

## Proceedings of Statistics Canada Symposium 2024: The Future of Official Statistics

### Data-driven Imputation Strategies and their Associated Quality Indicators in Economic Surveys

by Matei Mireuta, Ahalya Sivathayalan and Stephen Styles

Release date: September 8, 2025



Statistics  
Canada Statistique  
Canada

Canada

# Data-driven Imputation Strategies and their Associated Quality Indicators in Economic Surveys

Matei Mireuta, Ahalya Sivathayalan and Stephen Styles<sup>1</sup>

## Abstract

The use of modern “data”-driven imputation methods to treat non-response in the context of surveys processed in the Integrated Business Statistics Program at Statistics Canada have previously been explored. It has been observed that these methods can lead to high quality imputation and further have the potential to result in broad efficiencies when setting up a particular survey’s edit and imputation strategy. However, estimation of the associated total variance, more specifically the component due to imputation, remains a challenge. In this article, two methods for estimation of total variance are proposed and preliminary results are shown that gave motivation to pursue further research in this area.

Key Words: Cubist imputation; Random forest; Variance due to imputation; IBSP.

## 1. Introduction

First deployed in 2013 on around 40 surveys, the Integrated Business Statistics Program (IBSP) (Statistics Canada, 2015) now provides a standardized framework for approximately 130 economic surveys on topics as diverse as agriculture, manufacturing industries and services. The surveys now processed in IBSP also span a wide range of processing frequencies, from monthly to quinquennially, and a wide range of sample sizes, from several units to millions of units.

The IBSP system is built on the concept of a harmonized content model. Therefore, surveys must apply statistical standard classifications to categorize and collect business input and output data. Furthermore, processing in IBSP is predominantly carried out using Statistics Canada’s suite of generalized systems such as G-Sam for sampling (Statistics Canada, 2013), BANFF for edit and imputation (Statistics Canada, 2017), and G-Est for estimation (Statistics Canada, 2019). The methodology of IBSP is adapted to the Rolling Estimates Processing model, which is automated, iterative and runs regularly during the cycle of a survey until an acceptable level of quality is achieved (Godbout et al. (2011)).

One important goal of the IBSP system was to implement a methodology to estimate the total variance of estimates, i.e. the variance due to sampling and the variance due to imputation (VDTI). The variance due to imputation is estimated with Statistics Canada’s System for Estimation of Variance due to Nonresponse and Imputation (SEVANI) (Beaumont and Bissonnette (2011)) and quality indicators based on total variance are now produced for several economic surveys.

## 2. Context

Edit and imputation (E&I) strategies have become overly complex for most economic surveys processed in IBSP. Given the difference in availability of auxiliary data among units, composite imputation has been classically a method of choice for IBSP surveys. However, this leads to tens (and often more) of possible imputation models for

---

<sup>1</sup>Matei Mireuta, Statistics Canada, 100 Tunney’s Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6 ([matei.mireuta@statcan.gc.ca](mailto:matei.mireuta@statcan.gc.ca)); Ahalya Sivathayalan, Statistics Canada, 100 Tunney’s Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6 ([ahalya.sivathayalan@statcan.gc.ca](mailto:ahalya.sivathayalan@statcan.gc.ca)); Stephen Styles, Statistics Canada, 100 Tunney’s Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6 ([stephen.styles@statcan.gc.ca](mailto:stephen.styles@statcan.gc.ca))

each variable which makes implementation, support and analysis of a survey’s overall imputation strategy very difficult and costly.

