Working Paper.

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Workshop on Statistical Data Editing
(Geneva, Switzerland, 15-17 April 2020)

Evaluating Imputation Methods using ImpACT: First Case Study

Prepared by Darren Gray[1], Statistics Canada

I. Introduction

1. Missing data is a common issue facing National Statistical Organizations, occurring often in the context of a census, survey, or administrative data program. Imputation methods are frequently used to fill in missing values, as part of the data editing process, whose primary goal is the "treatment of the data to achieve fitness for use" (United Nations Economic Commission for Europe, 2019).

2. Designing an imputation strategy can be quite complex. Many different imputation methods exist, and methodologists must choose an appropriate one from amongst them, based on both the data needs and properties of the targeted dataset. Assumptions about imputation models and non-response mechanisms should be validated, if possible. Missingness can occur not just in variables of interest (the primary target of imputation) but also in auxiliary variables used for imputation itself. Ensuring relationships between variables, at both the aggregate and microdata level, poses another challenge.

3. To aid in the evaluation of imputation strategies, we introduce ImpACT: the Imputation Assessment and Comparison Tool. The goal of developing ImpACT is to provide survey methodologists with an easy-to-use, generalized tool to test and compare imputation methods in a controlled, simulation environment. We envision two primary uses of ImpACT at Statistics Canada:

(a) To test imputation methods, including new and emerging methods, on known data
(b) To aid in the design and evaluation of the imputation process for a production environment

The underlying framework of ImpACT was presented at the American Statistical Association Joint Statistical Meeting (Gray, 2019), along with examples of the possible applications, using publicly available data.

4. In this paper, we present our first case study, reviewing the imputation strategy used by Statistics Canada's Retail Commodity Survey (RCS) (Statistics Canada, 2020). The focus of this paper is not on the results of that review, but on the application and performance of the tool itself. We present some examples demonstrating the effectiveness of ImpACT in exposing the potential strengths and weaknesses of an imputation design. We also spend some time discussing new challenges that emerged when dealing with survey data, and how the tool can be improved going forward.

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

5. One of our objectives in developing ImpACT as a generalized tool is to design an assessment tool that can be used effectively without extensive knowledge of the target dataset and/or imputation methods being assessed. To that extent, we treat the data and imputation methods as unknowns, as much as possible. However, some guidance from the RCS team was required to properly set up the simulation, and to provide relevant background information to the reader.

6. We discuss the RCS imputation strategy in Section II, and give a brief overall of ImpACT in Section III. Section IV highlights some of the challenges faced in setting up an appropriate simulation study. Sections V and VI highlight two of the main studies performed using ImpACT. We end with a discussion of our findings and planned future work in Section VII.

II. Retail Commodity Survey

7. The RCS "collects detailed information about retail commodity sales in Canada" on a monthly basis, as a complement to the Monthly Retail Trade Survey. Sales estimates generated by the RCS are broken down by commodity, using the North American Product Classification System (NAPCS) (Statistics Canada, 2019), along with other classifiers such as industry and geography.

8. Within NAPCS, commodities are classified into mutually exclusive categories at four different levels:

Table 1: NAPCS category levels

Level	Coding	Number of categories
Group	3-digit code	157
Class	5-digit code	506
Subclass	6-digit code	1,409
Detail	7-digit code	2,737

The classification is hierarchical, with each category at a lower level contributing in aggregate to a category at the next higher level. To ensure that published estimates satisfy this additive constraint, a primary objective of the RCS data editing strategy (consisting of both imputing missing values and adjusting non-missing ones) is to ensure that all records individually satisfy these constraints at the microdata level, over all commodities.

9. To achieve this goal, the RCS imputation strategy consists of a multi-stage data editing approach with the following core steps:

(a) **Deductive imputation**[2]: missing values are imputed if only one possible value exists that can satisfy the additive constraints, given other, non-missing values.
(b) **Historical imputation**: missing values are imputed on a record-by-record basis based on the same variable collected in a previous iteration of the survey, adjusted in ratio to an auxiliary variable present in both timeframes. There are two attempts at historical imputation, using data from the previous year (first attempt) and previous month (second attempt).
(c) **Donor imputation**: within imputation classes based on geography, industry, and economic activity, missing values are imputed using nearest-neighbour donor imputation.
(d) **Ratio imputation**: within imputation classes, missing values are imputed based on the ratio (in aggregate) of non-missing values with respect to an auxiliary value. There are two attempts at ratio imputation, based on two different sets of imputation classes.

[2] Sometimes referred to as *deterministic imputation*.

(e) **Percentage imputation**: finally, any values still not imputed are set to a pre-determined percentage of total sales, set by the survey team.
(f) **Pro-rating**: imputed values are pro-rated such that all records satisfy the NAPCS hierarchical sum-of-parts constraint.

