Assessing the Automated Imputation of Missing and Erroneous Survey Data: A Simulation-Based Approach

Larkin Terrie
Bureau of Economic Analysis

Proceedings of the 2018 Federal Committee on Statistical Methodology (FCSM) Research and Policy Conference

I. Introduction

Statistical data editing, which is a form of data cleaning, is key to ensuring the accuracy and reliability of survey-based official statistics. For its annual direct investment surveys, the Bureau of Economic Analysis (BEA) relies on teams of accounting experts to verify the completeness and accuracy of data reported on survey forms. These surveys are used to produce widely-used statistics on U.S. direct investment abroad ("outward" investment) and foreign direct investment in the United States ("inward" investment).[1] A drawback, though, of the current approach to editing these surveys is that it is relatively labor intensive, as the experts have to carefully review large numbers of individual, completed forms. The demands placed on BEA's expert survey editors have also increased in recent years as the number of companies filing direct investment surveys has significantly increased.

BEA has sought to reduce the number of survey forms that must be manually edited, and simultaneously improve the efficiency and cost effectiveness of survey editing, by developing an automated data editing and imputation system for use with its direct investment surveys. This auto-editing system, the development of which began in 2015, is based on an implementation of the Banff system for data editing and imputation produced by Statistics Canada, and it has already performed successfully in two pilot tests conducted by BEA.[2] These tests, which used data from two past surveys – one of inward direct investment and one of outward direct investment – found that the aggregate statistical results derived from Banff-edited forms were very similar to the results obtained when the same forms were edited by expert survey editors (Xu, Kim, and Terrie 2017). Based on these results, the auto-editing system was used, in a limited manner, to help edit BEA's 2014 Benchmark Survey of U.S. Direct Investment Abroad.[3]

This paper builds on this previous work by further investigating the quality of the data produced by BEA's auto-editing system.[4] It presents a new and more rigorous approach to assessing the quality of the auto-editing system's imputations for missing and erroneously reported survey items. While the two pilot tests were useful in showing that auto-editing does not entail a loss in data accuracy compared to traditional manual editing, they were based on a less than ideal approach to testing the quality of auto-editing imputations. Ideally, a rigorous test of the auto-editing system's imputations should be based on an assessment of how close imputed values are to the true values of the survey items in question rather than how close they are to the values assigned by editors.

[1] As defined in these surveys, a direct investment relationship 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] A brief introduction to Banff is provided in Mohl (2007). Other studies that discuss practical applications of Banff include Barboza and Turner (2011), Johanson (2012), Kosler (2012), and Winkler (2006).
[3] These surveys covered U.S. direct investment abroad that occurred in 2014. They were due to BEA in 2015, and statistics based on them were published in 2016.
[4] In doing so, it also contributes to a body of work on assessing the quality of imputations for survey data and, more generally, the quality of data produced by auto-editing (Dasu and Loh 2012, Salvucci et al. 2012, Sarndal 1992, Scholtus et al. 2017, UNECE 2006, Van der Loo et al. 2017).

To be sure, when a response is missing or reported erroneously, its true value is generally unknown. It is, however, possible to estimate the degree to which imputed values resemble true values by using a simulation-based technique. In this paper, the approach adopted is to simulate the presence of missing and erroneous data in survey forms that have neither missing responses nor responses identified as erroneous by BEA's validity checks. From these forms, items are selected for imputation in a way that mimics the actual distribution of erroneous and missing responses in all received forms, and are then imputed by the auto-editing system. The resulting imputed values are then compared to the original, and presumably correct, reported values. By repeating this process a large number of times – each time with a different set of responses being treated as missing or erroneous – it is possible to draw conclusions about the average accuracy of the auto-editing system's imputations.[5]

It bears emphasis that this approach assumes that data that pass BEA's validity checks (explained in more detail in section two) are error free. Since these checks cannot identify every conceivable kind of reporting error, it is possible that this assumption may, at times, be violated. Nonetheless, this assumption is necessary because, other than directly contacting respondents to verify all submitted data, the validity checks are the most reliable way of identifying errors in the data. Indeed, these checks have been carefully designed to identify many of the most common types of reporting errors as well as those that could have the largest effect on published statistics. Data that pass these checks are thus, if not completely free of error, as close to error-free as the present study can realistically obtain.

The proposed simulation-based testing framework is used to evaluate auto-editing of BEA's Annual Survey of Foreign Direct Investment in the United States, the BE-15. This survey collects financial and operating data, such as balance sheet items, employment, and research and development expenditures, on foreign-owned U.S. business enterprises, referred to as "U.S. affiliates." The survey is completed by the affiliates, who submit a BE-15A, BE-15B, or BE-15C form – the type of form filed depending on the affiliate's size and whether it is majority foreign-owned.

The present study is based on data submitted on the 2014 and 2015 BE-15C forms. The BE-15C collects information on U.S. affiliates with assets, net income, or sales of over $40 million (positive or negative) but for which none of these is greater than $120 million (positive or negative). These are the smallest companies covered by BEA's annual survey of inward direct investment, and the survey form itself includes fewer items than the BE-15A and BE-15B forms, both of which cover larger companies.[6] Its relative simplicity was a major reason the BE-15C was chosen for this initial study using the simulation-based testing framework. The small number of items on the form makes the set-up of the testing framework – as well as the presentation of its results – relatively straightforward, thereby enabling BEA to evaluate the framework's usefulness before applying it to more complex surveys.

This study finds that auto-editing imputations are generally of high quality, as on average across all of the simulation-based results, imputed values tend to be similar to their corresponding true values. In this sense, the results of this study are similar to, and reinforce the findings of, the two pilot studies mentioned above. However, this study goes beyond testing the adequacy of current practice and uses the simulation-based framework to identify ways of improving auto-editing imputation procedures. Multiple tests based on this framework were run, each on a different version of the auto-editing system. The versions of the auto-editing system that were tested differed from one another in that each used different settings for two key imputation procedures – donor imputation and estimator imputation. The results from these tests could thus be compared to determine which combination of settings

[5] This approach is somewhat similar to that described by Di Zio et al. (2006), who also outline a method based on simulated data. These authors propose creating entirely synthetic datasets in order to circumvent the problem of not knowing the true value of items being imputed. The approach presented here provides an alternative in which, instead of requiring the creation of a synthetic dataset for each run of the simulation, actual data are used and only the presence of items in need of imputation is simulated.

[6] The BE-15A covers U.S. affiliates that are majority foreign-owned and have total assets, sales, or net income of more than $300 million (positive or negative). The BE-15B covers U.S. affiliates that are either (1) majority foreign-owned and have total assets, sales, or net income of more than $120 million (positive or negative) but for which none of these is greater than $300 million (positive or negative) or (2) minority foreign-owned and have total assets, sales, or net income of more than $120 million (positive or negative).

produces the best imputations, and these comparisons yielded two key findings. First, the overall quality of both donor and estimator imputations is improved when they are based on multiple years of previous survey responses, rather than only the responses from the year of the survey being processed. Second, estimator imputation performs better than donor imputation for certain survey items and so the usual order of imputation procedures, in which donor imputation precedes estimator imputation, should be reversed for these items.

The rest of the paper proceeds as follows. The second section provides an overview of BEA's general approach to auto-editing. The third section explains the procedures that were used by the simulation-based testing framework to simulate the presence of missing and erroneous data in forms that had neither missing nor erroneous data. Section four provides details on each of the tests that were conducted using the simulation-based framework and explains the metrics that were used to evaluate and compare the results of these different tests. Section five presents and discusses results. Section six concludes.

II. BEA's Auto-Editing Systems

Each survey form-specific auto-editing system at BEA is developed around a set of validity conditions, or "edits." These edits are logical and mathematical rules, originally developed by BEA for manually editing survey forms and then adapted to auto-editing.[7] Each survey has its own specific set of edits, though some edits are applied to more than one survey. Edits define the relationships between different survey items as well as the range of allowable values for individual items. For example, an edit might specify that the ratio of employee compensation to the number of employees must be within a certain range or that certain items, such as number of employees, cannot have a negative value. Edits are used to identify survey items that have been erroneously reported and need to be replaced by imputations, and they are also used by certain imputation procedures (see below) to find appropriate replacement values.