In a previous research project, our team has investigated three alternatives that are “data”-driven in the sense that all available auxiliary information was considered and virtually no prior knowledge of underlying variable relationships was assumed. The first method was a parametric linear regression method (LR), the second method was a random forest algorithm (RF) and the third method was a cubist imputation algorithm (Cub) (Quinlan 1992 and Quinlan 1993). Overall, we have observed that these methods led to high quality imputation in line with what other authors have reported using synthetic data (Dagdoug et al. 2023). We have also observed overall that estimates were very similar to previously published estimates where current, standard imputation methods were used. The obvious advantages of these “data-driven” methods are their automation capabilities, their relatively low cost of implementation and their potential to significantly reduce and simplify current IBSP E&I strategies. However, the main methodological challenge we have encountered is the estimation of variance due to imputation of these methods, especially of the random forest and cubist models. Given SEVANI was developed as an analytical solution for linear composite imputation, it cannot in its current form assist with VDTI estimation of statistical learning methods. The first option for measuring VDTI of the RF and Cubist methods (and other statistical learning methods in general) are replicate based techniques, of which several exist in the literature (see e.g. Davison et al. 2007). However, this is still a largely open problem and would require significant further methodological research and investments for IBSP. It may be ultimately warranted, but before going in that direction, we wanted to explore accuracy, feasibility and fitness for use of two approximate solutions that are implementable within the current infrastructure. These two methods are detailed in sections 3 and 4 and preliminary simulation results are shown in section 5.

### 3. Imputation using *a priori* Predictions

#### 3.1 Method Description

The general method proceeds as follows. Let a sample be of size  $n$  and let  $\mathbf{y}_t$  denote the vector of all  $n$  sample values for variable  $y$  at time  $t$ , some of which may be respondent and some of which may be imputed. Let  $\mathbf{y}_t^R$  and  $\mathbf{y}_t^I$  denote the vectors of respondent and imputed values for variable  $y$  at time  $t$ . The first step of the method fits a model  $M$  at time  $t - 1$  as follows:

$$\mathbf{y}_{t-1}^R \sim M(\mathbf{X}_{t-1}, \mathbf{X}_{t-2}, \mathbf{X}_{t-3}, \mathbf{H}_{t-2}, \mathbf{H}_{t-3}) \quad (1)$$

where  $\mathbf{X}$  and  $\mathbf{H}$  denote the matrix of observed (or available) auxiliary information and historical information respectively, at different time points. The results presented in section 3.3 were obtained with the set-up above (i.e. two time periods for historical data and three time points for auxiliary data), but in general the user may wish to include more or less auxiliary and historical information as deemed appropriate for a particular purpose. Then, the second step of the method obtains a prediction for each sample unit at time  $t$  as follows:

$$\mathbf{y}_t^P = M(\mathbf{X}_t, \mathbf{X}_{t-1}, \mathbf{X}_{t-2}, \mathbf{H}_{t-1}, \mathbf{H}_{t-2})$$

The vector of predictions  $\mathbf{y}_t^P$  does depend on auxiliary data and on historical data, but not on current data  $\mathbf{y}_t$  and therefore is available before the start of collection. The final step of the method fits and then applies, at time  $t$  and for each rolling estimate iteration, an imputation model  $F_I$  using solely  $\mathbf{y}_t^P$  as auxiliary data:

$$\begin{aligned} \mathbf{y}_t^R &= F_I(\mathbf{y}_t^P) + \boldsymbol{\varepsilon} \\ \mathbf{y}_t^I &= \hat{F}_I(\mathbf{y}_t^P) \end{aligned} \quad (2)$$

The model  $M$  in (1) can in general be quite complex, but we have observed in preliminary studies that often, for suitable  $M$ , a simple linear method for  $F_I$  in (2) can yield very accurate imputation results. We have explored linear regression and nearest neighbor methods for  $F_I$ , select results of which are shown in section 3.3. One major advantage of the method of imputation using *a priori* predictions is that, for linear  $F_I$  functions, the conditional

variance of estimates involving  $\mathbf{y}_t$  can be obtained using the current IBSP infrastructure. For estimates of totals for example, it is straightforward to estimate the total variance  $Var(\hat{T}_y - T_y | \mathbf{y}_t^P)$  using SEVANI as currently implemented. One crucial assumption of this method that can be challenged is whether the inference can reasonably be conditioned on  $\mathbf{y}_t^P$ .

### 3.2 Model M and Aggregation

The model  $M$  in (1) can be a single model or an aggregation of separate models, it can be parametric or not, in fact it can be any model or method that produces predictions that are reasonably correlated with respondent values. In general, we anticipate that the choice of  $M$  will largely be dictated empirically.

