Statistical Journal of the IAOS 38 (2022) 1177–1183
DOI 10.3233/SJI-220082
IOS Press

Predicting the quality and evaluating the use of administrative data for the 2021 Canadian Census of Population

Erin R. Lundy
Statistical Integration Methods Division, Statistics Canada, 150 Tunney's Pasture Driveway, Ottawa, Ontario, K1A 0T6, Canada
Tel.: +1 613 298 9867; E-mail: erin.lundy@statcan.gc.ca

Abstract. This paper presents the statistical contingency plan for the 2021 Canadian Census of Population, developed in response to the COVID-19 pandemic, wherein administrative data was to impute non-responding households in areas with a low response rate and where the administrative data were of sufficient quality. We describe the modeling approach for predicting the quality of data available for administrative households, including important extensions to existing approaches. As well, we provide a framework for evaluating direct imputation using administrative data, relative to traditional donor imputation, in the absence of a simulation study. We conclude by discussing the evaluation using preliminary data and subsequent implementation for the 2021 Canadian Census of Population.

Keywords: Administrative data, census non-response, direct imputation, Young Statistician Prize 2022

1. Introduction

Before the World Health Organisation declared a global pandemic in March 2020, natural disasters had impacted or limited Census field operations. In Canada, wild fires in 2016 and flooding in 2011 had necessitated that Statistics Canada prepare local contingency plans using administrative data as a way to compensate for non-response. These events launched a long term research agenda towards the use of administrative data in a combined census approach.

In 2020, the increased use of administrative data in census collection and research towards a combined census were also under development in other countries [1–3]. However, the advancement of the pandemic in March 2020 accelerated exponentially the research regarding the potential use of administrative data for the 2021 Canadian Census of Population, in light of this global emergency and the associated public health measures.

Statistics Canada developed a statistical contingency plan to mitigate a low response rate in the event that the pandemic affected collection. The plan was to use administrative data to impute non-responding households in areas with a low response rate and where the administrative data are of sufficient quality. The impact of the pandemic on the response rate was unknown and, therefore, the use of administrative data was reserved for processing stages following the traditional collection process. For this purpose, we adapted the modeling approach used by other countries, namely, the US Census Bureau [3] and Statistics New Zealand [4] to identify administrative households with good quality data.

Model development was based on data from the 2016 Census. However, the response rate for the 2016 Canadian Census of Population was a record high for the country (98%) and unlikely to reflect the response mechanisms observed during a pandemic. The contingency plan developed a timely but reliable framework to evaluate direct imputation using administrative data, relative to traditional donor imputation, under a variety of response mechanisms. Moreover, this framework allowed us to evaluate the identification of households

with good quality data using preliminary data from the 2021 Census during the collection period and adjust parameter specifications accordingly.

The remainder of this paper proceeds as follows. In Section 2, we describe the modeling approach for predicting the quality of data available for administrative households. Thereafter, in Section 3, we discuss the model development using 2016 Census data. In Section 4, we present the evaluation using preliminary data and subsequent implementation for the 2021 Census. Conclusions are provided in Section 5.

2. Modeling approach for predicting the quality of administrative households

Census data are essential for a country, as all layers of society use census data. In particular, it is often the only source of information for small sub-populations. Producing high quality census data is the objective of any National Statistics Organisation. It became evident that one integral part of the research on how to incorporate administrative data into a traditional enumeration census is the evaluation of the quality of the administrative data itself. We use a modeling approach to rank the quality of the available administrative data at the household level. Broadly, this approach is termed the household model and consists of three components: the person-place model, the household composition model and a distance metric.

The basis of the household model is a database of administrative persons, created for the sole purpose of the Census research, composed of multiple sources acquired by Statistics Canada from other government departments. This database includes a variable predicting if the administrative person is in-scope for the Census, the person's age and sex at birth, all of which are determined using probabilistic models. As well, auxiliary data are available from a variety of administrative data sources such as tax files, immigration files and vital statistics files. Some but not all of these data sources include detailed address information. From these, a list of unique person-address pairs is created. Note that all possible addresses are included in this list and, therefore, a person may have more than one administrative address. Conversely, a person may have no administrative address.

