 758–768.

Salvucci, Sameena, Eric Grau, Yuhong Zheng, and Julie Ingels. 2012. “Assessing the Validity of an Imputation Method Using Data from Comparable External Sources.” Paper presented at the *Fourth International Conference on Establishment Surveys*. Montreal, Canada.

Scholtus, Sander. 2014. “Donor Imputation.” *Handbook on Methodology of Modern Business Statistics*. Statistical Office of the European Union (EuroStat).

Scholtus, Sander, Bart Bakker, and Sam Robinson. 2017. “Evaluating the Quality of Business Survey Data Before and After Automatic Editing.” Paper presented at the *Conference of European Statisticians, Work Session on Statistical Data Editing*, The Hague, Netherlands, April 24–26.

Scholtus, Sander and Jacco Daalmans. 2020. “Variance Estimation After Mass Imputation with an Application to the Dutch Population Census.” Paper presented at the *Conference of European Statisticians, Work Session on Statistical Data Editing*, Geneva, Switzerland, April 15–17.

Seyb, Allyson, John Stewart, Grace Chiang, Ian Tinkler, Lee Kupferman, Val Cox and Darren Allan. 2009. “Automated Editing and Imputation System for Administrative Financial Data in New Zealand.” Paper presented at the *Conference of European Statisticians, Work Session on Statistical Data Editing*, Neuchatel, Switzerland, October 5–7.

Stekhoven, D.J. and P. Buehlmann. 2012. “MissForest—Non-Parametric Missing Value Imputation for Mixed-Type Data.” *Bioinformatics* 28(1): 112–118.

Terrie, Larkin. 2018. “Assessing the Automated Imputation of Missing and Erroneous Survey Data: A Simulation-Based Approach.” *Proceedings of the 2018 Federal Committee on Statistical Methodology and Policy Conference*.

Whitridge, P. and J. Kovar. 1990. “Applications of the Generalized Edit and Imputation System at Statistics Canada.” Ottawa: Statistics Canada.

Xu, Mark, Andy Kim, and Larkin Terrie. 2017. “Automated Data Editing and Imputation for Surveys of Multinational Enterprises, a Banff Implementation.” Paper presented at the *Conference of European Statisticians, Work Session on Statistical Data Editing*, The Hague, Netherlands, April 24–26.

Zou, Hui and Trevor Hastie. 2005. “Regularization and Variable Selection via the Elastic Net.” *Journal of the Royal Statistical Society* 67(2): 301–320.