At a programming level, the core of each auto-editing system is the Banff system for editing and imputation developed by Statistics Canada. Banff provides nine independent SAS procedures that can be used to identify and correct erroneous items in survey data. BEA has developed an approach to auto-editing its multinational enterprises survey data that relies on six of these procedures.[8] In the auto-editing systems that have been developed to date, they are run in the following order:

- **Proc VerifyEdits** checks that all edits are logically consistent with one another.
- **Proc ErrorLoc** is an error localization procedure that identifies fields to impute (FTIs) when a record violates one or more edits.[9]
- **Proc Deterministic** makes an imputation for an FTI when, based on the edits, the field in question has only one logically possible (integer) value.[10]
- **Proc DonorImputation** is a nearest-neighbor imputation procedure in which a record with one or more FTIs receives data from the valid record that is most similar to it in terms of key matching fields that are identified based on the edits.[11]

[7] Xu, Kim, and Terrie (2017) describe the adjustments that had to be made to the edits so that they could be used in auto-editing. These included linearizing non-linear edits and finding a way to process edits based on if-then conditions with Banff since Banff does not have the built-in capacity to process if-then edits.
[8] The choice of these six procedures is discussed in Xu, Kim, and Terrie (2017). Technical details on these procedures can be found in Banff Support Team (2012).
[9] This procedure chooses fields to impute based on the "rule of minimum change," which holds that as few fields as possible should be changed when altering a record so that it will pass the edits (Fellegi and Holt 1976, Sande 1979).
[10] For example, suppose x is a field requiring imputation for a particular record and that there are edits specifying that $x \geq y$ and $x \leq z$ and that y and z both have the valid value of 8 for this record. In this scenario, Proc Deterministic would impute a value of 8 for x.
[11] The matching fields are identified by an algorithm that takes into account edits that describe the relationship between the field requiring imputation and other fields. See Banff Support Team (2012).

- **Proc Estimator** enables imputation based on linear regression models and other kinds of estimator functions.
- **Proc ProRate** ensures that components of a sum, when summed, equal the desired total.

The relative order of the imputation procedures – deterministic, donor, and estimator – is particularly important because it reflects assumptions regarding the relative reliability of each type of imputation. Since each imputation procedure only creates imputations for FTIs (fields that have been identified by the error localization procedure as in need of imputation) that have not already been imputed by previous procedures, it is desirable to run the most reliable imputation procedures first. The relative ordering of donor and estimator imputation in particular is based on work demonstrating the highly robust and flexible nature of donor imputation (Chen and Shao 2000, Beaumont and Bocci 2009). However, as will be shown in the results presented below, the superiority of donor over estimator imputation is actually contingent on certain features of the field being imputed and it is thus desirable, under certain circumstances, to run estimator imputation before donor imputation.

An important limitation of the Banff system is that it can only identify errors in and make imputations for continuous, numeric data. The tests discussed in this paper are thus limited to survey items that take continuous, numeric values.[12] Due to the relative simplicity of the BE-15C form, it collects data on only eight continuous, numeric variables: assets; liabilities; net income; sales; research and development (R&D) expenditures; gross value of property, plant, and equipment (gross PP&E); number of employees; and total employee compensation.

With only eight fields to be edited by the Banff procedures, the way the BE-15C form is auto-edited has two unusual features in comparison to other forms for which auto-editing systems have been developed. First, none of these variables is equal to the sum of any of the others, so Proc ProRate is not used in auto-editing of the BE-15C. Second, the vast majority of the imputations for items on the BE-15C are made with either donor or estimator imputation. In general, deterministic imputation only occurs for a given FTI when there are multiple edits linking its value to the values of other fields with valid entries (see the example in footnote 10). With only eight numeric items on the BE-15C, the potential for these types of links between fields is limited since many of the items on the form have no necessary relationship to any of the other items on the form. For example, the only edits for gross PP&E and liabilities are those specifying upper and lower bounds. As a result, the potential for deterministic imputations is limited.

III. Modeling Missing and Erroneously Reported Data

The testing framework used in this paper involves selecting non-missing, non-erroneous survey responses to be treated as if they were either missing or erroneous – i.e., as if they were fields to impute (FTIs). A precondition for conducting tests using this framework is thus the creation of a dataset from which all records with missing data or erroneous data, as identified by Banff's error localization procedure, have been excluded. In each run of the simulation, a new set of responses is selected from this "clean" dataset to be treated as FTIs and then replaced by imputations produced by the auto-editing system.[13] By comparing the differences between the resulting imputed values and the corresponding actual values over a large number of runs, conclusions can be drawn about the average quality of the imputations for each of the eight numeric variables asked about on the survey.

[12] The categorical variables reported in BEA's surveys, such as the industry and country of the foreign parent in inward surveys, are edited using a stand-alone SAS program written by a BEA statistician in collaboration with the survey editors. This program identifies problematic variable values and recodes them based on a series of if-then statements, or, in the case of values that cannot be reliably recoded based on predetermined criteria, identifies records to turn over to the editors for manual editing.

[13] Since there are eight numeric fields on the survey form, each record will have between zero and eight of its responses selected for imputation in each run of the simulation.

The simulation requires a way to determine which fields will be treated as (simulated) FTIs in each run. One option is to assign FTI status completely at random, giving each of the $8 \times n$ fields (where n, in this case, is the number of "clean" records) an equal probability of being an FTI in each run.[14] Under this approach, though, the selection of items for imputation would not necessarily mimic the actual distribution of FTIs observed in all received forms since this approach does not involve modeling the distribution of erroneous and missing items (FTIs). Mimicking the actual distribution of FTIs is desirable because the distribution of FTIs (among fields and records) can affect how difficult it is to make accurate imputations and thus can affect the average quality of imputations. For the results of the simulation to be useful, the difficulty of making accurate imputations for simulated FTIs should be as close as possible to the difficulty of making imputations for actually observed FTIs.

Two aspects of the distribution of FTIs among fields and records have a particularly strong influence on the overall quality of imputations – in both cases because they affect the amount of information available for making imputations. First, if an FTI belongs to a company that did not report data in previous years – meaning its record lacks prior year values for the eight numeric variables – accurate imputations can be more difficult. Second, if a record has more than one FTI, it will generally be more difficult to impute its FTIs accurately. In regard to estimator imputation, these circumstances can reduce the quality of imputations by limiting the number of variables available for inclusion in the regression models used to estimate new values for FTIs.[15] In regard to donor imputation, these circumstances can limit the accuracy of imputations by reducing the information available for finding high quality donor-recipient matches.

The distribution of FTIs among the $8 \times n$ fields reported in a given year was modeled using logistic regression. The following model was fit for each of the j=1,…,8 numeric fields on the BE-15C form, where records (i.e., companies) are indexed by i=1,…,n and the binary dependent variable indicates whether the Y_{ij}th field is an FTI or not for a given record.

$$E\left[Y_{ij}\right] = \frac{\exp\left(\mathbf{X}_{ij}'\boldsymbol{\beta}\right)}{1 + \exp\left(\mathbf{X}_{ij}'\boldsymbol{\beta}\right)}$$

The resulting eight models were each fit using the unedited data that companies reported on the BE-15C in each of the six years from 2009 to 2011 and 2013 to 2015.[16] From these six years, there were 4,093 forms in total that were used to fit the models. Values for the dependent variables were obtained by using Banff's error localization procedure (Proc ErrorLoc) to identify which fields were FTIs in each year of the data.