2.1. Person-place model

The first component of the household model, the person-place model, predicts the probability that an administrative person is observed at the correct dwelling. The population of eligible persons consists of the set of persons deemed to be in-scope for the Census with a least one administrative address in the list of person-address pairs. Let

$$y_{ih}^{PP} = \begin{cases} 1 \text{ if person } i \text{ is found in administrative records and 2016 Census at dwelling } h \\ 0 \text{ otherwise} \end{cases}$$

We model the probability that person i is correctly placed at address h, $p_{ih} = P(y_{ih}^{PP} = 1)$, using logistic regression. For each person-address pair, we obtain a person-level estimated probability of coherence. If person i has administrative records at more than one dwelling, we assign the address with highest predicted probability, $\max_h \hat{p}_{ih}$, to that person. Next, we form administrative households, defined as all persons assigned to a given dwelling. For each dwelling h, we defined the dwelling-level estimated probability of coherence as

$$\hat{p}_h^{PP} = \min(\hat{p}_{1h}, \ldots, \hat{p}_{n_h h})$$

where n_h is the size of the administrative household at dwelling h. This provides a conservative estimate of the probability that every member of the administrative household is correctly placed at that dwelling.

2.2. Household composition model

The household composition model is used to predict the probability that an administrative household matches the household observed in the Census of Population. The household composition model applies to all dwellings with at least one administrative person. The outcome of interest, Y_h^{HC}, is categorical and has four levels, called coherence levels. The coherence levels characterize dwellings in terms of the degree to which the administrative household matches the census household at the person-level. These levels cover three dimensions of similarity: correct placement of administrative person(s), number of persons and household composition. The household composition indicates the presence of children less than 18 years old and/or the presence of adults 18 years or older. The four coherence levels for the household composition model are detailed in Table 1.

We model the probability that dwelling h belongs to each coherence level using multinomial logistic regression. In particular, the non-match coherence level is used as the baseline category and we specify three

Table 1
Coherence levels for the household composition model

Coherence level	Description
1	Perfect match – administrative household exactly matches census household.
2	Partial match (type 1) – At least one administrative person matches the census household, the administrative household count is greater or equal to the census count and the composition matches.
3	Partial match (type 2) – At least one administrative person matches the census household, the administrative household count is less than the census count and/or the composition does not match.
4	Non-match – No administrative person is matched to the census household.

independent binary logistic regression models:

$$\begin{cases} \log \dfrac{P(Y_h^{HC}=1)}{P(Y_h^{HC}=4)} = \beta_1 \boldsymbol{X_h} \\ \log \dfrac{P(Y_h^{HC}=2)}{P(Y_h^{HC}=4)} = \beta_2 \boldsymbol{X_h} \\ \log \dfrac{P(Y_h^{HC}=3)}{P(Y_h^{HC}=4)} = \beta_3 \boldsymbol{X_h} \end{cases}$$

This yields three sets of estimated regression coefficients. The primary estimate of interest is the probability of perfect match which we calculate as:

$$\hat{p}_h^{HC} = \frac{e^{\hat{\beta}_1 \boldsymbol{X_h}}}{1 + \sum_{k=1}^{3} e^{\hat{\beta}_k \boldsymbol{X_h}}}$$

Note that this specification of the household composition model differs from that proposed by [3] to identify households with good quality administrative data. This previous approach defined a household composition match based on number of adults and children and does not consider the person-level links.

2.3. Distance metric

Ideally, we want to accurately identify dwellings where high quality administrative data is available for every household member. This corresponds to a perfect match under the household composition model. However, a limitation of the household composition model is that the proportion of true perfect matches is overestimated. In order to address this limitation, we use a distance metric which incorporates both the estimated probability of a perfect match from the household composition model and the dwelling-level estimated probability of coherence from the person-place model into one measure of quality for dwelling-level administrative data.