10. This multi-stage strategy highlights the complexity typical of a Statistics Canada business survey data editing process. For many surveys, multiple attempts are often required to impute a missing data point, for a variety of reasons. Logical imputation is only successful under a very specific set of conditions. Historical and ratio imputation rely on the presence of historical and/or auxiliary data that may not be available for a particular record. Some methods may allow users to enforce post-imputation constraints; if these constraints cannot be met, imputation is not performed. Imputation within imputation classes may be subject to quality thresholds: for example, the generalized edit and imputation system Banff (Statistics Canada, 2017) allows users to restrict imputation to those classes that have a sufficient number or percentage of valid records.

11. Maintaining relationships between imputed variables poses another challenge. These relationships may be explicit, such as the additive relationship amongst commodities described above, or statistical, such as the correlation between two variables. Multivariate imputation methods, such as Banff's donor imputation, can be used to project these relationships from valid to missing data. Some methods allow for constrained imputation. In other cases, as in the RCS strategy, tools such as pro-rating are useful.

12. Both issues introduced unique challenges to our simulation.

III. Overview of ImpACT

13. ImpACT is a prototype tool developed in SAS as a follow-up to other simulation-based tools built at Statistics Canada by Haziza (2003) and Stelmack (2018). It consists of three modules: non-response, imputation, and analysis.

14. To begin, users provide an input dataset and identify a variable of interest y. Only those records for which y is non-missing are included in the assessment; we refer to them collectively as the training set.

15. Users must next generate Monte Carlo non-response trials. They may use ImpACT's non-response module to generate a non-response pattern in each trial, or provide a non-response pattern of their choosing. ImpACT offers a variety of choices to generate the non-response pattern. At the most basic, users can simulate *Missing Completely at Random* (MCAR), *Missing at Random* (MAR), or *Missing not-at-Random* (MNAR) non-response mechanisms by assigning a non-response probability to each unit in the training set, and selecting an independent Poisson sample in each trial. ImpACT offers additional methods to control both how many units are selected for non-response in each trial, and how many times each unit is selected over all trials, allowing for tailored analysis.

16. Once the non-response pattern is generated for each trial, users are responsible for imputing the resulting datasets. Each trial dataset should be imputed independently.

17. The analysis module includes three analysis types, based on a subset of the performance measures for imputation proposed by Chambers (2001):

(a) **Distributional Accuracy:** The imputation procedure should preserve the distribution of the true data values. That is, marginal and higher order distributions of the imputed data values should be essentially the same as the corresponding distributions of the true values.
(b) **Estimation Accuracy**: The imputation procedures should reproduce the lower order moments of the distributions of the true values. In particular, it should lead to unbiased and efficient

inferences for parameters of the distribution of the true values (given that these true values are unavailable).

(c) **Predictive Accuracy**: The imputation procedure should maximise preservation of true values. That is, it should result in imputed values that are 'close' as possible to the true values.

Combining non-response patterns with specific analysis types can lead to a variety of useful insights.

IV. Simulation Plan

18. In a multi-stage imputation strategy, the choice of sub-methods, parameters, imputation classes, and ordering of methods all affect the outcome. In this case study we attempt to assess the overall imputation strategy but also the individual sub-methods, via the following simulation studies:

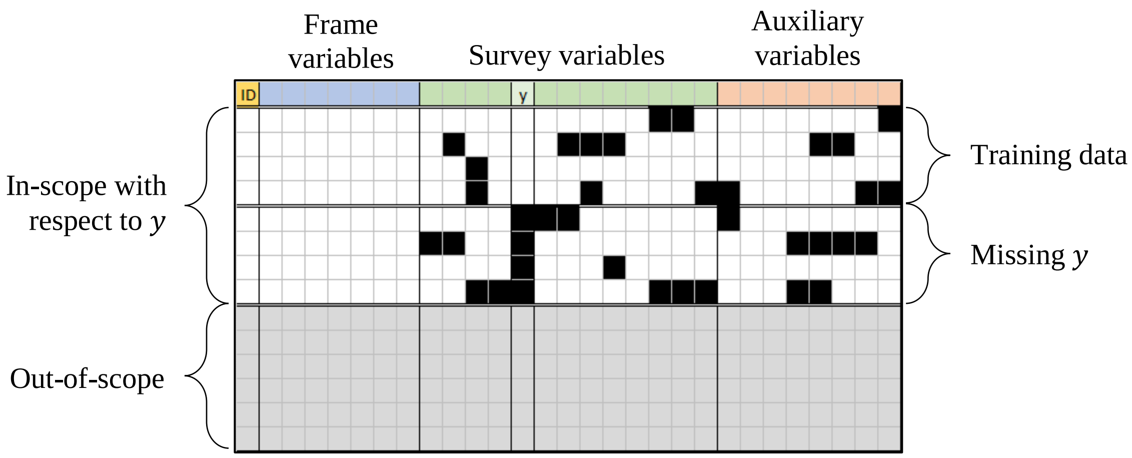
(a) A **general assessment**, designed to assess the performance of the multi-stage imputation strategy under a set of simulated MCAR non-response mechanisms at different rates of non-response.