The results in section 3.3 have been obtained via a convex linear aggregation of predictions obtained from three different models, namely a step-wise linear regression method (LR), a random forest imputation method (RF) and a cubist imputation method (Cub), as depicted in figure 3.3-1. In this case,  $\mathbf{y}_t^P = \alpha_1 \mathbf{y}_t^{PLR} + \alpha_2 \mathbf{y}_t^{PRF} + \alpha_3 \mathbf{y}_t^{PCub}$ , where  $\alpha_1, \alpha_2$  and  $\alpha_3$  are estimated using optimisation techniques on an independent subset of units, under the additional constraint that  $\alpha_1 + \alpha_2 + \alpha_3 = 1$ . In general, many other aggregation methods could be suitable, according to several possible quality measures, or none at all if one model significantly outperforms all others.

### 3.3 Select Preliminary Results

We have studied the method of imputation using *a priori* predictions for 10 variables for Statistics Canada's annual survey of manufacturing and logging (ASML) for reference period 2022.

As mentioned in section 3.2, an aggregated prediction  $\mathbf{y}_t^P = \alpha_1 \mathbf{y}_t^{PLR} + \alpha_2 \mathbf{y}_t^{PRF} + \alpha_3 \mathbf{y}_t^{PCub}$  was used in preliminary studies and results shown in this section are based on this aggregation method as well as several choices for  $F_I$ . Figure 3.3-1 shows results at the macroestimate level for one representative variable of the 10 examined, in this case the closing inventory of manufactured goods. In preliminary studies, we have explored several choices for  $F_I$  and have found that, often, many models lead to very similar macroestimates. Due to ease of implementation, a simple ratio model applied on aggregated predictions (third bar from the left in figure 3.3-1) or a nearest neighbor method (NNI) (last bar) would be recommended options.

The fourth bar in figure 5 shows a linear regression option for  $F_I$  applied on individual model predictions. The NNI method (last bar) can be particularly useful for variables that show significant nonlinear relationships to *a priori* predictions. There can also be two additional arguments made in favor of using NNI. First, in general, the model used to obtain *a priori* predictions is non-parametric, therefore a non-parametric method for  $F_I$  as well seems most conservative. Second, if a variable shows no correlation to *a priori* predictions, then using NNI as the model for  $F_I$  reduces the imputation procedure to a random hot-deck imputation which may be an acceptable worst-case scenario.