We use an extension of the Euclidian distance-based metric initially proposed by [5] with a penalty for administrative households of size 1. This penalty was implemented, since preliminary analyses indicated that single person households were overrepresented within the dwellings predicted to be high quality. The distance metric for dwelling h is defined as:

$$d_h = \sqrt{(1 - \hat{p}_h^{PP})^2 + (1 - (\hat{p}_h^{HC})^{e_h})^2}$$

where $\hat{p}_h^{PP}$ is minimum estimated probability from the person-place model for all persons placed at dwelling h, $\hat{p}_h^{HC}$ is the estimated probability that dwelling h is a perfect match from the household composition model and the penalty term $e_h = 1$ for households with $n_h = 1$ and $e_h = 1/2$ otherwise. A smaller value for the distance metric indicates dwellings with better quality administrative data.

The use of a distance metric allows us to rank dwellings by a single measure of quality. Once appropriate threshold value(s) are determined, all dwellings below the specified threshold(s) are deemed of good quality and, therefore, eligible for further use. This methodology is quite flexible. Once the distance metric values are calculated for all dwellings, we can easily change the specified threshold value to suit the intended data use. As well, if other relevant statistical models are available, we can readily incorporate additional inputs into the distance metric.

3. Model development: Retrospective study of the 2016 Census

In preparation for the 2021 Census, we evaluated the identification of dwellings where good quality administrative data is available, using data from the 2016 Census. The person-place and household composition models were fitted using auxiliary data that reflects the vintages of administrative data that were available prior to the 2016 Census. Both regression models included person-level auxiliary variables such as age, sex at birth and number of addresses as well as dwelling-level auxiliary variables such as dwelling type and geography.

For the person-place model, the set of eligible persons corresponded to over 118 million unique person-address pairs. Due to computational constraints, the logistic regression model was fit using a training data set of 1% of the unique addresses in the provinces and

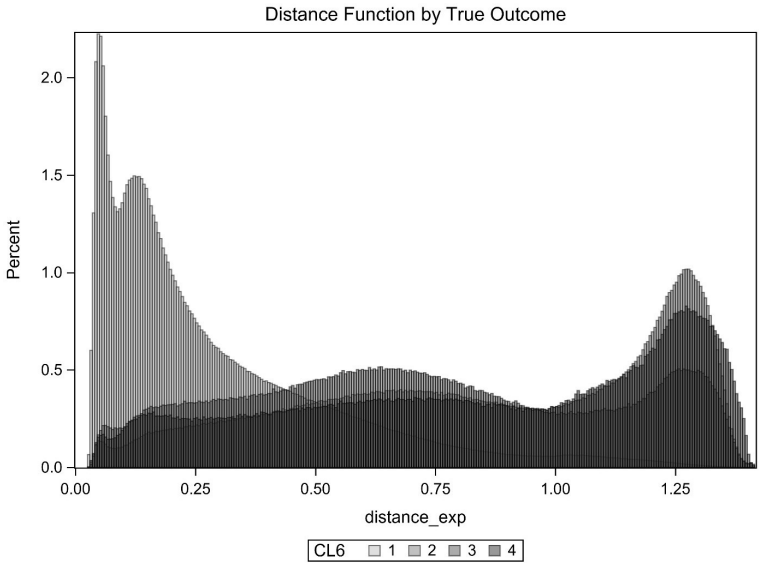

Fig. 1. Empirical distribution of distance metric.

20% of the unique addresses in the sparser populated territories. Variable selection was conducted using a forward step-wise procedure. The estimated coefficients were then applied to the entire data set to obtain person-level estimated coherence probabilities for each eligible person-address pair. Analyses using classification trees and random forests yielded similar results as the logistic regression model. This is consistent with previously conducted analyses [6].

Similarly, for the household composition model, the multinomial logistic regression model was fit using a training data set that corresponded to 2.5% of the dwellings in the provinces and 33% of the dwellings in the territories. Here, variable selection and model fitting was done simultaneously using the grouped LASSO approach. This approach is a variation of LASSO which uses a constraint to force all parameters corresponding to the same effect to be either included or excluded simultaneously. The estimated regression coefficients were then applied to all eligible dwellings. Finally, we calculated the distance metric for every dwelling with at least one administrative person.

