

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

The Usage of the Relief Algorithm for Edit and Imputation in the Canadian Census of Population

by Irwin Khuu

Release date: September 8, 2025



Statistics
Canada

Statistique
Canada

Canada

The Usage of the Relief Algorithm for Edit and Imputation in the Canadian Census of Population

Irwin Khuu¹

Abstract

Historically, the Canadian census of population Edit and Imputation (E&I) process has operated using a nearest-neighbour donor imputation methodology wherein the distance between a failed unit and a potential donor is obtained through a weighted combination of auxiliary variables. Revision to the model between cycles can be a complicated and time-consuming process given there is no standard approach to variable selection and weighting between topics. This paper will illustrate the potential of the Relief variable selection algorithm to create a machine learning-driven approach to variable selection and weighting that is standardized and comparable between census cycles and among the many topics of the census. An overview on how this process may be applied in practice will be presented, followed by results on several topics that indicate a general improvement over previous methods.

Key Words: Edit and imputation; Canadian Census; Variable selection; Relief; Machine learning.

1. Introduction

Edit and imputation is a complicated and time-consuming process within the Canadian census of population, with hundreds of edit rules for any given census questionnaire topic. In the census, imputation is conducted using Statistics Canada's generalized system known as the Canadian Census Edit and Imputation System (CANCEIS), which operates using a nearest-neighbour donor imputation methodology. This methodology allows for the efficient determination of minimum change imputation actions for simultaneous and multivariate hot-deck imputation for a mixture of quantitative and qualitative variables.

Responses are split between records failing one or more edit rules and thus requiring imputation (failed records) and those that pass the edit rules (passed records). For each failed record, the most similar passed records (potential donors) are found. Subsets of a potential donor's data are determined such that when imputed into the failed record would cause it to no longer fail any edit rules (imputation actions).

Resultingly, the quality of the imputation depends on the ability of the imputation model to successfully measure the similarity between neighbours. In CANCEIS, the distance between a failed-donor pair is defined by a weighted combination of distance functions:

$$D_{fp} = \sum_i w_i D_i(V_{fi}, V_{pi}) \quad (1)$$

where for each variable in the model i , $w_i \geq 0$ is the weight assigned and $D_i(V_{fi}, V_{pi})$ is a function on $[0,1]$ representing the distance between the value of the failed record V_{fi} and the value of the potential donor V_{pi} . The weights allow specification of the importance of similarity between the failed and potential donor values for each variable.

As a result, the problem of developing a high quality CANCEIS model centers around both the selection and weighting of auxiliary variables. Within the Canadian census, the decision-making strategies for this problem are developed on a per-topic basis through the collaboration of subject matter and methodology. This tends to be an intricate and time-consuming process, further complicated by the potential addition of questions, variables, or changes in scope between

¹Irwin Khuu, Statistics Canada, 150 Tunney's Pasture Driveway, Canada, K1A 0T6 (irwin.khuu@statcan.gc.ca)

census cycles. Hence, the development of a systemized and generally applicable process that can be standardized and compared across the numerous census topics merits further exploration.

Previous investigations (Stelmack, 2019) identified the Relief family of algorithms to be a promising strategy for the imputation of the Admission Category topic in the 2016 Canadian Census, outperforming other feature selection methods such as Information Gain Ratio and Random Forest. This paper extends the usage of Relief algorithms to other topics present in the 2021 Census and compares their performance to the 2021 Census models.

2. Relief Overview

The Relief family of algorithms are a set of feature selection algorithm that, similarly to CANCEIS, use a nearest-neighbour approach to determining variable importance. That is, auxiliary variables that are useful for the prediction of a variable of interest should have similar values among records that have the same value in the variable of interest. The original Relief algorithm (Kira and Rendell, 1992a) has generally been supplanted by ReliefF (Kononenko, 1994), an extension that allows the algorithm to handle missing data and multi-class variables of interest, as well as generally providing more reliable weight estimates.

In general, for a given target variable, the algorithm calculates auxiliary variable weights by iterating through a list of records and in each case (the target record) finding the nearest “hits” (neighbours with the same target variable value) and “misses” (neighbours with a different target variable value). For each neighbour, it then updates the weights of the auxiliary variables, either positively if the auxiliary variable differs between the target record and a nearest miss, or negatively if the auxiliary variable differs between the target record and a nearest hit. The result is a vector of weights in $[-1,1]$ where each element represents the relevance of an auxiliary variable, i.e. its ability to predict the target variable. Weights close to zero, or negative, indicate irrelevant predictors, which can effectively be removed from a prediction model without compromising prediction quality. More thorough explanations can be found in (Kononenko, 1994), (Kira and Rendell, 1992b) and (Urbanowicz, 2018).

