SSC Annual Meeting, May 2022
Proceedings of the Survey Methods Section

Modernization of the Canadian Census: An Administrative Data-Driven Approach to Defining Households

Thomas Yoon[1], Karelyn Davis, Erin Lundy, and Arthur Goussanou

ABSTRACT

Many national statistical offices are conducting research to better utilize administrative records, defined as data collected as part of administering a program or service. Administrative records offer the possibility to complement the traditional survey enumeration approach and potentially improve quality and efficiency in estimation. A combined census is currently under research at Statistics Canada whereby administrative data and traditional data collection are used jointly to enumerate the population. One part of ongoing census research is the household model, which aims to group administrative individuals into ‘’households’’ using statistical models, and to evaluate their quality as compared to traditional census outputs. The paper will host the methodology and the evaluation of the key quality indicators of the household model approach.

KEY WORDS: administrative data, combined census, population estimation, Canadian Census of Population, multinomial logistic model.

RÉSUMÉ

De nombreux bureaux nationaux de statistiques mènent des recherches pour mieux utiliser les données administratives, définis comme des données recueillies dans le cadre de l'administration d'un programme ou d'un service. Les données administratives offrent la possibilité de compléter l'approche traditionnelle de collecte par enquête et d'améliorer potentiellement la qualité et la précision de l'estimation. Un recensement combiné est actuellement à l’étude à Statistique Canada dans le cadre duquel les données administratives et la collecte traditionnelle de données sont utilisées conjointement pour dénombrer la population. Une partie de la recherche en cours sur le Recensement est le modèle des ménages, qui vise à regrouper les individus administratifs en « ménages » à l'aide de modèles statistiques, et à évaluer leur qualité par rapport aux résultats traditionnels du Recensement. L'exposé présentera la méthodologie et l'évaluation des indicateurs clés de qualité du modèle des ménages.

MOTS CLÉS : données administratives; recensement combiné; estimation de la population, Recensement de la population canadienne, modèle logistique multinomial.

1. INTRODUCTION

1.1 Background

Every five years, Statistics Canada conducts the Canadian Census of Population to collect demographic and social information vital for planning governmental services and analyzing demographic trends (Census of Population (statcan.gc.ca)). The Census of Population enumerates the entire population of Canada on Census Day, the most recent being May 11, 2021. From a statistical standpoint, the Census in-scope population includes Canadian citizens (by birth or by naturalization); landed immigrants and non-permanent residents and family members residing with them; and Canadian citizens temporarily out of Canada on Census Day. Each household, defined as the collection of individuals living in the same address containing at least one Census in-scope person is legally required to complete the census. The Canadian Census is seen as a traditional census, whereby respondents answer Census questions directly via a questionnaire.

While traditional enumeration has worked very well in the past, many countries around the world have noted challenges pertaining to traditional enumeration, (Skinner, 2018) from increased costs for non-response follow-up (NRFU) to rapidly changing migration patterns, and most recently the public health restrictions placed on traditional collection due to the COVID-19 pandemic and natural disasters (e.g., forest fires). In Canada, research into a combined census begun prior to the COVID-19 pandemic was used to study the statistical integration of administrative data into traditional census operations

[1] All authors, Statistical Integration Methods Division, Statistics Canada, Ottawa, Canada, K1Y 0T6; thomas.yoon@statcan.gc.ca

to mitigate the aforementioned risks. In particular, the impact of post-collection imputation activities was studied and results are presented in this paper.

1.2 Benefits and Drawbacks of Administrative Data

Administrative data are data generated during the course of an administrative operation and then retained in a database (Hand, 2018). One example is a Canadian citizen filing his or her income tax each year. The filing of the income tax demands more than just his or her earned income; it also asks for his or her address, birthday and Social Insurance Number (SIN). Statistical data integration of administrative data is achieved by linking different sources of administrative data at the unit level- for example, an organization or individual- or at the micro level- for example, a small geographical area- to compile and organize information that was traditionally collected from manual censuses, for statistical and research purposes (Telford, 2017).