The distribution of the distance metric by the true coherence level (CL) is displayed in Fig. 1. As expected, the majority of dwellings with a low distance metric value are indeed true perfect matches. As well, the distributions for partial matches and non-matches are left-skewed, meaning that true partial and non-matches tend to have higher distance metric values. This skewness is most pronounced for partial match type 2.

3.1. Threshold determination

In order to determine appropriate threshold value(s) we first needed to specify the key measures of quality for our evaluation. We defined several levels of agreement between the administrative household and the true census household as follows:

- *Perfect match*: The administrative household exactly matches the census household.
- *Composition match*: The administrative and census households have the same household composition.
- *Near match*: The number of persons in the administrative household is within 1 of the number of persons in the census household and the household composition matches.
- *Non-match*: No person in the administrative household matches the census household.

Subsequently, we considered the measures of quality described in Table 2. These measures were calculated at the national level as well as for a few important domains of interest.

We specified thresholds by percentile for each geographical region, province or territory, and by minimum age of the household members according to administrative data. For households with a minimum age of 0–64 years, the threshold was set as the 75^{th} percentile and for households with a minimum age of at least 65 years, the 40^{th} percentile was used as the threshold. A lower threshold was specified for older households due the presence of lower specificity for this population in preliminary analyses.

Table 2
Measures of quality for dwelling-level administrative data

Measure of quality	Definition
Number of eligible dwellings	The number of dwellings below the threshold value
Proportion perfect match	The proportion of eligible dwellings that are perfect matches
Proportion near match	The proportion of eligible dwellings that are near matches
Sensitivity	The proportion of perfect matches that are below the threshold value
Specificity	The proportion of non-matches that are above the threshold value

Table 3
Measures of quality for eligible dwellings

Measure of quality	
Perfect match	74.3%
Near match	91.3%
Sensitivity	91.6%
Specificity	56.2%

Table 3 displays the measures of quality for the chosen thresholds for all eligible dwellings. Overall, the proportion of eligible dwellings that are perfect matches is reasonably high at 74%. Approximately 90% of the eligible dwellings are near matches (includes perfect matches) meaning that the administrative household is the same or similar in composition to the census household. Furthermore, the vast majority (91.6%) of true perfect matches are considered eligible and over 50% of the true non-matches are correctly excluded from the set of eligible dwellings.

3.2. Assessing fit for use

The statistical contingency plan used dwelling-level administrative data for direct imputation within the existing edit and imputation process. Here, direct imputation refers to the process of imputing certain demographic variables, primarily age and sex at birth, using administrative data for non-respondent households. Importantly, this plan would only be implemented in a scenario were the use of administrative data was deemed likely to provide more accurate results than the existing edit and imputation process alone. Due to time constraints and the complexity of the imputation process, only a single iteration involving one pre-specified non-response scenario could be fully implemented and evaluated. Therefore, a comprehensive simulation study was not feasible, necessitating the development of an alternative methodology to determine scenarios under which it is advantageous to use direct imputation in place of traditional donor imputation.

It is essential to maintain the age distribution during the edit and imputation process. As such, this served as the basis for our assessment of the household model approach, relative to traditional donor imputation. We simulated a non-response scenario in which late respondents to the 2016 Census, defined as households who responded after June 15$^{\text{th}}$, were considered non-respondents, corresponding to a situation in which non-response follow-up was terminated early. In order to evaluate the potential differences in the use of direct versus donor imputation we compared the age distributions for:

- Eligible dwellings who were late respondents using the age variable from the Census response database (RDB)
- Eligible dwellings who were late respondents using the age variable from the administrative data
- Early respondents using the RDB

We summarized the differences in the age distributions between the RDB, administrative data and early respondents, who can be viewed as potential donors, using a chi-square difference measure:

$$D = \sum_{l} \frac{(q_l - \widehat{q}_l)^2}{q_l}$$

where l indexes the age groups, q_l is the true proportion of late respondents from eligible dwellings in age group l, and $\widehat{q}_l$ is the proportion of late respondents from eligible dwellings in age group l estimated from either the corresponding administrative data or the early respondents.