For implementation in CANCEIS, Relief weights can serve as a relevant measure and can be used to both discard variables that seem unproductive for predicting a given target variable, as well as transformed and implemented as CANCEIS imputation weights.

3. Methodology

For each topic investigated, high-quality responses were identified to use for training and testing a Relief-based imputation model. Within each data set, the records are split into a training set, to train the Relief model, and a test set, to evaluate the model’s performance and compare it to the 2021 Census model.

Creation of a Relief model involves both selection of auxiliary variables and weighting. For variable selection, the use of automatic variable importance evaluation allows the introduction of new potential auxiliary variables to the model without requiring onerous efforts by subject matter experts and methodology to determine how best to include them in the imputation model. The Relief algorithm itself is implemented in R using the **CORElearn** package (Robnik-Šikonja and Savicky, 2022).

The Relief algorithm implementation only considers a single target variable in its training. In the case of multivariate imputation, a separate Relief model is generated for each target variable. The weights are standardized and averaged across all models to obtain a general set of weights, then rescaled for implementation in CANCEIS.

A small proportion of the test set data, dependent on the topic, has artificial non-response induced in the variable(s) of interest. Non-response was set through a Missing Completely At Random (MCAR) response propensity mechanism. The non-responses are imputed in CANCEIS using both the Relief model with variable selection and the 2021 Census model. The results are checked against the true value to obtain estimates of overall accuracy,

macro-recall, and macro-precision. Furthermore, distributional goodness-of-fit were checked using a chi-square test and inspection of the Cramér's V statistic.

4. Datasets

The census topics investigated are Official Language and Mother Tongue. Only responses which did not require either deterministic edits or donor imputation were used for analysis so as to reliably estimate the quality of our donor imputation model without being confounded by other edit and imputation steps. Furthermore, many CANCEIS donor modules include their own stratification of the investigated population. These are also adopted into the selection criteria for the Relief model. That is, separate Relief models are created for each unique stratum.

4.1 Official Language

The Official Language topic imputes a single variable of interest: the unit's known official languages. The response can be English, French, bilingual or neither. Four Relief models corresponding to Indigenous communities, collective dwellings, single-person households, and the rest of the population were created. Due to training time considerations, although around 5.65 million responses were eligible respondents for training the Relief model, a reduced and class balanced training set containing 1.42 million responses was used to generate the Relief models. Approximately 870,000 responses were used in the test set, including Indigenous communities, collective dwellings, and members from the general population. The data were imputed in three separate strata: Indigenous communities, collective dwellings, and the general population. Non-response rate in the test set was set at 2% for the general population, while due to the small number of cases non-response for Indigenous communities and collective dwellings was set at 10%.

The 2021 Census model used seven auxiliary variables including demographics, administrative data, and previously processed language responses. Using Relief, three to five additional auxiliary variables were found to be suitable, incorporating additional census-family- and household- level information, geographical data, as well as information on Indigenous status for respondents in Indigenous communities.

4.2 Mother Tongue

The Mother Tongue topic imputes three variables: two binary variables indicating whether English or French are mother tongues, as well as a high-cardinality write-in variable for non-official languages, with hundreds of potential imputable values. Five Relief models corresponding to Indigenous communities on reserves, collective dwellings, single-person census families, single-person households, and two adult, two children census families were created. The Relief models were training on a total of 2.03 million records. Approximately 1.33 million responses were used in the test set, including Indigenous communities on reserves, collective dwellings, and the general population. The data were imputed in three separate strata: collective dwellings, single-person census families, and census families containing two adults and two children. Non-response rate in the test set was set at 7% for private dwellings, while due to the small number of cases non-response for collective dwellings was set at 15%.

The 2021 Census model used ten auxiliary variables relating to administrative data, home language questionnaire response, and household-level information for all respondents, and an additional three auxiliary variables involving Indigenous status for those on Indigenous reserves. The Relief models included an additional six variables including demographic and geographical information as well as official language questionnaire response.

5. Results

When checking model results, various types of imputation-related quality metrics are used: record-level accuracy, class-level precision and recall, and distributional goodness-of-fit. At the record level, accuracy is simply determined by checking if the missing-response record has the same value imputed as it was originally. However, in class-imbalanced data sets, accuracy can be a poor measure of imputation quality. Instead, the following class-level

precision are examined: the proportion of correctly imputed records out of all imputed records; and recall: the proportion of correctly imputed records out of all records in the class. As analysis of individual classes can be strenuous in high cardinality imputed variables, macro-averaged values of precision and recall among all classes are provided.