(b) A **sub-method comparison**, using a leave-one-out approach, to assess the individual methods comprising the multi-stage approach in isolation.

19. To begin, the RCS team identified a variable of interest, representing sales of a commodity at the most detailed level of categorization. For the purposes of data confidentiality, we leave the variable unidentified and simply refer to it as y.

20. The presence of missingness in our input dataset, alongside the multivariate nature of the RCS imputation strategy, required special consideration. To help illustrate why, we provide the following schematic (Figure 1) of the input dataset representing RCS survey data, with missing data indicated in black:

Figure 1: Schematic of RCS data



21. We began by removing records that were out-of-scope with respect to y, based on available frame information. These records do not contribute to any estimates involving y and are not imputed. (Note that the survey distinguishes between records that are out-of-scope with respect to the sale of a commodity, and those that are in-scope but report sales of zero.)

22. The remaining records were split into those for which y was missing or non-missing. The latter formed our training set: the records upon which we simulate non-response, perform imputation, and conduct analysis. The other records – in-scope records for which y is missing – were included in the imputation process, but excluded from analysis. This was done so as to mimic the production

environment as much as possible; in particular, we wanted to maintain the size of existing imputation classes in our simulation.

23. The multivariate nature of the existing strategy, driven by the donor imputation stage, also introduced some challenges in our simulation. The RCS imputation process is designed to impute all missing data in the input dataset, over hundreds of variables, and preliminary tests showed that running this complete process over many trials would not be feasible. Unfortunately, isolating the imputation process with respect to y was not a trivial task, as all variables were linked by the NAPCS additive structure. Our tests showed that the imputation of y depended both on the missing pattern of other commodity variables, and on the value of other variables already imputed. In order to run our simulation, we modified the donor imputation sub-method in the following way:

(a) Imputation was restricted to y, the associated sub-class commodity (the next highest NAPCS categorization to which y contributes) and all other variables contributing to the same sub-class.
(b) We removed all post-imputation constraints not related to these variables.

24. In addition to identifying the training set and target population, we needed to identify which variables would undergo simulated non-response. On the RCS dataset, we identified three main variable types:

(a) **Frame variables**, which included the unique identifier (ID) and variables used to construct the imputation classes.
(b) **Survey variables**, collected from respondents, consisting mostly of commodity sales numbers, including y.
(c) **Auxiliary variables,** including historical data, used for imputation.

Non-response occurred in both the survey and auxiliary variables, but not in the frame variables.

25. An early test showed that if non-response was limited to y, then it was often imputed in the first stage, by deductive imputation. (This did not align with what we witnessed in the actual imputation process.) An analysis of the non-response pattern revealed that when y was missing, so were most other variables that contributed to the same sub-class. To mimic this pattern, we decided to simulate non-response in all these variables (including the associated sub-class).

26. Finally, we removed deductive imputation and pro-rating from our study, as both steps were rendered obsolete due to our simulation choices. The modified strategy, arranged by imputation stage, is given in the table below.

Table 2: Simulated imputation process by stage

Imputation Stage	Sub-method
S1	Historical imputation, using data from previous year
S2	Historical imputation, using data from previous month
S3	Donor imputation (modified)
S4	Ratio imputation, within first set of imputation classes
S5	Ratio imputation, within second set of imputation classes
S6	Percentage imputation

V. General Assessment

27. For this study, we conducted three simulations of 500 trials each. Each trial was designed to represent the outcome of a MCAR non-response mechanism, with non-response rates in each simulation of 10% (simulation 1), 30% (simulation 2) and 50% (simulation 3).

28. For analysis purposes, we fixed the number of non-respondents in each trial. As a result, the set of non-respondents in each trial can be viewed as a simple random sample of the training set. We also coordinated the samples such that all records were selected the same number of times, within a difference of one. Coordinating trials in this way removes sample size variance (which we'd normally see in a Poisson or Bernoulli sample), and ensures that all respondents are equally represented in our analysis. We note that all possible samples of the specified fixed size are equally likely of being selected, despite this coordination.

29. After imputation, we performed distribution and estimation accuracy tests. These tests are designed to investigate how well the imputation method does at reproducing properties of the original data. For confidentiality purposes, values of y are only displayed on a transformed log scale, and no numbers are included in any axes.

A. Initial analysis

30. We began by comparing the distribution of missing values (i.e., original values set to missing in simulation) against imputed values, over all trials. The comparison for each simulation is shown in Figures 2 through 4, with missing values in grey and imputed values in blue. ImpACT offers a variety of visual comparison tools for this purpose; in this example we've included a kernel distribution, jitter plot, box plot, and the mean value. We note that the kernel distribution and box plot are generated from the transformed data (on a log scale) while the means are plotted on this same scale but calculated from the original data.