Three types of independent variables (whose values for each record and model are represented by the vector $\mathbf{X}_{ij}$ in the above equation) are included in each model. The first is a set of dummies indicating which of the seven other numeric fields, if any, are also FTIs.[17] The second is a single dummy variable indicating whether or not prior year data exist for the record in question. The third is a set of dummies indicating to which major industry group the company belongs. The first two categories of independent variable are included because, as mentioned above, the presence of multiple FTIs or lack of prior year data affects the quality of imputations by reducing the amount of

[14] In a given year, the proportion of BE-15C forms that is "clean" in the sense of having no FTIs (as identified by Banff's error localization procedure) is generally a little over 80 percent. In the years 2009, 2010, 2011, 2013, 2014, and 2015, between 81 and 85 percent of BE-15C forms were clean in this sense.

[15] The lack of prior year data particularly affects the quality of imputations produced by estimator imputation. With the partial exception of net income, the values of the eight numeric variables tend to be strongly correlated with their prior year values. As a result, regression models that include as independent variables the prior year version of the dependent variable tend to produce better estimates of actual values than those that do not.

[16] Data from 2012 were excluded because 2012 was a benchmark year and the benchmark survey form differs in important ways from the annual survey form.

[17] In order to take into account which other fields have been identified as FTIs for a given record, the eight models are run sequentially. See also footnote 19.

information available on which imputations can be based.[18] The third is included because previous research on responses to BEA's inward investment surveys indicates that there is a relationship between industry and the quality of the responses received by BEA (Limés 2017).

The determination of which fields will be treated as FTIs in each run of the simulation is based on probabilities generated by these eight logistic regression models. That is, the models' estimated coefficients are used to "predict" the probability of a given survey item being an FTI based on the values of the independent variables for the record in question. In other words, an estimated probability of being an FTI, $p_{ij} = E[Y_{ij}]$, is assigned to each of the j=1,…,8 fields within each of the i=1,…,n records included in the test. In each run of the simulation, the assignment of FTI (or not FTI) status to each of the individual $8 \times n$ fields is determined by performing a random draw from a Bernoulli distribution where the parameter p is set equal to p_{ij}.[19,20]

This procedure was used to generate simulated FTIs for the three tests of the BE-15C auto-editing system that are discussed in subsequent sections. Each test was run twice, once with data from the 2014 survey and once with data from the 2015 survey, and each test is based on 5,000 runs of the simulation – meaning 5,000 separate sets of simulated FTIs. The usefulness of the results presented below depends, in part, on how closely these simulated FTIs resemble actual FTIs and on whether there tends to be meaningful variation among the 5,000 separate sets of simulated FTIs. The remainder of this section uses 10,000 separate sets of FTIs generated with the procedure outlined above (5,000 for the 2014 data and 5,000 for the 2015 data) to show that there is in fact a high degree of similarity, on average, between simulated FTIs and actual FTIs and that there tends to be significant variation between different sets of simulated FTIs.

Table 1 provides the distribution of FTIs by field, in both observed and simulated data, in terms of each field's average percentage share of all FTIs. The percentage share for observed data is averaged over the six years of unedited, reported data used to fit the logistic models discussed above, and the shares for simulated data are averaged over the 5,000 separate sets of simulated FTIs generated for the 2014 and 2015 data, respectively. The results indicate a high degree of similarity in regard to how frequently each field is identified as being in need of imputation in the actual and simulated data. In addition, table 2 indicates that the average number of FTIs per form is consistent across the actual and the simulated data. Taken together, these results indicate that the procedure presented above is doing a reasonably good job of generating simulated FTIs whose distribution mirrors that of actual FTIs. This finding is particularly important because section six will present results regarding how close the aggregate auto-edited total for each field is to its actual aggregate total, and these results are a more meaningful measure of the quality of imputations if the number of times each field is imputed in the simulation is similar to the number of times it is imputed in actual data.

[18] The first set of dummies also makes it possible to model whether certain fields are particularly likely to be missing or erroneous together, which is important to model in case fields that are highly correlated with one another are also particularly likely to be FTIs simultaneously. For example, (number of) employees and employee compensation tend to be strongly correlated with one another, so each provides useful information for making an imputation for the other in the case one is an FTI. If both are FTIs, though, accurate imputations will be more difficult for both.

[19] The draws were performed using the RAND function in SAS.

[20] The values of the p_{ij} are generally not constant across different runs of the simulation. The random draws for the eight fields are performed in a set sequence, which means, in essence, that the eight fields take turns having their FTI status determined for each of the n records. According to the logistic models described above, the probability that a given field for a given record will be assigned an FTI status depends in part on whether other fields in that record are also FTI. As a result, the values of the p_{ij} for the fields whose draws occur later in the sequence depend on the outcomes of the draws for the fields earlier in the sequence. For example, if the probability of employment being an FTI goes up when it is already known that employee compensation is an FTI, then the likelihood of employment being an FTI for a given record would be higher if employee compensation has already been assigned FTI status for that record.

Table 1: Distribution of FTIs among Survey Items in Actual vs. Simulated Data

Field Selected as FTI	Observed Average Percent Share	Simulated Average Percent Share (2014)	Simulated Average Percent Share (2015)
Assets	0.3	0.3	0.3
Liabilities	1.6	1.7	1.8
Sales	23.2	23.0	23.6
Net Income	4.5	5.1	5.5
Employee Compensation	23.9	25.4	25.2
Gross PP&E	18.2	16.0	16.2
R&D	9.2	8.4	8.2
Employees	19.1	20.1	19.2

Table 2: FTIs per Form in Actual vs. Simulated Data

Observed Average	Simulated Average (2014)	Simulated Average (2015)
0.234	0.251	0.237

A key motivation for modeling the distribution of FTIs is to ensure that FTIs arise as frequently in the simulated as in the actual data among records that tend to be harder to impute – meaning records with multiple FTIs and without prior year data. Tables 3 and 4 compare the tendency of observed and simulated records to have multiple FTIs and to lack prior year data when they have at least one FTI. Table 3 provides the frequency with which observed and simulated records with at least one FTI had between one and eight FTIs. The key result in table 3 is that, among records with at least one FTI, similar percentages have two or more FTIs in both the observed and simulated data, indicating the records with multiple FTIs tend to arise as frequently in the simulated as in actual data.

Table 3: Distribution of FTIs among Records with at least One FTI, Actual vs. Simulated Data

Number of FTIs	Percent of Records (Observed)	Percent of Records (Simulated, 2014)	Percent of Records (Simulated, 2015)
1	74.032	70.284	73.728
2	15.352	21.493	18.906
3	7.174	5.889	5.320
4	3.300	1.733	1.507
5	0.143	0.510	0.459
6	0.000	0.081	0.067
7	0.000	0.009	0.011
8	0.000	0.001	0.001

Table 4 provides the average percentage of FTIs that belong to records with no prior year data in both the simulated and observed data. The percentages are similar in the simulated as compared to the observed data, though the FTIs simulated in the 2015 data do belong to records with no prior year data about 25 percent less often than do observed

FTIs and FTIs simulated in the 2014 data. This discrepancy is not a cause for grave concern given the overall similarity in the distributions of the FTIs across the observed data and both years of simulated data. However, it should be kept in mind as a possible explanation in case imputations made for the 2015 simulated data appear to be systematically more accurate than those made for the 2014 simulated data.

Table 4: Average Percent of FTIs that Belong to Records with no Prior Year Data

Observed FTIs	Simulated FTIs (2014)	Simulated FTIs (2015)
21.4	21.8	15.4

Finally, table 5 presents evidence that there is significant variation among the fields selected as FTIs in each run of the simulation. This table categorizes each of the $8 \times n$ fields in terms of how many times it was selected as an FTI in 5,000 runs of the simulation. With both the 2014 and 2015 data, the vast majority of the $8 \times n$ fields were selected as FTIs in ten percent or fewer of the 5,000 runs (i.e., in 500 or fewer runs). This finding indicates that the different runs of the simulation are distinct from one another, which means that conducting multiple runs of the simulation does in fact, as desired, test how the imputation procedures perform on average under varying circumstances.