Distributional goodness-of-fit is an even more important measure of imputation quality with respect to impacts on further analysis. It should be expected that under the MCAR non-response model used here that the distribution of the imputed data should closely match the distribution of the unimputed data. Distributional goodness-of-fit is measured by examining the results of a chi-square goodness of fit test on the imputed data compared to the relative frequencies in the unimputed data. The chi-square test can be unreliable in cases of high cardinality with small or zero cell counts, so sparse class realizations are collapsed into a single “Other” category when relevant. Under well-fitting imputation, the resulting chi-square test statistics are expected to be low, indicating a close match between the imputed and unimputed data.

Approximate distributional quality is further examined using Cramér’s *V* statistic. Although it is typically used to measure the association between two categorical variables, it can be adapted for a goodness-of-fit role by comparing observed frequencies to expected counts. There is no specific hypothesis test associated with the measure, but similarly to the chi-square test, that the Cramér’s *V* is expected to be low, and differences between the 2021 Census model and relief model may be assessed heuristically.

5.1 Official Language

As previously mentioned, the Official Language topic imputes the record’s known official languages, with four possible responses: English, French, bilingual or neither. The data was separated into three strata: Indigenous communities, collective dwellings, and the general population.

Table 5.1-1 provides accuracy, macro-precision, and macro-recall results between the 2021 production and Relief models for each stratum. In all strata, the Relief model outperforms the 2021 Census model on all metrics. Inspection of the imputation results show that while the gains were present among all four classes, they were especially prevalent among the “bilingual” and “neither” classes, which had particularly low precision and recall.

**Table 5.1-1
Official Language Imputation: Accuracy, Precision and Recall**

Stratum	Accuracy		Macro-Precision		Macro-Recall	
	2021 Production	Relief	2021 Production	Relief	2021 Production	Relief
Indigenous Communities	90.27%	94.80%	57.53%	75.26%	55.08%	70.62%
Collective Dwellings	89.07%	94.26%	57.21%	76.04%	57.64%	77.04%
General Population	82.87%	89.88%	70.48%	82.43%	68.60%	81.77%

Table 5.1-2 provides measures of distributional goodness-of-fit between the 2021 production and Relief models for each stratum. In all strata, for both the chi-square test and Cramér’s *V* statistic the Relief model outperforms the 2021 Census model with respect to distributional goodness of fit. Most notably, among the general population, inspection of the imputation results shows that the Census model overestimates the number of records in the “English” class, while underestimating the “bilingual” and “neither” classes.

Table 5.1-2
Official Language Imputation: Chi-Square Test and Cramér's V

Stratum	2021 Census model				Relief Model			
	Chi-Square Test			Cramér's V	Chi-Square Test			Cramér's V
	Test Statistic	df	p-value	Statistic	Test Statistic	df	p-value	Statistic
Indigenous Communities	5.869	3	0.12	0.0465	4.78	3	0.18	0.0420
Collective Dwellings	4.953	3	0.18	0.0672	0.747	3	0.86	0.0260
General Population	22.961	3	4.12×10 ⁻⁵	0.0212	2.216	3	0.53	0.0066

df degrees of freedom

5.2 Mother Tongue

As previously mentioned, the Mother Tongue topic imputes three variables: two binary variables for English and French mother tongues, as well as a write-in variable for non-official languages. The data was separated into three strata: collective dwellings, single-person census families, and census families containing two adults and two children. Although results were obtained for all imputable variables in each stratum, for brevity only the results for the imputation of the write-in are presented here.

Table 5.2-1 provides accuracy, macro-precision, and macro-recall results between the 2021 production and Relief models for each stratum when imputing Mother Tongue write-ins. In all strata, the Relief model outperforms the 2021 model on all metrics, especially in precision and recall, which the 2021 model fared relatively poorly.

Table 5.2-1
Mother Tongue Write-In Imputation: Accuracy, Precision and Recall

Stratum	Accuracy		Macro-Precision		Macro-Recall	
	2021 Production	Relief	2021 Production	Relief	2021 Production	Relief
Collective Dwellings	90.38%	94.18%	38.51%	48.55%	31.14%	34.21%
Single-Person Census Families	89.92%	93.13%	35.83%	57.55%	27.35%	46.05%
Two-Adult, Two-Children Census Families	89.94%	94.44%	36.51%	74.73%	19.72%	46.47%

Table 5.2-2 provides measures of distributional goodness-of-fit between the 2021 production and Relief models for each stratum when imputing Mother Tongue write-ins. In general, there were mild write-in improvements in the Relief model over the census model for private dwellings, but this relationship reversed in collective dwellings. That said, a heuristic examination of the measures used suggest that neither model performed especially well in maintaining similarities between the imputed and unimputed data.