In addition to individual population enumeration, household-level characteristics are important concepts to obtain. However, there are considerations to be made when using administrative data to form households in the Census context since administrative data begins with individual information and households must be 'created' as a result. Moreover, administrative data was created for a different purpose than traditional Census collection, thus require a different methodology for population and household enumeration. Broadly, there are three main considerations: Extraction of Administrative Data, Grouping using Administrative Relationships, and Matching to the Correct Address in Administrative Data. Provided below is a brief description of and the potential issues of each step.

1. Extraction of Administrative Data: Individuals are identified and extracted from various sources of administrative data. It is imperative to accurately remove the non-Census target population such as foreign residents and recent deaths.
2. Grouping using Administrative Relationships: In this step, individuals who appear on family relationship administrative data are grouped together into a household unit. Double counting of same people must be avoided. For example, children under multiple custody (with divorced parents, etc.) may appear in multiple household units.
3. Matching to the Correct Address in Administrative Data: The household unit formed in the previous step is assigned an address in the administrative geography database. As some administrative records have poor address quality, statistical models are needed to rank administrative households for future Census use, so as to avoid removal of otherwise high-quality household units.

For the Canadian Census, these steps formed the methodology of what is known as the 'Household Model' which forms part of the modernization of the Canadian Census of Population by complementing some of these problems through the use of administrative data. The remainder of this paper discusses the household model and demonstrates its use in the 2021 Census of Population imputation activities.

2. METHODOLOGY

The Household Model forms household units from administrative data based on statistical models to place administrative individuals into the most probable address at the time of Census. It consists of three main stages: Creation of the Person-Address File, the Person-Place Model, and Household Composition Model. In this paper, the methodology and results of the Household Model for the 2016 Census will be discussed, along with an application to the 2021 Census.

2.1 Creation of the Person-Address File

Various sources of administrative data from 2011 to 2016 were linked probabilistically to create the Person-Address file, similar to an approach used by the US Census Bureau and Statistics New Zealand (Morris, 2017 and Bycroft and Matheson-Dunning, 2020). The sources include tax files (T1 personal master file, miscellaneous tax files), Canada Child Tax Benefit, various Pension Plan files, immigration files, some provincial driver's license files and the Indian Register. The file was formatted so that each observation represents a unique combination of person-address pair.

2.2 Person-Place Model

The Person-Place Model takes in all different person-address pairs from the previous step, and aims to predict a probability of a match between the administrative address and the Census address for each individual. The Person-Place Model employs a forward step-wise logistic regression model with response variable as follows:

$$y_{ih} = \begin{cases} 1 \text{ if person } i \text{ is found in admin data and 2016 Census at the same address } h \\ 0 \; otherwise \end{cases} \quad (1)$$

The predictors can be subdivided into two categories:

1. Individual attribute variables include age, sex at birth, indigenous status, source of address (which indicates the administrative source the address was extracted from), mobility status based on postal code, multiple address status (which indicates whether the person appeared in more than one address since 2011), and total number of administrative data sources from which the person-address pair was extracted.
2. Dwelling attribute variables include urban area status, province, collection area (which indicates whether the dwelling is a mailable, list-leave, canvasser area or other types of address), and dwelling type (which specifies whether the dwelling is a house, apartment or other type of building structure).

For each person-address pair, the Person-Place Model provides an estimate of the coherence probability, $\hat{p}_{ih}$where $p_{ih} = P(y_{ih} = 1)$. Administrative households are then formed as follows:

1. if the administrative person, *i*, has only one address, select the address as their usual place of residence
2. if the administrative person, *i*, has multiple addresses, select the address with the highest $\hat{p}_{ih}$as their usual place of residence
3. The administrative household, *h*, is formed by grouping individuals whose usual place of residence is *h*.

The creation of administrative households allows for the analysis at the household level.