Table 5: Frequency with which Each of the $8 \times n$ Fields Was Selected as FTI

Number of Runs Selected as FTI (out of 5,000)	Percent of $8 \times n$ Survey Items (2014)	Percent of $8 \times n$ Survey Items (2015)
0 to 5	13.8	17.1
6 to 50	22.7	27.3
51 to 100	16.7	22.7
101 to 150	15.1	8.2
151 to 201	5.0	11.3
201 to 250	5.1	2.2
251 to 500	16.5	9.0
501+	5.2	2.1

IV. Testing Framework

This section explains the three different versions of the auto-editing system that are tested using the simulation framework laid out above. It also presents the metrics used to evaluate and compare the results of these tests. The three versions of the auto-editing system being tested differ from one another in the approaches they take to donor and estimator imputation. The first version represents the initial, or base, settings for the imputation procedures in that these are the settings that grew out of the early development of auto-editing systems at BEA. The next two versions incrementally introduce innovations to this initial set-up in an attempt to improve the average quality of imputations for each of the eight numeric fields. Two sets of results are presented for each of the three tests, one based on conducting the test with 2014 data and one based on conducting the test with 2015 data.

In the first, or base, version of the auto-editing system, the imputation procedures are run essentially as explained in section two above. Donor imputation is performed on all eight fields and then estimator imputation is performed on all eight fields. Moreover, both imputation procedures use only data from the survey year being processed. That is, in the auto-editing of, say, the 2015 forms, donor imputation is performed using only records from 2015 as potential donors and estimator imputation is performed using only data from 2015 forms to estimate the regression parameters on which imputations are based.

The second version of the auto-editing system tested differs from the first in that donor and estimator imputation are both performed using multiple years of data. In addition to using data from the survey year being processed, both imputation procedures use edited data from the three previous years of surveys.[21] Doing so increases the pool of potential donors in donor imputation by roughly a factor of four, which should, on average, increase the quality of donor-recipient matches and of the resulting imputations. For estimator imputation, the use of multiple years of data means that the regression models used to make imputations are fit using panel data. As a result, company-level fixed effects can be estimated and the estimates of other regression parameters may also be more accurate.

The third version of the auto-editing system being tested builds on the second version by altering the order of estimator and donor imputation for certain fields. In particular, instead of running donor imputation on all eight fields followed by estimator imputation on all eight fields, first estimator imputation is run on assets, liabilities, and gross PP&E, followed by donor imputation on all fields and then estimator imputation on all remaining fields. The decision to investigate the effect of running estimator before donor imputation for these three fields was based partly on empirical and partly on theoretical considerations. Empirically, exploratory tests using the simulation-based framework indicated that estimator imputation may perform better than donor imputation for these three fields. Theoretically, it appears that the poor performance of donor imputation with these three fields is due to a built-in limitation of the algorithm that Banff uses to identify donor-recipient matches. The algorithm identifies matching fields – i.e., fields on which donors and recipients should have matching or nearly matching values – based on edits that involve the field requiring imputation, in particular edits that describe relationships between the field requiring imputation and other fields.[22] However, assets, liabilities, and gross PP&E each has relatively few relationships, as described in the edits, with the other seven fields, which means that Banff is likely to have difficulty identifying appropriate matching fields when one of these fields requires imputation for a given record. As a result, the quality of donor-recipient matches and of resulting imputations may be systematically suboptimal when one of these three fields is being imputed with donor imputation.

Metrics are needed to analyze the average proximity of imputed values to actual values in these tests. The results in the following section are presented in terms of two closely related measures, which are applied separately to each of the eight numeric fields. The first is the percentage difference – averaged over all 5,000 runs of the simulation – between each field's actual aggregate value and its aggregate value after simulated FTIs have been replaced by imputations. This measure, denoted $\bar{y}$, can be represented mathematically as follows,

$$\bar{y}_j = \frac{\sum_{k=1}^{5,000} \left[\left(\left[\frac{\sum_{i=1}^{n} s_{ijk}}{\sum_{i=1}^{n} o_{ij}} \right] - 1 \right) \times 100 \right]}{5,000}$$

where $\sum_{i=1}^{n} s_{ijk}$ is the aggregate value of field j in run k of the simulation (summed over records i=1 to n) and $\sum_{i=1}^{n} o_{ij}$ is the actually observed aggregate value of field j for "clean" BE-15C forms for the year in question.

[21] The test based on 2014 data thus uses data from 2011, 2012, 2013, and 2014 for imputations, while the test based on 2015 data uses 2012, 2013, 2014, and 2015 data.

[22] For example, if, say, field x requires imputation for a given record, edits that indicate its relationship to other fields, such as $x \leq y$ or $x + z \geq y$ are used to identify matching fields. A detailed explanation of the algorithm can be found in Banff Support Team (2012).

While $\bar{y}$ provides a useful measure of how close aggregate estimates are, on average, to their true values, it does not necessarily provide an accurate view of how close individual imputations are to the true values they are replacing.[23] For example, an aggregate estimate will be close to its true value even if half of the imputations systematically overestimate their true values by a wide margin and the other half systematically underestimate their true values by a wide margin, since positive and negative differences will cancel each other out. The second measure avoids this problem and thereby provides a potentially more accurate take on the quality of individual imputations.

The second measure focuses on the absolute value of the differences between imputed and actual values. In particular, it measures the average absolute difference between imputed and true values for each field. To make these differences comparable across the eight different fields, they are then standardized by dividing them by the actually observed aggregate value for each field. The formal expression for this measure, denoted $\bar{x}$, is the following,

$$\bar{x}_j = \frac{\left[\frac{\sum_{k=1}^{5,000}\left[\frac{\sum_{l=1}^{m_{jk}}\left|s_{l(jk)} - o_{l(j)}\right|}{m_{jk}}\right]}{5{,}000}\right]}{\sum_{i=1}^{n} o_{ij}} \times 100$$

where the l=1,…,m_{jk} records with simulated FTIs are nested within the j=1,…,8 fields and k=1,…,5000 runs. The similarity between $\bar{x}$ and $\bar{y}$ can be seen by noting that the equation for $\bar{x}$ becomes equivalent to the equation for $\bar{y}$ if $\left|s_{l(jk)} - o_{l(j)}\right|$ is replaced with $\left(s_{l(jk)} - o_{l(j)}\right)$. Both statistics can thus be interpreted as measuring the average distance between imputed and actual values by field as a percentage of the field's true aggregate value – the difference between the two measures being that one treats distance in absolute terms and one does not. Moreover, while both measures are used in the presentation of results in the subsequent section, greater weight is given to $\bar{x}$. A value of $\bar{y}$ that differs greatly from zero would certainly be a cause for concern, but, for the reasons discussed above, the average absolute difference between imputed and actual values ($\bar{x}$) generally provides a better indication of the overall quality of imputations than does the average (non-absolute) difference ($\bar{y}$).

V. Results

Results are presented for six different tests, each involving 5,000 runs of the simulation-based testing framework. As explained above, three different versions of the imputation procedures were tested, with each version being tested once on 2014 data and once on 2015 data. In addition to the summary measures ($\bar{x}$ and $\bar{y}$) explained in the preceding section, a series of graphs are presented that show how the values of these summary measures are affected by the number of simulation runs. These graphs help to demonstrate that sufficient runs have been conducted for drawing conclusions based on the summary measures. As the number of runs on which they are based increases, the summary measures should come to approximate more and more closely their "true" values. If a sufficient number of runs has been conducted, there should be an identifiable point at which the measures converge to their true values and additional runs no longer have a significant effect on their values.

Overall, the results are satisfactory, as all six tests indicate relatively close agreement between actual and estimated/imputed values. Even the two tests based on the initial version of the imputation procedures, before any innovations were introduced, show that aggregate estimates and individual imputations are both relatively close to their true values. As shown in tables 6 and 7, the average difference between actual and auto-edited aggregate values

[23] The proximity of individual imputations to their true values matters for BEA's purposes because it publishes not only survey-wide aggregates but also sub-aggregates that may be based on only a small number of records.