31. Each comparison type (kernel distribution, jitter plot, box plot, and mean) provides slightly different information, and together can be very informative. We begin by noting that the outputs look very similar over all three simulations, implying that the accuracy of the imputation method (as tested) is not particularly sensitive to non-response rate. Because of this similarity, we discuss the observed properties without distinguishing between simulations.

32. The kernel distributions show nice overall agreement between the original and imputed values, with a few discrepancies worth noting. First, it appears that the imputation method generates fewer values in the tails than it should. Second, the curve takes on a bimodal distribution, and the method shifts the lower mode slightly upwards.

Figure 2: Distribution accuracy, Simulation 1

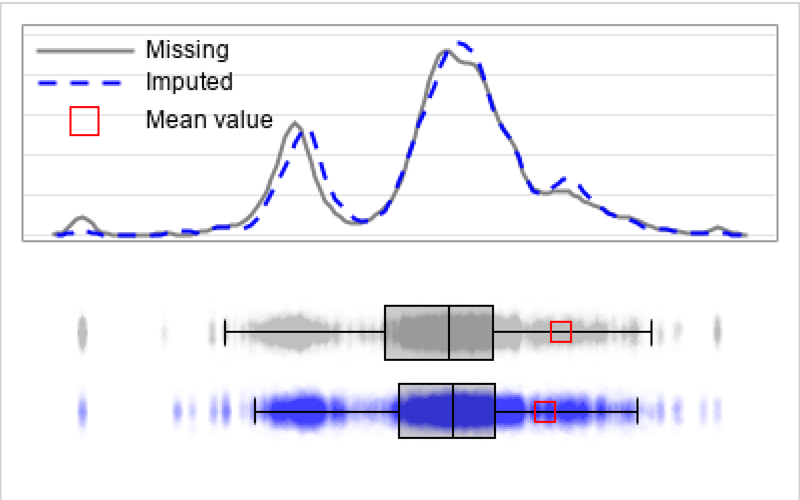


Figure 3: Distribution accuracy, Simulation 2

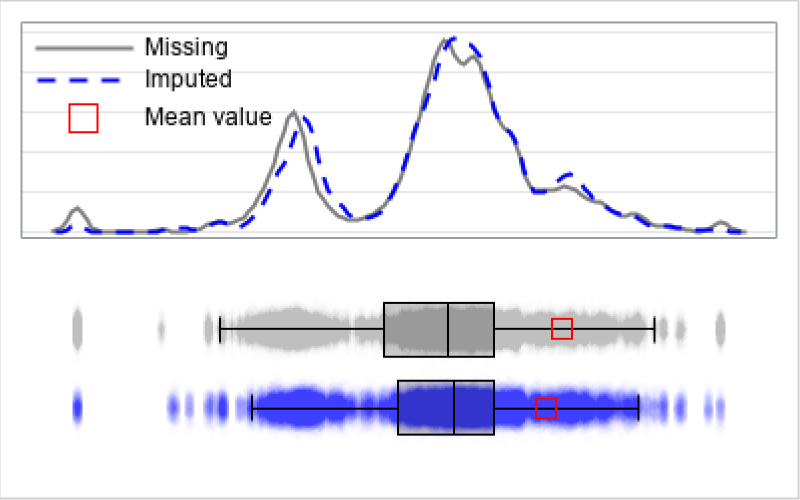
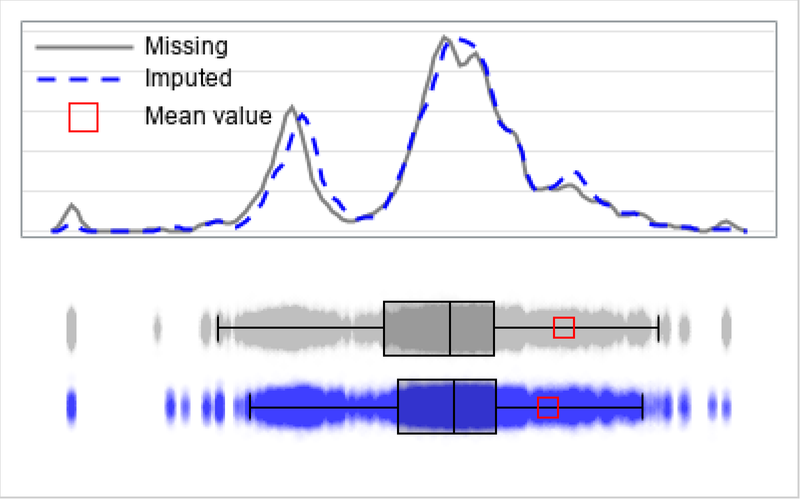


Figure 4: Distribution accuracy, Simulation 3



33. The jitter plot is a plot of univariate values, randomly shifted in the vertical plane to provide some visual spacing. It tells a similar story to the kernel distribution comparison, but here we can see where original values are disappearing, and new ones are being introduced. The overlaid boxplot provides a nice summary; we can see the slight reduction in the spread of the data, in particular at the lower end; this aligns with what we noticed from the other visual comparisons.