## 4. Imputation via Cubist Linearization

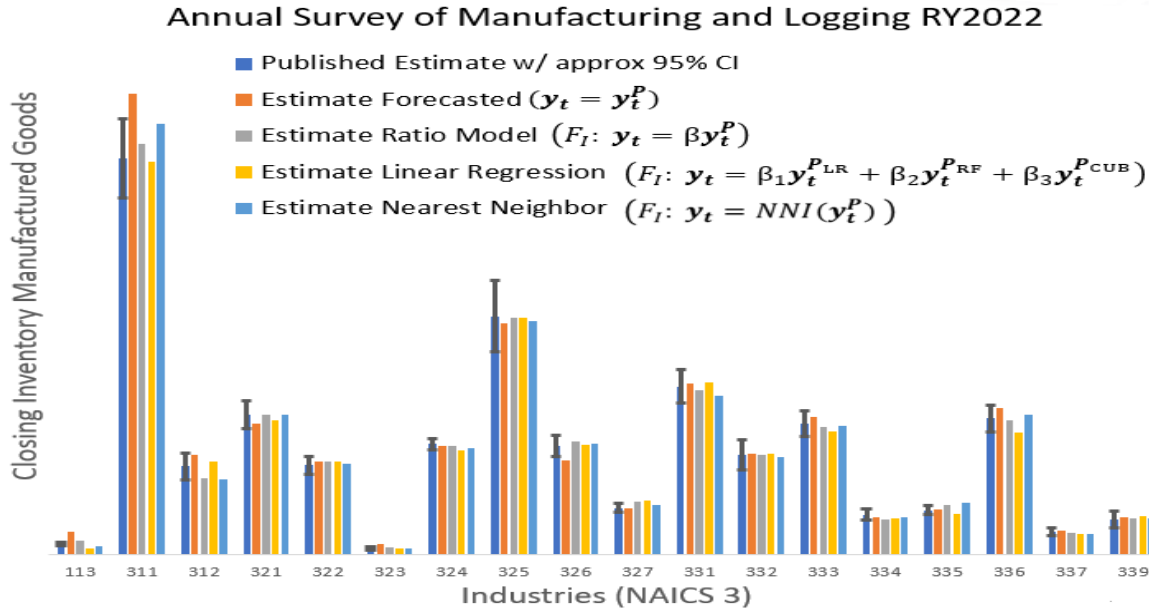
### 4.1 Cubist Model Description

The cubist imputation approach is a rule-based model that combines many methodologies (Quinlan 1992 and Quinlan 1993) and outputs linear regression models at each end node of a rule-based tree (see figure 4.1-1 as an example). The smoothing procedure at each node considers variance and covariance of child and parent nodes and collects the sequence of linear models at each node into a single, smoothed model per node. Additionally, overfitting is mitigated by incorporating an adjusted error rate as the criterion for pruning and/or combining rules. The cubist method can be used as a single tree, which will be denoted by  $M_{cub(1,0)}$  in this section, or as an ensemble method. As shown in figure 4.1-1, it is straightforward to automatically map the arborescence outputted by a model  $M_{cub(1,0)}$  into an IBSP compatible linear E&I strategy and make use of all IBSP functionalities (including SEVANI) without additional system infrastructure investments. The resulting E&I strategy does not necessarily have less steps than a

classical E&I strategy currently used in IBSP, as a matter of fact it often is much longer. However, such a strategy is simpler in that it is automatically generated without the need for manual intervention.

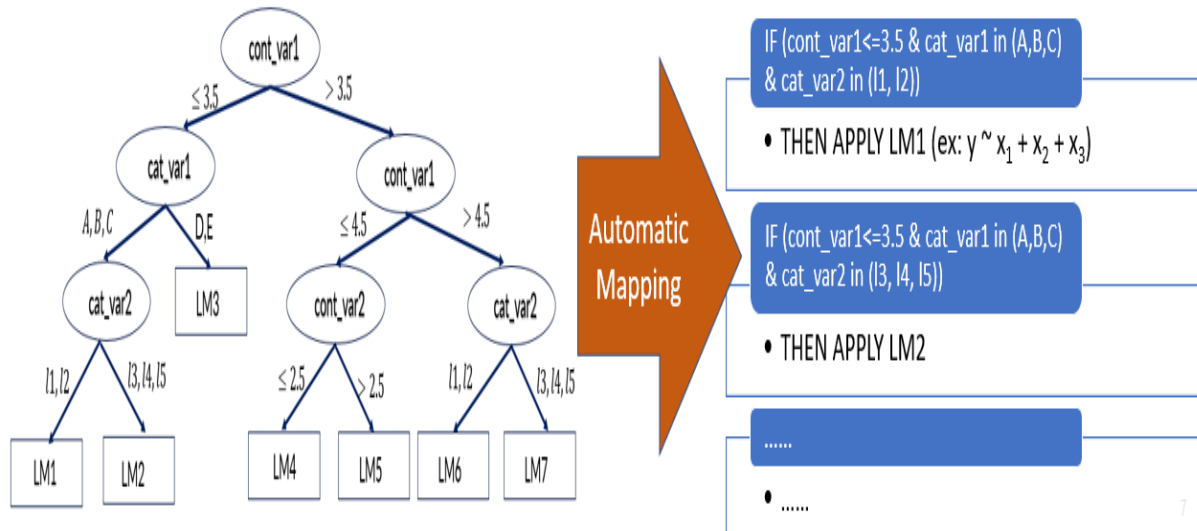
**Figure 3.3-1**

**Estimates of Total Closing Inventory of Manufactured Goods by Industry for Survey ASML 2022. The dark blue bars are current published production estimates with error bars representing an approximate 95% confidence interval.**