Table 5.2-2
Mother Tongue Write-In Imputation: Chi-Square Test and Cramér's *V*

Stratum	2021 Census model				Relief Model			
	Chi-Square Test			Cramér's <i>V</i>	Chi-Square Test			Cramér's <i>V</i>
	Test Statistic	df	p-value	Statistic	Test Statistic	df	p-value	Statistic
Collective Dwellings	30.881	25	0.193	0.0382	43.82	26	0.016	0.0447
Single-Person Census Families	482.10	248	≈ 0	0.0066	261.71	248	0.149	0.0049
Two-Adult, Two-Children Census Families	740.13	213	≈ 0	0.0085	411.28	213	≈ 0	0.0063

df degrees of freedom

Although not presented, measures of imputation quality were obtained for the official language and mother tongue imputation. In general, the accuracy, precision and recall measures were very high for both models, although there were consistent improvements found in the Relief model over the 2021 Census model. However, on a distributional level, results were generally mixed with neither model having a consistent advantage. Furthermore, according to the measures used, the ability of the models to maintain distributional goodness-of-fit in general was somewhat dubious.

6. Discussion

Comparisons of the Relief and 2021 Census models indicated general improvements at the unit level and on a per-class basis for both univariate and multivariate imputation. At the distributional level, the Relief model also improved upon the 2021 model in the univariate case, while results were inconclusive in complex, high-cardinality multivariate imputation cases such as the Mother Tongue topic. The results for distributional goodness-of-fit in the case of multivariate imputation should be taken with some caution, as the measures used cannot reliably assess multiple distributional changes in multiple variables simultaneously and may be conservative as a result.

There are many potential avenues for further research. It should be highlighted that these results were obtained for MCAR non-response. The quality of the Relief model, in comparison to census models, under missing at random (MAR) or missing not at random (MNAR) non-response mechanisms is a topic of further research, although research (Stelmack, 2019) does suggest these results may still hold in such cases. The imputation of metric variables, such as income, is also not present in this analysis—and ReliefF would be unsuitable in this case, but regression-based Relief algorithms such as RReliefF (Robnik-Šikonja and Kononenko, 2003) are also available in the CORElearn package and could be considered. Regarding variable selection, only those with negative Relief weights were removed from the Relief imputation model, but alternative decision criteria can be considered. For example, (Kira and Rendell, 1992b) suggest an inclusion threshold based on Chebyshev's inequality. Finally, the analysis shows that Relief is still imperfect for maintaining distributional goodness-of-fit in multivariate imputation. Whether this can be reliably improved remains unresolved.

With all these caveats in mind, the results show that the inclusion of Relief in the creation of imputation models has great potential in improving imputation quality in future Canadian censuses. In more complex imputation problems, Relief model outputs should serve primarily as a baseline and used to advise subject matter and methodology on potential variable inclusion decisions, rather than being taken wholesale.

7. Acknowledgements

The author would like to thank the other members of the census edit & imputation methodology team, who provided valuable feedback and intuition in the development of this research.

References

- Kira, K. and Rendell, L. A. (1992a), "A Practical Approach to Feature Selection", *Machine Learning Proceedings 1992*, pp. 249-256.
- Kira, K. and Rendell, L. A. (1992b), "The Feature Selection Problem: Traditional Methods and a New Algorithm", *Proceedings of the AAAI Conference on Artificial Intelligence*, 10, pp. 129-134.
- Kononenko, I. (1994), "Estimating Attributes: Analysis and Extensions of Relief", *Machine Learning: ECML-94*, pp. 171-182.
- Kononenko, I., Šimec, E. and Robnik-Šikonja, M. (1997), "Overcoming the Myopia of Inductive Learning Algorithms with RELIEFF", *Applied Intelligence*, 7, pp. 39-55.
- Robnik-Šikonja, M. and Kononenko, I. (2003), "Theoretical and Empirical Analysis of ReliefF and RReliefF", *Machine Learning*, 53, pp. 23-69.
- Robnik-Šikonja, M. and Savicky, P. (2022), CORElearn: Classification, Regression and Feature Evaluation. R package version 1.57.3.1. <https://CRAN.R-project.org/package=CORElearn>.
- Stelmack, A. (2019), "Improving Edit and Imputation Strategies Through Feature Selection", *2019 Joint Statistical Meetings – Government Statistics Section*, pp. 3083-3095.
- Urbanowicz R.J., Meeker., La Cava W., Olson R. S., and Moore J. H. (2018), "Relief-based feature selection: Introduction and review", *Journal of Biomedical Informatics*, 85, pp. 189-203.