2.3 Household Composition Model

The purpose of the Household Composition Model is to assess the coherence of administrative households with Census households through multinomial logistic regression with LASSO covariate selection of four outcome levels. The outcome levels, labelled as CL (coherence level), are distinguished by degree of match in three criteria: administrative household member matching to Census household member, counts of people in the household, and the household composition. The household composition mentioned below refers to whether the household includes at least one child. The CL levels are described as follows:

- Perfect Match, ($CL_h = 1$), occurs when the administrative household exactly matches the Census household in all three criteria; 60.58% of households from 2016 Census belonged to this category.
- Type 1 Partial Match, ($CL_h = 2$), is when at least one administrative person matches the Census household, the administrative household count ≥ Census count, and the household composition matches; 21.42% of 2016 Census households were Type 1 Partial Matches.
- Type 2 Partial Match, ($CL_h = 3$), is when at least one administrative person matches the Census household, and either administrative household count < Census count or the household composition does not match; this group corresponded to 9.75% of 2016 Census households.
- No Match, ($CL_h = 4$), occurs when there is no administrative person matching to the Census household, covering about 8.26% of 2016 Census households.

The explanatory variables used for the Household Composition Model include age (minimum age, maximum age, proportion of people in various age groups for each dwelling), administrative sources of the address, number of addresses, number of Census out-of-scope individuals assigned to the address, geography, dwelling type of the address, and historical family relationships derived from administrative data. The Household Composition Model produces a predicted probability for each coherence level as follows:

$$\hat{p}_h^{CL=i} = \begin{cases} \frac{e^{\hat{\beta}_i X_h}}{1 + \sum_{k=1}^{3} e^{\hat{\beta}_k X_h}} \; for\; i \in \{1,2,3\} \\ \frac{1}{1 + \sum_{k=1}^{3} e^{\hat{\beta}_k X_h}} \; for\; i = 4 \end{cases} \tag{2}$$

2.4 Distance Metric

To determine the overall quality of administrative data as a function of the Household Composition Model and the Person-Place Model, a Euclidean distance metric variable was created (adapted from Keller et al, 2018):

$$d_h = \sqrt{(1 - \hat{p}_h^{PP})^2 + (1 - (\hat{p}_h^{HH})^{e_h})^2} \tag{3}$$

where:

- $\hat{p}_h^{PP}$ is the minimum predicted probability from the Person-Place Model of all people placed at address *h*
- $\hat{p}_h^{HH}$ is the predicted probability from the Household Composition Model at address *h* for perfect match
- e_h is a penalty function to correct the over representation of single person households. As such, e_h is 1 for single person households and 0.5 for bigger households.

2.5. Implementation in Census 2021 Whole Household Imputation

Since the impact of the pandemic on Census response rates was unknown, the use of administrative data was earmarked for unit imputation of Census 2021 non-responding households after other collection activities had ceased. In particular, administrative households of high quality were given priority in the Whole Household Imputation (WHI) process, which aims to impute occupancy status, household information and short-form variables (e.g., household size, age, sex at birth, gender, language, etc.) for non-responding dwellings. The existing WHI methodology is based on control totals provided by estimates of the Dwelling Classification Survey (DCS), a survey of non-respondent dwellings performed later in the collection period. For the WHI process, administrative data was used to impute household size, age and sex at birth while other short-form characteristics were imputed using donor imputation. Note that an administrative record was permitted to act as a donor for other non-responding households. While other approaches have investigated the use of administrative data to inform donor imputation (Farnell and Darby, 2020), our approach considers the formation of administrative households within the imputation methodology.

3. RESULTS

3.1 Person-Place Model

The Person-Place Model's goal is to predict the probability of a match between the administrative address and the Census address for each individual. More Specifically, separate logistic regression models were built for provinces and territories. The provinces model used 1% of addresses as the training set. This accounted for roughly 487,600 individuals from 129,600 addresses. The analogous number is 20% for the territories model from 15,000 individuals and 3,800 addresses. Table 1 illustrates the results of the two models from 2016 Census. The administrative address was considered a match if the estimated probability is greater than 0.5. The false negative rate refers to the proportion of people whose administrative address matched to Census address but was predicted to be a mismatch. The false positive rate refers to the proportion of people whose addresses mismatched but was predicted to be a match.