($\bar{y}$) in the two initial tests was, for seven of the eight fields, less than one percent and between one and two percent for the eighth field (R&D). The average absolute differences between imputed and actual values ($\bar{x}$) were also relatively low – less than one percent of the aggregate for four of the eight fields in both years and less than two percent for six of the eight fields. To be sure, the imputations for net income and R&D do appear to be somewhat less accurate than those for the other six fields, but the average absolute difference between imputed and actual values is still well below ten percent for these two fields.

Table 6: Results Based on 2014 Data

	Initial Test		Second Test		Third Test	
Field	Avg. % Diff. ($\bar{y}$)	Avg. Abs. % Diff. ($\bar{x}$)	Avg. % Diff. ($\bar{y}$)	Avg. Abs. % Diff. ($\bar{x}$)	Avg. % Diff. ($\bar{y}$)	Avg. Abs. % Diff. ($\bar{x}$)
Assets	-0.01	0.06	-0.01	0.05	-0.02	0.03
Liabilities	0.04	0.24	0.03	0.23	0.01	0.18
Sales	-0.02	1.59	-0.02	1.45	0.00	1.44
Net Income	-0.32	4.35	-0.09	4.02	-0.09	3.99
Employee Compensation	-0.28	0.93	-0.44	0.97	-0.45	0.99
Gross PP&E	0.12	1.41	0.15	1.32	0.17	1.12
R&D	-1.04	6.79	-1.24	5.63	-1.17	5.57
Employees	-0.17	0.66	-0.22	0.67	-0.23	0.67

Table 7: Results Based on 2015 Data

	Initial Test		Second Test		Third Test	
Field	Avg. % Diff. ($\bar{y}$)	Avg. Abs. % Diff. ($\bar{x}$)	Avg. % Diff. ($\bar{y}$)	Avg. Abs. % Diff. ($\bar{x}$)	Avg. % Diff. ($\bar{y}$)	Avg. Abs. % Diff. ($\bar{x}$)
Assets	-0.01	0.06	-0.01	0.05	-0.02	0.03
Liabilities	0.01	0.23	-0.01	0.20	-0.02	0.14
Sales	0.09	1.47	0.08	1.31	0.07	1.30
Net Income	-0.04	6.22	0.35	5.84	0.37	5.80
Employee Compensation	-0.08	0.70	-0.23	0.71	-0.24	0.71
Gross PP&E	0.12	1.17	0.26	1.20	0.31	1.02
R&D	-1.81	5.04	-2.05	3.88	-2.19	3.95
Employees	0.01	0.51	-0.03	0.52	-0.03	0.52

Moreover, the plots in figures 1 and 2 indicate that sufficient runs were conducted for the results based on the initial version of the auto-editing system to be trustworthy. These plots show how the value of the average absolute percent difference ($\bar{x}$) evolved for each field as more and more runs were conducted.[24] To be precise, these graphs were created by recalculating the average absolute percent difference between imputed and actual values after every 50 runs. The key takeaway from these plots is that they are all relatively flat, indicating that, for each field, the measure

[24] Plots for $\bar{y}$ are not shown, but they follow the same pattern as those for $\bar{x}$.

is converging to its true value and that the estimates reported in tables 6 and 7 can be considered reflective of their true values.

Figure 1: Initial Test, 2014

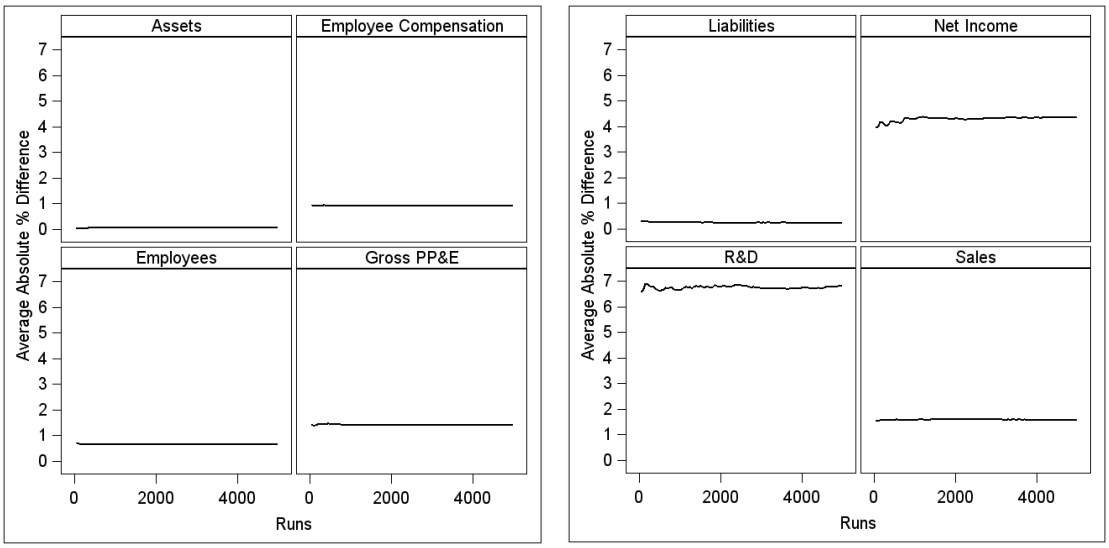
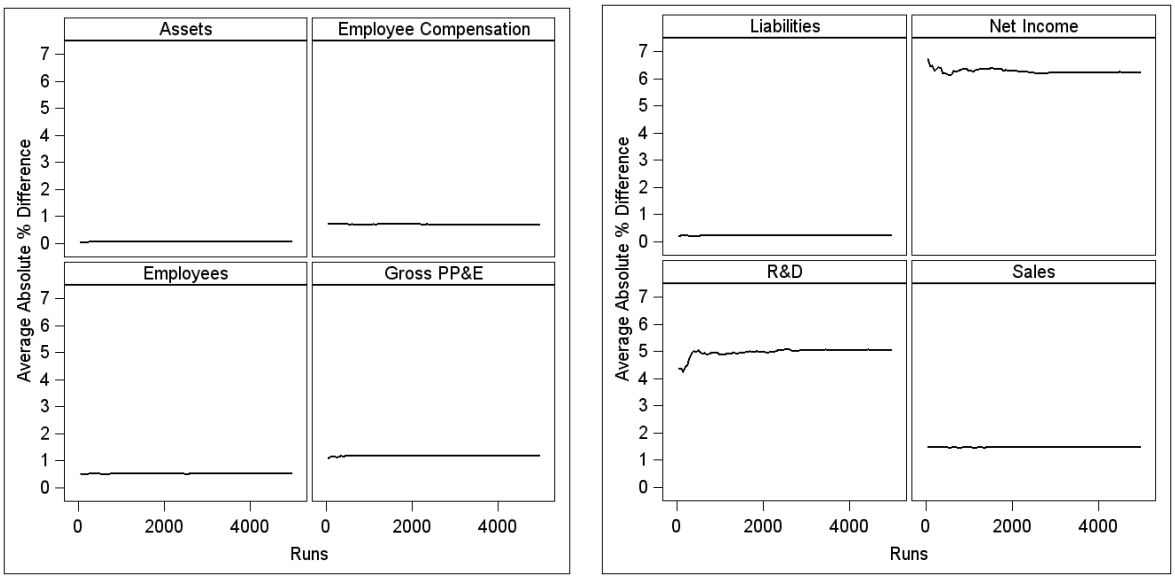