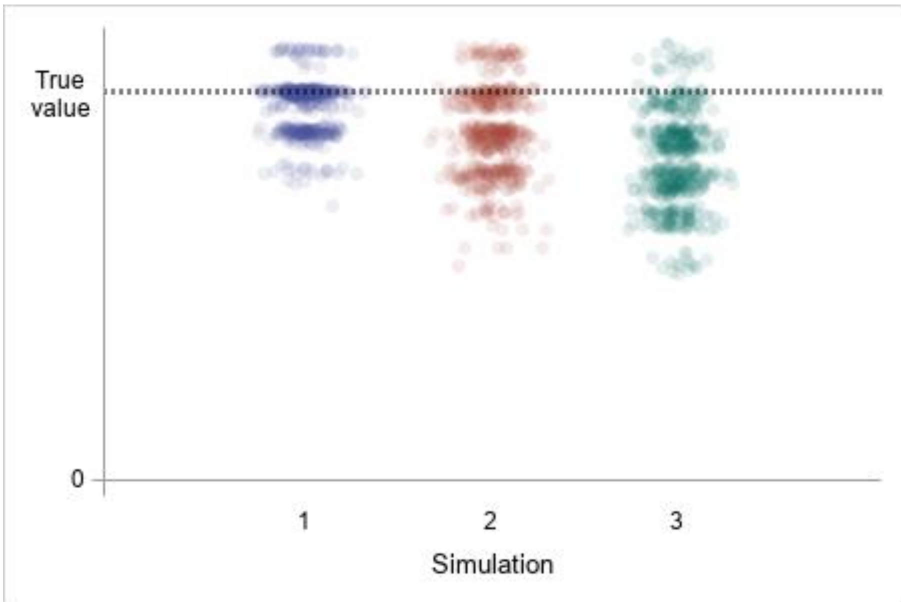
34. Because the RCS survey team intends to produce estimates of total sales, we also plotted the mean value of all data over all trials. We immediately note the location of the mean with respect to the boxplot, indicating the skewed nature of the data. Comparing the means of missing and imputed values, we also note a negative bias over all simulations, introduced during imputation. Note that because the data is plotted on a log scale, this bias could be significantly larger than it appears in the figures.

35. The distribution analysis is effective at showing how an imputation method performs in aggregate, taken over all trials. However, it does not show the variance between trials; it's entirely possible for an imputation method to perform well asymptotically, but to produce poor results in individual instances. One way to investigate the variance from trial to trial is to inspect the results with respect to an estimator of the user's choosing.

36. Figure 5 compares the mean of y generated in each trial to the mean of y taken from the original values in the training set. Each dot represents an estimate generated from one trial, while the horizontal line represents the true parameter. We've included the origin for relative comparison purposes. The

horizontal axis has no meaning; the trials are grouped by simulation and artificially jittered to create visual space. From these outputs, we can analyse the variance and bias introduced by non-response and imputation, within our simulation.

Figure 5: True value vs. trial estimates of population mean, by simulation



37. This output gives us a better sense of the negative bias we noticed in our distribution analysis (now plotted on a relative scale) and how the effect increases with non-response rate. Also noticeable is the apparent clustering of individual estimates: there appears to be four clusters in simulation 1 (10% non-response) and six in simulation 3 (50% non-response). This type of pattern often indicates an extreme imputation error for one or more individual values; the clustering occurs depending on the number of such errors that occur in an individual trial.

38. Investigation of imputation errors of individual units (not shown here) confirmed that a small number of outliers with large errors were highly influential in our results. In all cases, these imputation errors were generated by historical imputation.

B. Secondary analysis

39. To determine the impact of these outliers on our results, we ran the simulation a second time, having removed the offending units from the training set. The results (Figure 6) show drastically reduced bias and variance over all simulations, with respect to population mean.

40. Along with estimates of sales, the RCS also publishes quality estimates that take into account estimated sampling variance. To assess the potential impact of non-response and imputation on sampling variance is outside the scope of ImpACT's current functionality; however, we can compare the standard deviation of y in each trial to the standard deviation on the training set's original values (Figure 7).

41. While this doesn't take into account sample design, it does show that the tested imputation strategy does, after the removal of outliers, maintain the approximate spread of original data, which bodes well for estimates of sampling variance.