Table 4 summarizes the age distributions and difference measure for late respondents. Overall, the age distribution of the administrative data is closer to the true distribution as reported on the RDB than that of the potential donor pool. In particular, the early respondents tend to be older than the late respondents. These results indicates that direct imputation should better preserve the age distribution compared to donor imputation in this non-response scenario.

However, the performance of a given imputation method depends on the response mechanism. Next, we considered the best case scenario for donor imputation, completely random non-response. We randomly set 7% of the dwellings as non-respondents. This is comparable to the level of non-response that was observed at June 15$^{\text{th}}$ for the 2016 Census. In this scenario (not shown),

Table 4
Difference measure for late respondents

	Late respondents in eligible dwellings		Early respondents
	Reported data RDB %	Administrative data %	Donor pool %
0–4	6.69%	7.32%	5.37%
5–17	18.51%	18.44%	14.71%
18–29	15.72%	16.51%	14.52%
30–64	46.50%	49.75%	48.54%
65–79	5.38%	5.74%	12.93%
80+	1.93%	2.24%	3.51%
Missing age	5.27%	0.00%	0.42%
Difference measure (D)		0.0040	0.1309

Table 5
Quality measures for eligible dwellings identified using 2016 thresholds and adjusted thresholds

	2016	Preliminary 2021	Preliminary 2021 with adjustment
Perfect match	74.3%	66.1%	71.6%
Near match	91.3%	89.3%	92.1%
Sensitivity	91.6%	93.8%	89.4%
Specificity	56.2%	35.5%	48.8%

Table 6
Proportion of perfect matches by minimum age of administrative household

Minimum age of administrative household	2016	Preliminary 2021	Preliminary 2021 with adjustment
0–17 years	72.1%	60.7%	67.2%
18–29 years	56.0%	46.1%	52.4%
30–64 years	75.6%	66.6%	71.7%
65–79 years	93.7%	89.2%	89.2%
80 years or older	90.6%	86.0%	86.0%

the age distributions of the administrative data and the potential donor pool are very similar. This indicates that either imputation method should preserve the age distribution.

4. Adaptive implementation: 2021 Census

Prior to implementing the household model approach for the 2021 Census, it was necessary to evaluate the performance of this approach using up-to-date data. The parameters for the household model were specified using data from the previous Census and, importantly, prior to the COVID-19 pandemic which greatly impacted the daily lives of Canadians. This evaluation was executed during the collection period using a preliminary version of the RDB while the non-response follow-up and verification processes were still ongoing. Nevertheless, we were able to evaluate our proposed approach for identifying households with good quality data using recent data, reflecting the Canadian population during the pandemic.

Table 5 displays the various quality measures for eligible dwellings for the 2016 Census (first column) and for the 2021 Census (second column) using the thresholds proposed in Section 3. There is a notable decrease in the proportion of perfect matches and the specificity when the household model approach is applied to the preliminary 2021 data.

However, the magnitude of this decrease is not uniform across different types of households. The proportion of perfect matches by minimum age of the administrative household for 2016 and 2021 Censuses are shown in columns 1 and 2 of Table 6. The decrease in the proportion of perfect matches is more pronounced for younger households. In particular, the proportion of perfect matches remains above 85% for households with a minimum age of 65 years or older but drops below 50% for households with a minimum age of 18–29 years old.

Within the collection period, it was not feasible to rerun the entire household model, particularly refit the person-place and household composition models, using the 2021 preliminary data. However, the flexibility of the distance metric allowed us to easily change the threshold specifications to suit our data requirements. For the 2021 Census, we lowered the threshold for administrative households with a minimum age of 0–64 years from the 75th percentile to the 65th percentile. The results for the measures of quality are given in column 3 of Table 5 and the proportion of perfect matches by minimum age is given in column 3 of Table 6. This adjustment to the thresholds yields data of a similar quality as that obtained for the 2016 Census data in Section 3. As a result, we excluded proportionally more large households with 6 or more persons and more complex households than initially anticipated. Even so, only 4.8% of the eligible true perfect matches under original thresholds were excluded with the threshold adjustment. Of the 15.40 million dwellings with administrative data available, 9.23 million dwellings were

below the final threshold and eligible for use in the contingency plan.