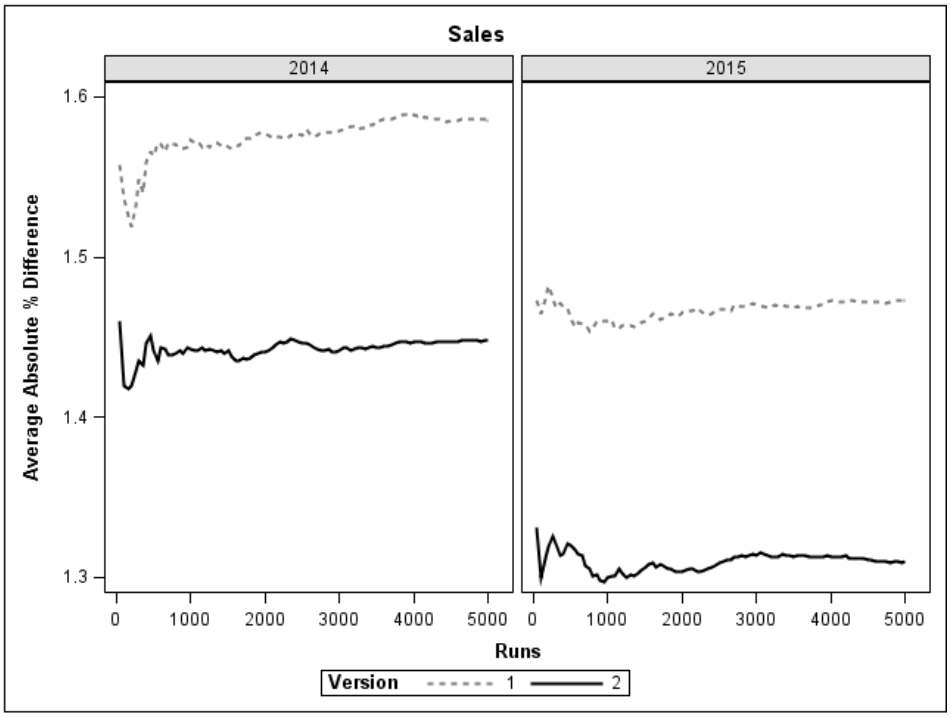
Figure 2: Initial Test, 2015



The next two columns of tables 6 and 7 report the results from the test based on the second version of the auto-editing system, which builds on the initial version by incorporating multiple years of data into both donor and estimator imputation. The results for the eight fields are of three distinct types. First, the use of multiple years of data appears to lead to clear improvements in the quality of imputations for sales, net income, and R&D

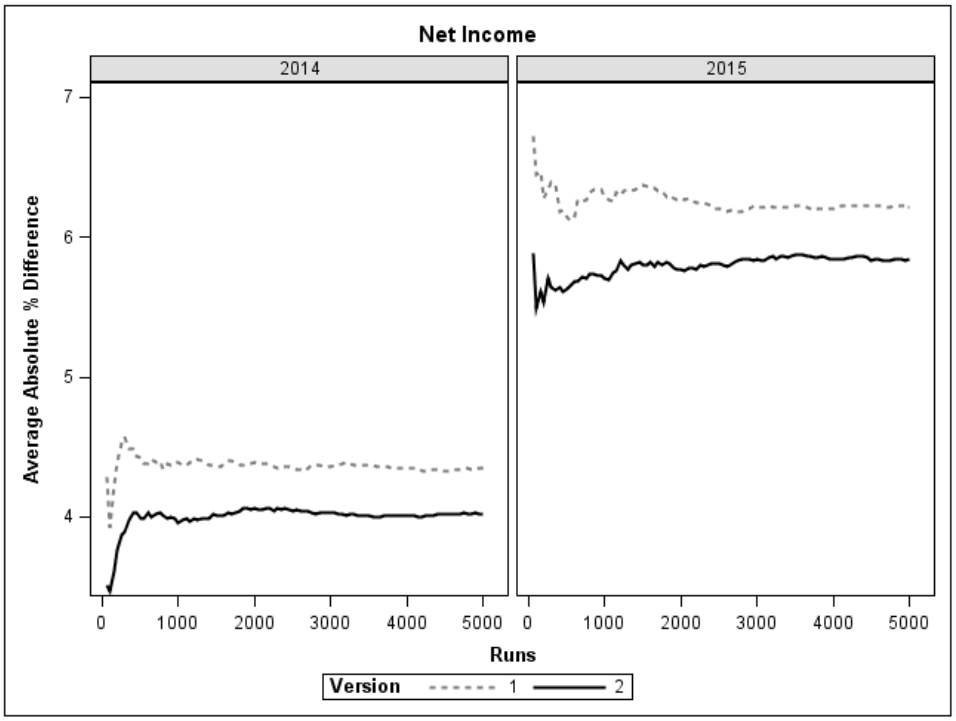
expenditures. The plots in figure 3 compare version one and version two of the imputation procedures in terms of average absolute percent difference for these three fields. For all three, the difference between the two versions appears to be stable and enduring as the number of runs increases. Moreover, Welch's *t* test[25] for difference in means indicates that, for all three fields and with both 2014 and 2015 data, the differences in average absolute percent difference for the two versions of the imputation procedures are significant at less than the α=0.0001 level.[26] For these three fields there are thus clear indications that the use of multiple years of data leads to real improvement in the accuracy of imputations.

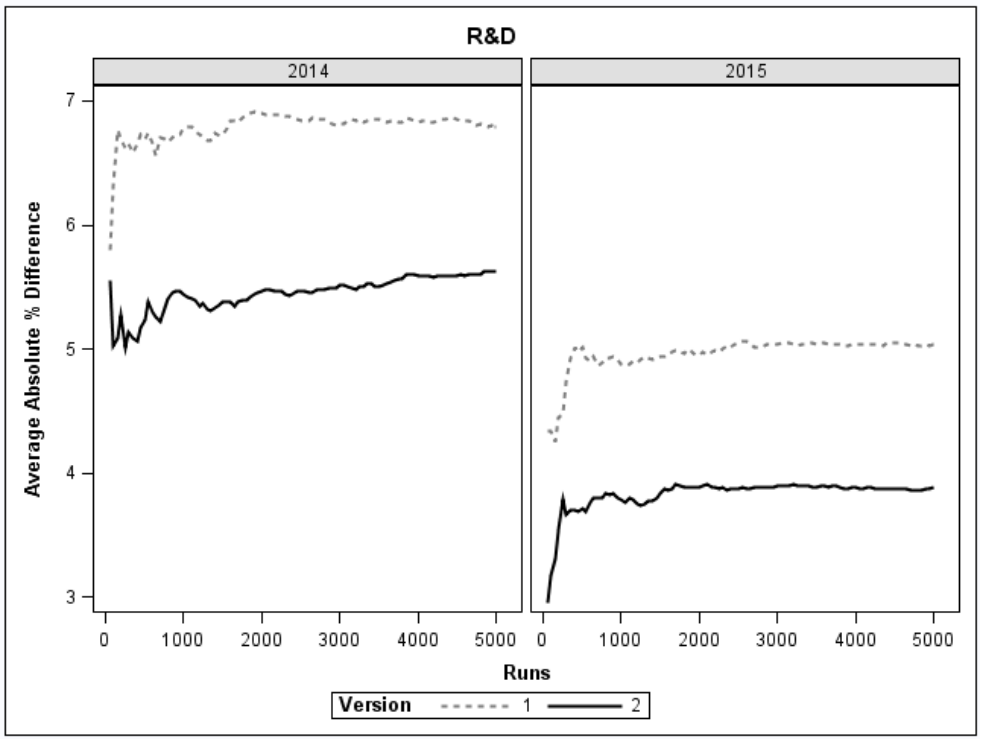
Figure 3: Initial vs. Second Test, 2014 and 2015 Data



[25] Following Rasch et al. (2011), pre-testing for normality was not conducted. On the unimportance of the normality assumption in large datasets, see also Lumley et al. (2002).

[26] The relevant test statistics are as follows. For the 2014 data, Sales (t = 19.04, df = 9,979.2), Net Income (t = 5.73, df = 9,988.4), R&D (t = 10.31, df = 9,909.4). For the 2015 data, Sales (t = 23.05, df = 9,971.6), Net Income (t = 4.75, df = 9,966), R&D (t = 13.90, df = 9,040.6).