Figure 6: True value vs. trial estimates of population mean, by simulation (outliers removed)

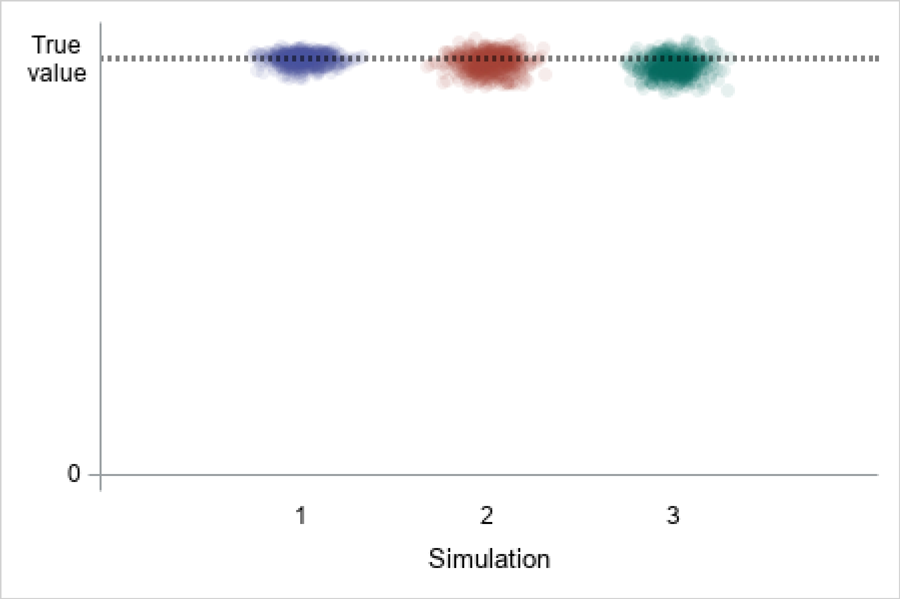
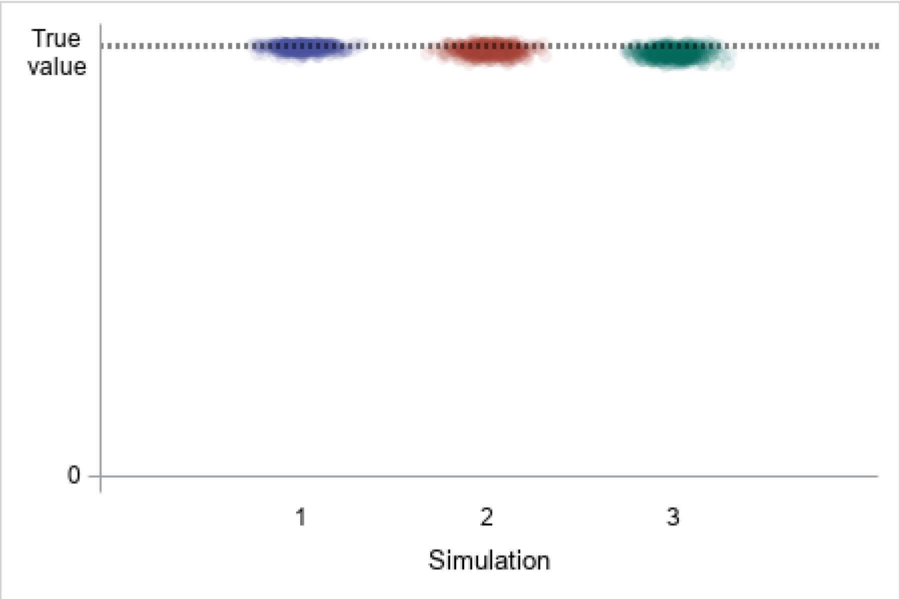


Figure 7: True value vs. trial estimates of population standard deviation, by simulation (outliers removed)



C. Inferential limitations

42. This result is based on a controlled simulation, using known data. In general, inferences about the true performance of an imputation method using ImpACT are limited, for two significant reasons:

(a) The training set data consists of survey data *after* non-response has occurred, and may not be representative of the true population data.
(b) The simulation is based on a simulated non-response mechanism, which may not be representative of the true non-response mechanism.

43. It is possible to investigate these properties to some extent: for example, an analysis of auxiliary data might show where the training data and population data differ, or give an indication as to the type of

non-response mechanism. In fact, simply running an ImpACT test can in some cases indicate a non-representative training set or non-response mechanism. In Table 3, the middle three columns display the imputation rate by stage and simulation, on the training set. The final column shows the same imputation rates, when the process was run on the population records not in the training set, i.e., the records that actually require imputation. The discrepancy is quite apparent: over 98% of all simulated missing values were successfully imputed by one of the historical imputation methods, while in reality less than 20% were imputed by one of these methods.

44. Because the imputation method is identical over all columns, we can conclude that the discrepancy must arise from differences in the training set and remaining in-scope records. An investigation of the original data revealed a significant difference in the rate of missingness in historical data between these two sets. This could indicate that the training set is not representative of the larger population, or that our simulated non-response mechanism should have been applied to historical data as well. We leave that possibility for a future study.