**Figure 4.1-1**

**Schematic Representation of Mapping Output from  $M_{cub(1,0)}$  into an IBSP E&I Strategy.**



## 4.2 Method Description

We propose to utilize  $M_{cub(1,0)}$ , as a linear approximation of any other machine learning methods or model aggregation methods proposed in section 3.2. Consequently, M could be any method, as given in Equation (1), used to obtain predictions  $\mathbf{y}_t^P$  in the same way as in section 3.1. Then, in a second step, a model  $M_{cub(1,0)}$  can be fitted to  $\mathbf{y}_t^P$  and its output mapped into  $F_I$ :

$$y_t^P \sim M_{cub(1,0)}(X_t, X_{t-1}, X_{t-2}, H_{t-1}, H_{t-2})$$

$$F_I \cong M_{cub(1,0)}(X_t, X_{t-1}, X_{t-2}, H_{t-1}, H_{t-2})$$

## 5. Unconditional Variance – Simulation Results

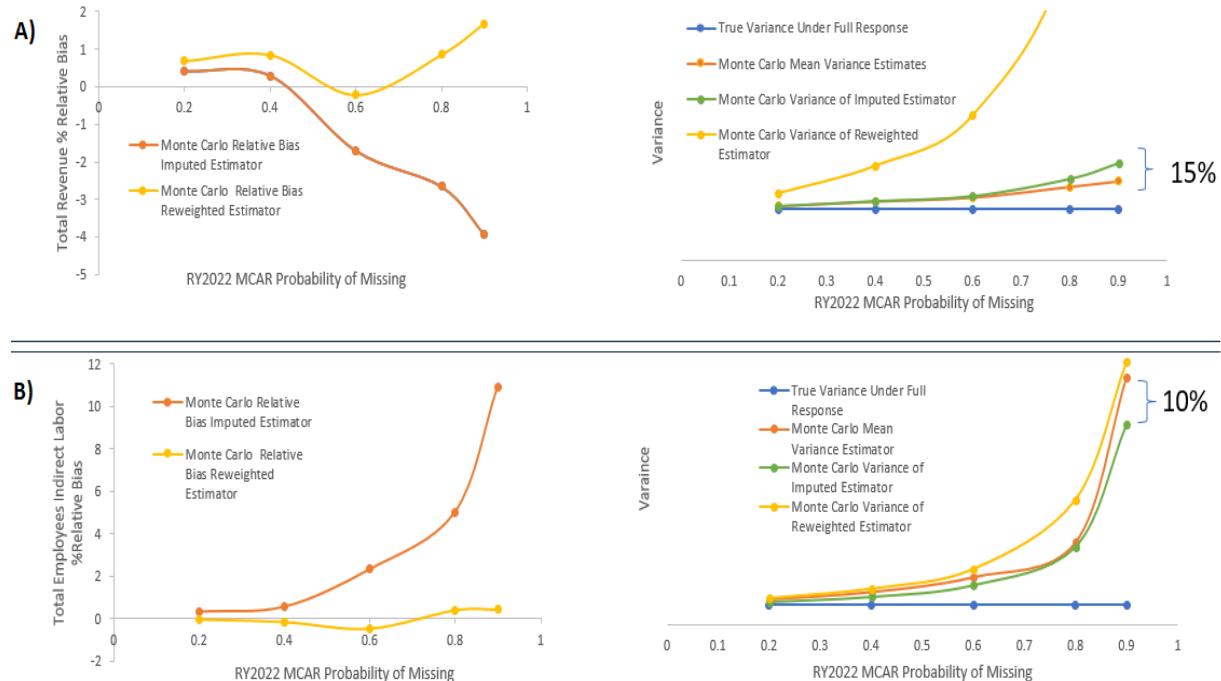
One concern of both proposed methods is the fact that variance estimation is conditional, on either the predictions or on the tree and associated node models, both of which depend on observed historical and auxiliary data. In order to assess the unconditional variance as well as the bias of the imputed estimator and of the variance estimator, we conducted a short simulation study. We evaluated the method of *a priori* predictions described in section 3, even though we used the cubist imputation method. The ideas described in section 4 will be evaluated in future studies.