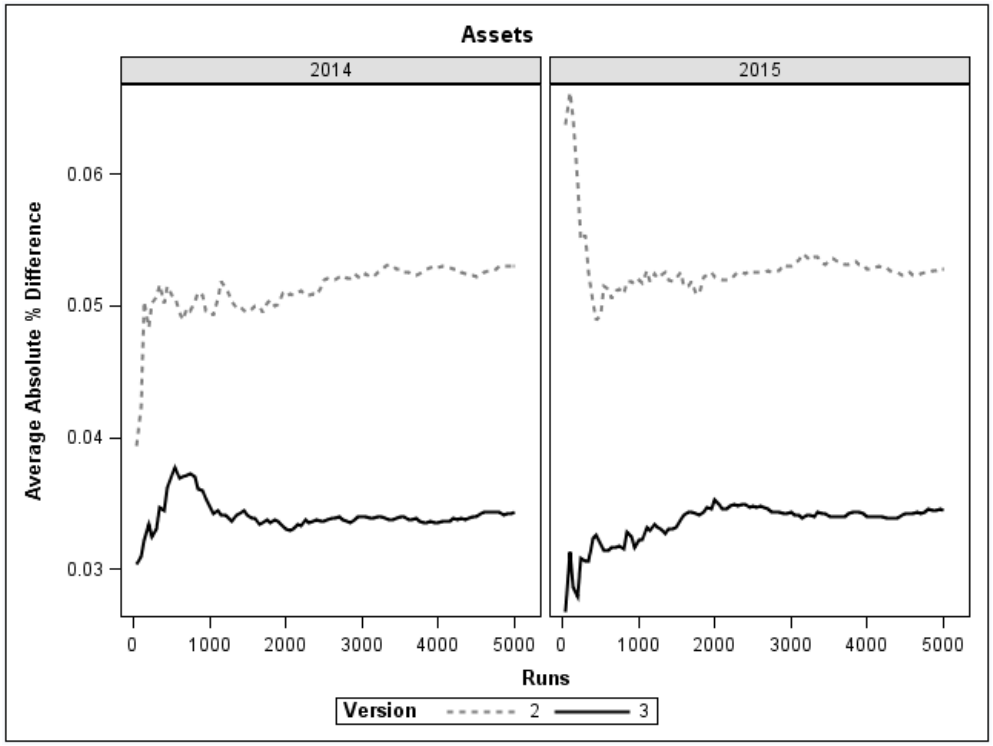


In contrast, employee compensation and number of employees show a small decrease in the accuracy of imputations in 2014 and 2015 as indicated by increases in $\bar{x}$ in tables 6 and 7 between the first and second tests. These increases are slight, and difference-in-means tests indicate that only the increase corresponding to employee compensation in

the test with 2014 data is statistically significant at the α=0.05 level.[27] Nonetheless, these results suggest that it might be advisable not to use multiple years of data when obtaining imputations for employee compensation and number of employees.

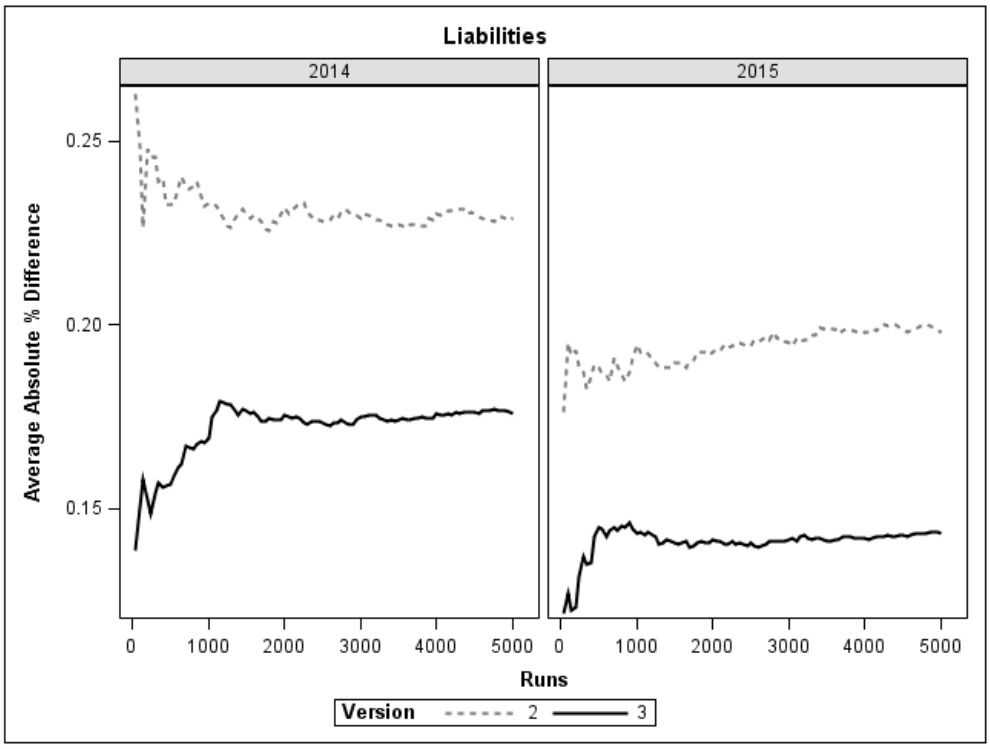
In a third, intermediate category are assets, liabilities, and gross PP&E, as the evidence for improved imputations as a result of using multiple years of data is somewhat more ambiguous for these three fields than it is for sales, net income, and R&D. However, for these three fields, the differences between the results for the second and third versions of the imputation procedures, presented in the last two columns of tables 6 and 7, are not at all ambiguous. In addition to using multiple years of data, the third version of the procedures alters the order of estimator and donor imputation for assets, liabilities, and gross PP&E. As expected, there are indeed reductions in the absolute percent difference between actual and imputed values for these three fields. Moreover, the plots in figure 4 indicate that these reductions are enduring and stable over the 5,000 runs of the simulation. Difference-in-means tests also indicate that, for both 2014 and 2015 data, the differences in $\bar{x}$ for versions two and three of the imputation procedures are, for all three fields, significant at less than the α=0.0001 level.[28] These results thus indicate that altering the order of the imputation procedures for these three fields does lead to improvement in the quality of their imputed values.

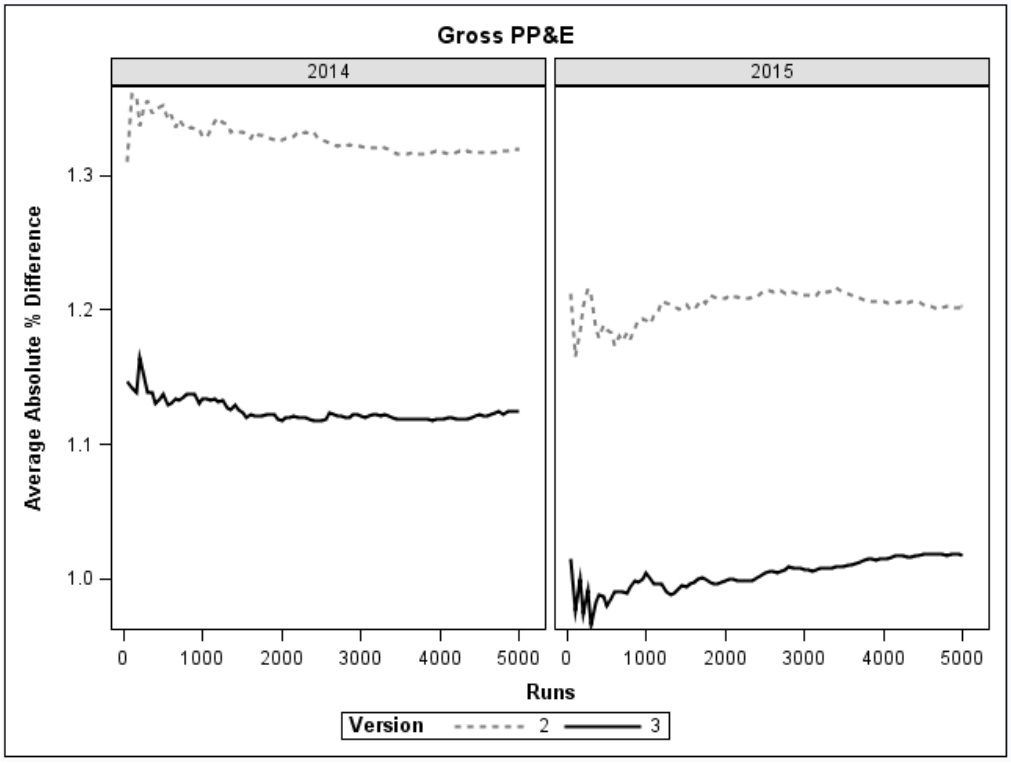
Figure 4: Second vs. Third Test, 2014 and 2015 Data



[27] The relevant test statistics are as follows. For the 2014 data, employees (t = -1.30, df = 9,995.8), employee compensation (t = -5.85, df = 9,986.3). For the 2015 data, employees (t = -1.70, df = 9,990.4), employee compensation (t = -1.70, df = 9,998).