Table 3: Imputation rate comparison

Imputation stage	Imputation rate (%) by simulation			Actual rate (%)
	1	2	3	
S1	72.7	72.7	72.7	16.4
S2	25.5	25.5	25.5	3.0
S3	0.6	0.4	0.1	32.0
S4	0.4	0.6	0.9	31.6
S5	0.6	0.6	0.6	5.7
S6	0.3	0.3	0.3	11.3

45. We also note that while our simulation proved very sensitive to influential outliers, business surveys involving skewed data often include steps (such as targeted follow-up) to identify and address potentially influential units that could lower the rate of non-response or imputation error in these units. These actions would not be reflected in an ImpACT simulation. The pro-rating step, also excluded from the simulation, could further affect results.

VI. Comparing sub-methods

46. In the second part of our investigation, we decided to test each imputation sub-method separately. The objective of this study was to assess these methods individually, and to compare them, to see if a re-ordering of the overall strategy would increase overall imputation accuracy.

47. For this paper, we present a predictive analysis, under a leave-one-out scheme. Unlike our simulations that are meant to mimic a real-world non-response mechanism, the leave-one-out approach systematically assigns non-response to exactly one record in each trial. As a result, the number of trials is identical to the number of records in the training set.

48. The leave-one-out approach offers a number of benefits with respect to predictive analysis. Under the leave-one-out scheme, the non-response pattern within each trial differs from the original non-response pattern in only one record – it is as close to the original pattern as possible under the limitations of the ImpACT framework. With regards to predictive accuracy for an individual unit, we can view this as a best-case scenario given the input dataset: when all information (except for the record in question's value) is available.

49. Another feature of leave-one-out analysis is that the imputation errors (the difference between imputed and original values) associated with each unit should be good approximations of the underlying model residuals. (This assumes the imputation model is robust with respect to the removal of a single

data point.) This approach allows ImpACT users to perform residual analysis even when the underlying model is unknown.

A. Overall comparison

50. After running imputation on each trial, we plotted the imputation errors against the original values of y, on a log scale (Figure 8). Perfect imputation (an error of zero) is plotted as a horizontal line.

Figure 8: Imputation error, sub-method comparison

51. In general, a good imputation method should produce small errors randomly distributed around zero. Inspecting the results, we note the following:

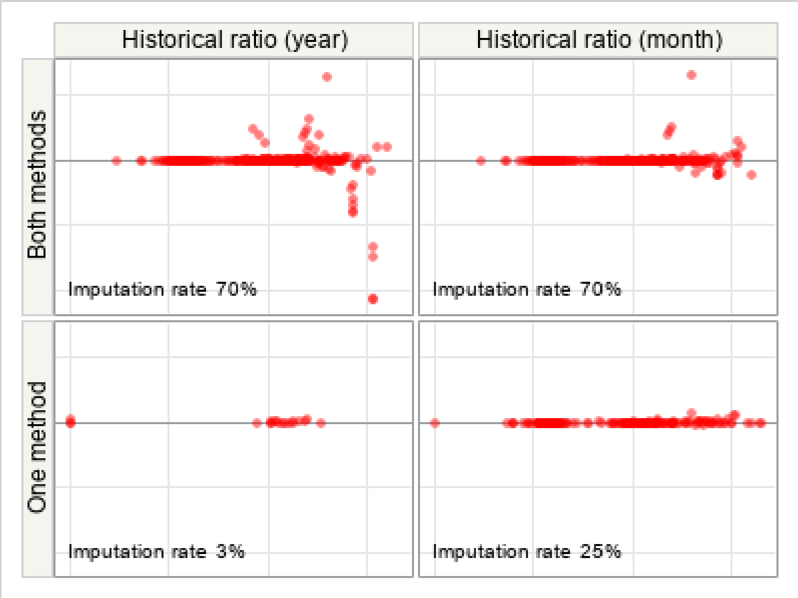
(a) Errors tend to increase in magnitude as y increases; this is in general expected for imputation methods in skewed data.
(b) As y increases, some of the methods show a trend towards negative errors. This trend is significant in the percentage imputation method but also more visibly apparent in donor imputation.
(c) The percentage imputation method is the only method that raises some red flags: in particular the negative trend in errors as y increases, and what appears to be a positive systematic imputation error for some lower values.

This chart gives an overall comparison of the predictive accuracy of each individual method. However, it is not perfect for comparing methods, as they are not all being compared on the same set of data. As discussed in Section II, not every method is capable of imputing all missing values; we've overlaid each sub-method's respective imputation rate on the chart. To directly compare two methods, a head-to-head comparison is required.

B. Head-to-head comparison

52. For a head-to-head comparison of two methods, we divide the simulation data into values that were successfully imputed by both methods, and those that were only imputed by one of the two methods. (This categorization is dependent on the two methods being compared.) For example, we compare the two historical methods (using previous month and previous year's data) in Figure 9.