As part of the 2021 Canadian Census of Population, administrative data was used for the direct imputation of number of residents, age and sex at birth for non-respondent dwellings in geographic areas with lower response rates [7]. Specifically, direct imputation using administrative data was implemented at a detailed geography level where response rates were less than 90% and only for dwellings where good quality administrative data was available, as defined by the methodology detailed here. In total, approximately 12,000 non-responding households were imputed using administrative data which corresponds to less than 0.1% of occupied private dwellings in Canada.

5. Conclusion

We have presented a modeling approach for identifying high quality household administrative data for use in direct imputation. This flexible approach allows us to incorporate multiple statistical models and tailor the definition of "good quality data" to fit the intended use of the administrative data. We extended existing approaches in two important aspects. First, the inclusion information on person-level links into the household composition model, resulting in a stricter definition of a household match. Second, the use of a penalty term in the distance metric to limit the overrepresentation of single person households within the pool of households eligible for direct imputation.

Furthermore, we have provided a framework for evaluating the resulting direct imputation, relative to donor imputation, under different response mechanisms in the absence of a simulation study. This evaluation can be conducted quickly with relatively little computational requirements. As such, it is suitable for use within a production environment as illustrated through its use for the 2021 Canadian Census of Population.

Additional analyses indicated that some overcoverage, particularly for older persons, is expected due to limitations of the administrative data such as timeliness and potential differences in mailing address and usual place of residence. This underlines the importance of understanding the source of administrative data and careful consideration of the appropriateness of its use in a given context. Here, administrative data were used for non-respondents after non-response follow-up was completed and in scenarios where the use of administrative data was deemed likely to provide more accurate results than the existing edit and imputation process alone. Future work is planned to assess additional uses of administrative data within the Census, and Statistics Canada is researching the possibility of combined census options whereby administrative data would be used more extensively and earlier in Census collection, similar to the approach of other countries, yet recognizing the particularities of the Canadian context. Additional research into adjustments to other imputation methodologies is also under development.

Further, within the database of administrative persons and the household model methodology, not all persons could be linked to an exact address. Ongoing research examines the possibility of extending the person-place model to a higher level of geography, similar to the mesh-block approach used by Statistics New Zealand [4].

Acknowledgments

The author thanks Karelyn Davis, Arthur Goussanou and Thomas Yoon for their many contributions to the household model project. She also thanks Michelle Simard for her support of this work and for her constructive comments and suggestions.

References

[1] Blackwell L, Charlesworth A, Rogers NJ. Linkage of census and administrative data to quality assure the 2011 census for England and Wales. Journal of Official Statistics. 2015; 31(3): 453–73.

[2] Bycroft C. Census transformation in New Zealand: Using administrative data without a population register. Statistical Journal of the IAOS. 2015; 31(3): 401–11.

[3] Morris DS, Keller A, Clark B. An approach for using administrative records to reduce contacts in the 2020 Decennial Census. Statistical Journal of the IAOS. 2016; 32(2): 177–88.

[4] Bycroft C, Matheson-Dunning N. Use of administrative records for non-response in the New Zealand 2018 Census. Statistical Journal of the IAOS. 2020; 36(1): 107–16.

[5] Keller A, Mule VT, Morris DS, Konicki S. A distance metric for modeling the quality of administrative records for use in the 2020 US Census. Journal of Official Statistics. 2018; 34(3): 599–624.

[6] Morris DS. A modeling approach for administrative record enumeration in the decennial census. Public Opinion Quarterly. 2017; 81(S1): 357–84.

[7] Statistics Canada. Guide to the Census of Population, 2021 [Internet]. Ottawa (CA): Statistics Canada; 2022 [updated 2022 Feb 8; cited 2022 Sept 25]. Available from: https://www12.statcan.gc.ca/census-recensement/2021/ref/98-304/2021001/app-ann1-7-eng.cfm