[28] The relevant test statistics are as follows. For the 2014 data, assets (t = 11.86, df = 3,441.3), liabilities (t = 8.85, df = 8,732.3), gross PP&E (t = 20.78, df = 9,595.4). For the 2015 data, assets (t = 11.65, df = 3,187.8), liabilities (t = 11.18, df = 9,340), gross PP&E (t = 20.34, df = 9,728.6).





Taken together, the results in this section show that the innovations introduced in versions two and three of the auto-editing system lead to clear improvements in the accuracy of imputations for six of the eight numeric fields on the

BE-15C survey form. To be sure, the magnitude of $\bar{y}$ sometimes increased with the introduction of the innovations in versions two and three of the auto-editing system. However, these increases were relatively small and are outweighed by the accompanying decreases in $\bar{x}$, which is a superior measure of the overall accuracy of imputations. While two fields, employee compensation and number of employees, did not show improvement in the accuracy of imputations with the introduction of these innovations, the tests still yielded valuable information about the average accuracy of imputations for these fields. Indeed, it is notable that, of all six fields that account for at least two percent on average of all FTIs (see table 1), these two fields have the lowest values of $\bar{x}$ both before and after the innovations in versions two and three of the imputation procedures were introduced.

VI. Conclusion

This paper has proposed a new approach to assessing the quality of imputations obtained through BEA's auto-editing systems. This approach is based on simulating the presence of missing and erroneous data – i.e., simulating FTIs – in survey forms that have neither missing data nor data identified as erroneous by BEA's validity checks. By obtaining imputations for simulated FTIs and comparing these imputed values to their corresponding original values – and repeating this process over and over again with new sets of simulated FTIs – conclusions can be drawn about the average quality of imputations for each field on a survey form.

The results of the tests conducted with this simulation-based framework were highly satisfactory. These tests were conducted on three different versions of the imputation procedures that could be used to auto-edit BEA's BE-15C survey forms. Overall, the tests revealed close agreement between imputed values and the actual values they were replacing, even when the simplest version of the imputation procedures was used. The tests also proved to be useful in comparing the three different versions of the imputation procedures. They provided relatively unambiguous evidence that imputations can be improved, for most fields, by incorporating multiple years of previous survey data into donor and estimator imputation and by reversing the usual order of donor and estimator imputation for assets, liabilities, and gross PP&E.

This testing technique is an important addition to BEA's regime for testing its auto-editing systems. Previously, testing was based on auto-editing survey data that had already been manually edited and then comparing the results of auto-editing to the results of manual editing. While this approach offered valuable insights, it provided only a partial picture in regard to the accuracy of auto-edited data. In particular, it did not provide a means of estimating the average difference between the values of auto-edited survey items and their corresponding true values – a shortcoming for which the current approach has attempted to correct.

Finally, it bears mentioning that a limitation of the present approach to testing the results of auto-editing is that it does not provide a test of the error localization procedure. The quality of the output of the auto-editing system depends both on the accuracy of its imputations and on its ability to correctly identify which field (or fields) to impute when a record fails one or more edits. The approach presented here only examines the quality of the auto-editing system's output under the assumption that the fields in need of imputation have already been correctly identified by the error localization procedure. It thus does not measure the extent to which the auto-editing system might introduce error into aggregate estimates by incorrectly identifying which fields need to be imputed.

Even without measuring the potential for error due to misidentifying FTIs, the assessment of the imputation procedures provided by the present framework is still highly valuable, as indicated by its ability to adjudicate between alternative versions of these procedures. Moreover, the inability of this framework to test error localization does not preclude its testing and refinement through other means. The testing framework proposed in this paper should indeed be seen as a supplement, not a replacement, to other methods of validating and improving BEA's auto-editing systems. For example, in the course of developing the BE-15C auto-editing system, its ability to

correctly identify FTIs was tested by auto-editing data that had previously been manually edited and comparing the auto-editing system's choice of FTIs with fields that had been altered by expert survey editors. These comparisons led to the incorporation of additional edits (i.e., validity rules governing the values that survey items are allowed to take) into the auto-editing system when it was felt that doing so could improve the error localization process.

References

Banff Support Team. 2012. *Functional Description of the Banff System for Edit and Imputation*. Version 2.05. 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.

Chen J. and Shao J. 2000. "Nearest Neighbor Imputation for Survey Data." *Journal of Official Statistics,*16(2): 113-131.

Dasu, Tamraparni and Ji Meng Loh. 2012. "Statistical Distortion: Consequences of Data Cleaning." *Proceedings of the VLDB* [Very Large Databases] *Endowment*, 5(11): 1674-1683.

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.

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.

Granquist, L. and J. G. Kovar. 1997. "Editing of Survey Data: How Much Is Enough?" In *Survey Measurement and Process Quality*, eds. L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, and D. Trewin, New York: Wiley, 415-435.

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

Kosler, J.S. 2012. "Survey Process Control with Significance Editing: Foundations, Perspectives, and Plans for Development." National Agricultural Statistics Service, Joint Statistical Meetings.

Limés, Ricardo. 2017. "Effect of Mode Choice and Respondent Characteristics on Data Quality - Profiling Respondents to BEA's Annual Survey of Foreign Direct Investment in the United States." Paper presented at *Joint Statistical Meetings*, Baltimore, MD, July 29 – August 3.

Lumley, Thomas, Paula Diehr, Scott Emerson, and Lu Chen. 2002. "The Importance of the Normality Assumption in Large Public Health Data Sets." *Annual Review of Public Health* 23: 151-169.

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. Available at: http://www.amstat.org/meetings/ices/2007/proceedings/ICES2007-000135.PDF

Rasch, Dieter, Klaus D. Kubinger, and Karl Moder. 2011. “The Two-Sample *t* Test: Pre-Testing its Assumptions Does not Pay Off.” *Statistical Papers* 52: 219-231.

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.

Sande, G. 1979. “Numerical Edit and Imputation.” Presented at the *42nd International Statistical Institute Meeting*, Manila, Philippines.

Sarndal, Carl-Erik. 1992. “Methods for Estimating the Precision of Survey Estimates when Imputation Has Been Used.” *Survey Methodology* 18(2): 241-252.

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.

UNECE [United Nations Economic Commission for Europe]. 2006. *Statistical Data Editing, Vol. 3: Impact on Data Quality*. New York: United Nations.

Van der Loo, Mark, Jeroen Pannekoek, and Lisanne Rijnveld. 2017. “Computational Estimates of Data-Editing Related Variance.” Paper presented at the *Conference of European Statisticians, Work Session on Statistical Data Editing*, The Hague, Netherlands, April 24-26.

Winkler, W.E. 2006. “Data Quality: Automated Edit/Imputation and Record Linkage.” U.S. Census Bureau, Research Report Series (Statistics #2006-7).

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.