Table 1. Person-Place Model Accuracy Measures

	Accuracy (%)	False Negative Rate (%)	False Positive Rate (%)
Provinces	85.77	8.89	22.94
Territories	85.66	20.19	11.14

Some machine learning algorithms including the Classification Tree and the Random Forest model were fit. Due to the limitation in the size of the training set, the models were constructed only on the provinces training set. R programming's rpart and randomForest packages at their default values were used to construct the models. In the case of the classification tree model, the tree was pruned until the first addition of an insignificant variable. Table 2 compares the accuracy of the three models.

Table 2. Model Comparisons of Person-Place Model

	Logistic Regression	Classification Tree	Random Forest
False Negative Rate (%)	8.89	8.52	8.08
False Positive Rate (%)	22.94	25.63	21.69
Sensitivity (%)	91.11	91.48	91.92
Specificity (%)	77.06	74.37	78.31
Accuracy (%)	85.77	84.85	86.79

The logistic regression model was used for future analyses as the benefit of being able to interpret the coefficient estimates outweighed the cost of having 1.02 percentage points lower accuracy than the random forest model.

3.2 Household Composition Model

The Household Composition Model estimates the coherence between the administrative households and the Census households. Similar to the Person-Place Model, separate models were created for provinces and territories. The training set for provinces and territories accounted for 2.5% and 33%, respectively, of private dwellings. Figure 1 shows the result of the Household Composition Model on provinces for all four predicted response levels, by $\hat{p}_h^{CL=i}$. The red area indicates correctly classified addresses while the blue represents incorrectly classified addresses.

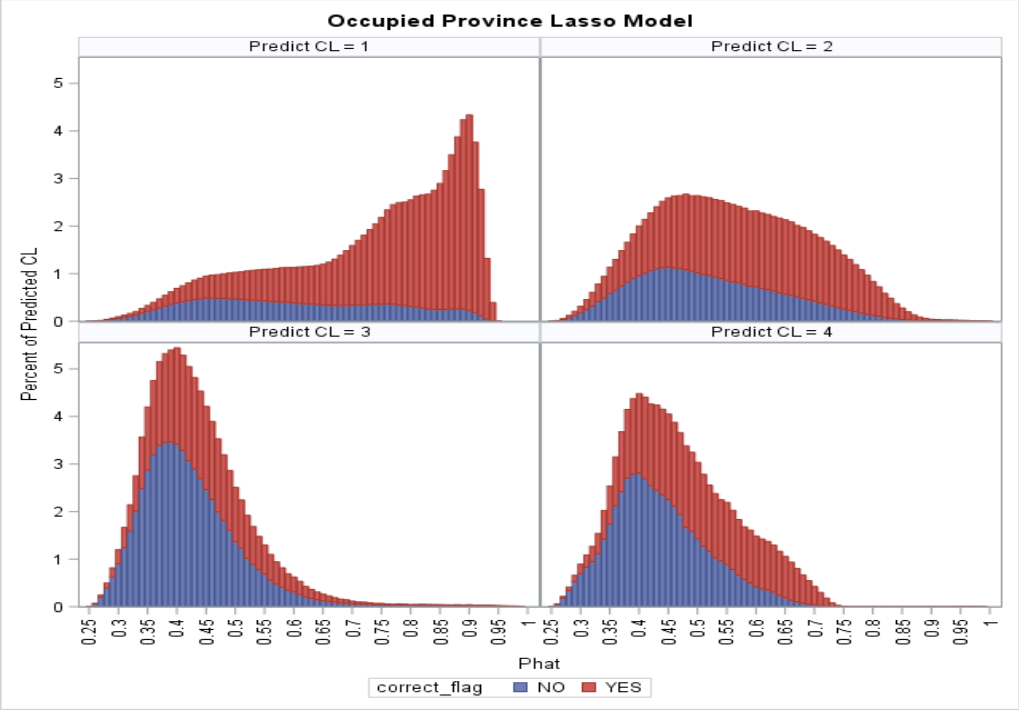

Figure 1. Household Composition Model Result on Provinces

3.3 Distance Metric

Combining the results of the Person-Place Model and the Household Composition Model, the distance metric provides a means to evaluate the quality of both models for each address. Figure 2 depicts the histogram for each true outcome level

by d_h. The dotted vertical line represents the 75th percentile of distance metric, which was used as a preliminary threshold. The addresses whose distance metric is lower than the threshold are considered eligible for administrative data imputation.