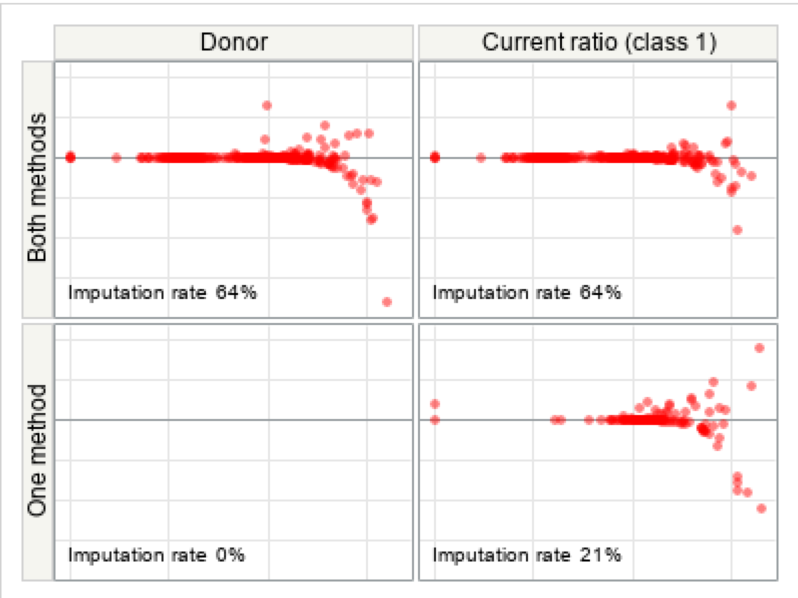
Figure 9: Head-to-head comparison, historical imputation



53. We note that 70% of the missing records were successfully imputed by both methods, and that on this subset, imputation using data from the previous month performed slightly better. Errors were quite small for records imputed by only one method. Based on the figure above, it appears that reversing the order of these two methods in the imputation strategy would reduce imputation error.

54. In the original chart, it also appears that donor imputation performs at least as well as the first run of ratio imputation. However, we note the relatively low (64%) imputation rate of donor imputation. Again, fair comparison of these two methods can only be run using a head-to-head analysis, as shown in the Figure 10. This result shows that ratio imputation outperforms donor imputation when taken over records imputed by both methods.

Figure 10: Head-to-head comparison, donor vs ratio imputation



C. Assessment

55. Under the leave-one-out simulation, we did not see any major concerns in the individual methods, other than the percentage imputation method, which we would suggest revisiting or replacing.

56. Errors were smallest using historical imputation methods, suggesting that the overall strategy used by the RCS is correct to employ them first. Results suggest it might be beneficial switching the order of these two sub-methods, and running historical imputation using previous month's data first. (It is also possible that this result is specific to the month we analysed, and that previous year's data might be a better predictor overall.) We also identified some re-orderings in subsequent methods that might improve results, but more testing is required.

VII. Conclusions and future work

57. This case study proved to be a good testing ground for evaluating the effectiveness of ImpACT for assessing survey imputation strategies. We ran two studies, each of which provided insight into the tested imputation strategies, and which we hope will prove useful to the RCS survey team for future iterations of the survey. While there are limitations as to inferences one can make from a simulation study applied to an individual reference period, we believe that these insights are still useful in understanding the overall performance of the tested methods, and for identifying areas for further investigation.

58. We encountered challenges running ImpACT on survey data that we had not faced in previous tests. Missingness in the variable of interest y was to be expected, but missingness in auxiliary variables was new, and limited the conclusions we could make about the overall imputation strategy. Investigating the pattern of missing data over all records and all variables would prove useful for future studies.

59. Another challenge that required careful consideration was how to assess the performance of a multivariate imputation strategy on a single variable of interest, in a simulation framework. This required modification of the imputation strategy, and careful thought as to which variables should be included in the non-response simulation. The lessons learned should help simplify this process in future projects.

60. A key objective of this case study was to determine the effectiveness of ImpACT as a generalized tool; in particular what one could learn without an in-depth knowledge of either the methods being tested or the target dataset. This aspect of the project was a success, as no knowledge of the methods tested was required to run our analysis – they could effectively be a set of black boxes. However, an understanding of the underlying nature of each method was required in order to set up informative tests, in particular with respect to donor imputation.

61. Finally, a comment on our choice of visual over numerical outputs: while many survey methodologists and statisticians are comfortable with numerical measures such as Monte Carlo bias and mean squared error, we feel that the concepts these numerical measures aim to capture can be sufficiently – and in most cases more efficiently – represented in visual medium. When properly designed, data visualizations can be effective at highlighting important information, and can capture patterns that numerical outputs might not. While ImpACT is capable of outputting numerical measures, the focus going forward is on data visualization outputs for analysis purposes.

References

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

Gray, D. (2019). A Generalized Framework to Evaluate Imputation Strategies: Recent Developments. *In JSM Proceedings, Government Statistics Section. Alexandria, VA: American Statistical Association. 1861-1870.*

Statistics Canada (2020). Retail Commodity Survey. https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=2008

Statistics Canada (2019). North American Product Classification System (NAPCS) Canada 2017 Version 2.0. https://www.statcan.gc.ca/eng/subjects/standard/napcs/2017v2/index

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

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

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

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