A pseudo-population of 540 units was created containing respondents of the Annual Survey of Manufacturing and Logging for years 2021 and 2022. For each iteration, a sample of 108 units was selected using SRS for year 2021 and non-response was generated missing completely at random (MCAR). Several choices for the probability of non-response were tried without much impact on results (data not shown). Then, a cubist method was used to obtain predictions for year 2022 using 2021 respondent values and a set of auxiliary data. The same sample was selected for year 2022 and non-response was generated once more (MCAR). Finally, estimates of the population total for year 2022 were obtained together with corresponding variance estimates as per the methodology described in SEVANI (Beaumont and Bissonnette (2011)) and using a ratio imputation model ( $F_I: y_t = \beta y_t^{CUB}$ ).

Figure 5-1 shows the Monte Carlo (MC) mean relative bias as well as the MC variance and MC mean of variance estimates for a well predicted variable (Total Revenue, figure 5-1 A) and a difficult to predict variable (Total Employees in Indirect Labor, figure 5-1 B). In both cases, the imputed estimator remains reasonably unbiased unless conditions are extreme (>80% non-response). Similarly, the conditional variance estimator remains reasonably unbiased for the unconditional variance unless conditions are again extreme. In both cases, MC properties of the simple reweighted estimator (assuming MCAR) are shown as a comparison (in yellow). In future studies, our aim is to explore the behavior of these estimates under non-response models that are more realistic and for other prediction methods and imputation methods.

**Figure 5-1**

**MC relative Bias, MC Variance and MC Mean of Variance Estimates of several Estimators obtained via Simulation for two variables of the Survey ASML 2022. The Graphs are a Function of Year 2022 Probability of Non-Response (Missing).**



## 6. Conclusion

In this article, we have described preliminary ideas for two general approaches that could be used to bring modern imputation methods into the current IBSP framework in a scalable and automated way. We have observed promising results in preliminary studies and have observed several other advantages of these methods, notably in the simplification of production E&I strategies and in their usefulness in the IBSP collection methodologies. However, we are cognizant that both approaches rely on assumptions and, in general, that more work will be required to fully evaluate and fine tune these methods.

## Acknowledgements

The authors would like to thank the following colleagues for their advice and discussion: Steve Matthews, Marie-Claude Duval, Jean-François Beaumont, Keven Bosa and David Haziza as well as Pierre Daoust and Darren Gray for their much appreciated review of this article.

## References

- Beaumont, J-F., and Bissonnette, J. (2011), “Variance estimation under composite imputation: The methodology behind SEVANI”, *Survey Methodology*, Statistics Canada, 37(2), pp. 171-179.
- Dagdoug, M., Goga, C., and Haziza, D. (2023), “Imputation procedures in surveys using nonparametric and machine learning methods: an empirical comparison”, *Journal of Survey Statistics and Methodology*, 11, pp. 141-188.
- Davison, A. C., and Sardy, S. (2007), “Resampling variance estimation in surveys with missing data”, *Journal of Official Statistics*, 23 (2007), pp. 371-386.
- Godbout, S., Beaucage, Y., and Turmelle, C. (2011), “Achieving Quality and Efficiency Using a Top-Down Approach in the Canadian Integrated Business Statistics Program”, *Proceedings of the Conference of European Statisticians*. Ljubljana, Slovenia.
- Quinlan, J. (1993), “Combining instance-based and model-based learning”, *Proceedings of the tenth International Conference on machine learning*, pp. 236–243.
- Quinlan, J. (1992), “Learning with continuous classes”, *Proceedings of the 5th Australian Joint Conference on Artificial Intelligence*, 92, pp. 343–348, World Scientific.
- Statistics Canada (2013), G-Sam Methodology Guide, Technical Report (Draft).
- Statistics Canada (2015), Integrated Business Statistics, Program Overview. Catalogue no. 68-515-X, ISBN 978-0-660- 02486-8.
- Statistics Canada (2017), Functional Description of the Banff System for Edit and Imputation, Technical Report.
- Statistics Canada (2019), G-Est Methodology Guide, Technical Report.