Table 3 illustrates the overall accuracy of the Household Composition Model and distance metric on either side of the threshold. A 'Near match' corresponds to households where the composition matched and the number of people differed within count of 1. The sensitivity refers to the proportion of perfect matches below the distance metric threshold, while the specificity refers to the proportion of non-matches above the threshold.

Table 3. Household Composition Model and Distance Metric Accuracy Measures

Measure of Quality	Perfect Match	Near Match	Sensitivity	Specificity
Percent (%)	74.1	90.49	95.3	51.8

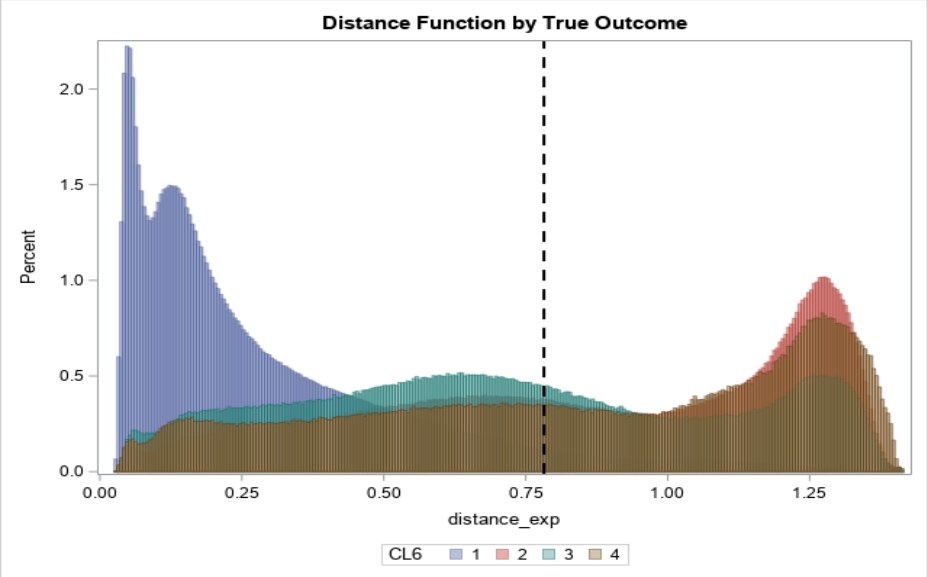

Figure 2. Distance Metric by True Outcome Levels

3.4. Simulation of Increased Non-Response for Whole Household Imputation

In the spring of 2021, to assess the usage of eligible administrative dwellings in the WHI process, an empirical study was conducted using information from late-Census 2016 respondents. The aim of the study was to better understand non-response in a pandemic context, and dwellings representing the last 10% of Census respondents were considered as non-respondents for the purpose of simulation. Furthermore, additional non-respondents were targeted in certain areas where public health restrictions would have prevented traditional in-person NRFU activities.

Respondents in private dwellings were randomly set to be non-respondents and administrative eligible dwellings were investigated for these areas. Where a high-quality administrative dwelling existed, the household size, age and sex of respondents were imputed using administrative data as opposed to donor imputation. Overall, three scenarios were considered for these non-responding households: Scenario 0 (Census 2016 reference - Respondents and donor imputation); Scenario 1 (Donor imputation) and Scenario 2 (Imputation using high quality administrative households where available). Table 4 displays a comparison of the bias, using the mean absolute percent error (MAPE) between Scenario X and the reference scenario, where X = 1,2.

Table 4. Comparison of Mean Absolute Percent Error (MAPE) of Population counts for Scenario 1 and Scenario 2 by Collection Method and Average Response Rates

Collection Method	Response rate	Number of Collection Units (CUs)	MAPE Scenario 1	MAPE Scenario 2
Mail-out	1-RR 0% to 79%	2,690	7.42%	4.42%

	2-RR 80% to 84%	2,912	3.71%	2.59%
	3-RR 85% to 89%	6,125	2.62%	1.91%
	4-RR 90% to 94%	11,675	1.69%	1.22%
	5-RR 95% +	12,452	0.93%	0.76%
Non Mail-Out	1-RR 0% to 79%	2,076	10.73%	9.81%
	2-RR 80% to 84%	1,009	4.11%	4.09%
	3-RR 85% to 89%	1,629	2.73%	2.70%
	4-RR 90% to 94%	2,659	1.93%	2.04%
	5-RR 95% +	4,669	1.20%	1.71%

Scenario 2 noted the greatest improvement in MAPE when the response rate was less than 80% and for mail-out areas, which represent the vast majority of dwellings and where address quality is of the highest. While little to no improvement was noted for non-mail-out areas, the subgroup is more difficult to enumerate in general and is the subject of future research.

4. CONCLUSION

Regardless of the high response rate of 2021 Census at 98%, the administrative data imputation mentioned previously was triggered to ensure high-quality population and dwelling counts in areas where census collection was affected by Covid-19, a natural disaster, or low response rates. Approximately 12,000 non-responding households from 1,045 collection units accounting for about 0.1% of private dwellings were imputed using administrative data. For more information, please refer to https://www.census.gc.ca/census-recensement/2021/ref/98-304/2021001/app-ann1-7-eng.cfm.

The Household Model is currently limited to individuals with detailed level of geography (i.e., dwelling). Even with the presence of administrative signal at a broader geographic level (i.e., collection unit), the Household Model does not include these individuals if their dwelling addresses are missing. A new model will be created, similar to the New Zealand meshblock approach (Bycroft and Matheson-Dunning, 2020) to provide quality indicators for this situation. In addition, quality indicators are planned to be adapted to incorporate record linkage impacts.

While the 2021 Census of Population had a high response rate at the dwelling level (98%), the methodology developed in this paper provides an alternative using administrative data, which may lead to timelier analyses and potential efficiencies. The Household Model will be at the forefront in modernizing the usage of administrative data to complement traditional Census enumeration.

REFERENCES

Bycroft, C. and Matheson-Dunning, N., 2020, Use of administrative records for non-response in the New Zealand 2018 Census. *Statistical Journal of the IAOS* **36(1)**, pp 107-116.

Farnell, J., Darby, P. 2020. Administrative data informed donor imputation in the Australian Census of Population and Housing. *Statistical Journal of the IAOS* 36, pp. 117-124.

Hand, D.J. (2018). "Statistical Challenges of Administrative and Transaction Data". *J.R. Statist. Soc. A.* **181 (3)**, 555-605.

Keller, A., Mule, V.T., Morris, D.S. and Konicki, S. (2018). "A Distance Metric for Modeling the Quality of Administrative Records for Use in the 2020 U.S. Census". *Journal of Official Statistics*. **34 (3)**, 599-624.

Morris, D.S., 2017, A modeling approach for administrative record enumeration in the decennial census. *Public Opinion Quarterly* **81(S1)**, pp 357-384.

Skinner, C. (2018). "Issues and Challenges in Census Taking". *Annual Reviews of Statistics and Its Applications*. **5**,49-63.

Statistics Canada. (2022, November 18). *Census Program: Census of Population*. Retrieved November 21, 2022, from https://www12.statcan.gc.ca/census-recensement/index-eng.cfm

Telford, J., Araghi, R. and Samson, P. (2016). “Modernization Processes in National Statistical Offices – Transforming the Australian Bureau of Statistics”. *IAOS 2017-2019*.