IAOS-ISI 2024 MEXICO CONFERENCE

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

Erin Lundy

IAOS-ISI Mexico City

May 16th 2024

INEGI

International Statistical Institute

IAOS IMPROVING OFFICIAL STATISTICS

Outline

- Context
- Household model approach
- Model development using 2016 Census data
- Adaptive implementation in the 2021 Census
- Future work

Context

- In 2020, long term research agenda towards use of administrative data in a combined census approach already underway at Statistics Canada.
- The COVID-19 pandemic accelerated exponentially research related to use of administrative data in the Canadian Census of Population.
- Statistics Canada developed a statistical contingency plan to mitigate for potentially lower response rates.

Context

- Planned to use administrative data to impute non-responding households in areas with a low response rate and where administrative data was of sufficient quality.
- Adapted existing modeling approach to identify households with good quality data.
- Contingency plan would be implemented in a scenario where the use administrative data was deemed likely to provide more accurate results than existing edit and imputation process alone.
- Developed framework to evaluate direct imputation using administrative, relative to donor imputation, in the absence of a comprehensive simulation study.

Household model approach

- 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.
- Used a modeling approach to create administrative households and rank the quality of the available administrative data for these households.
- Consists of three components: person-place model, household composition model and distance metric.

Household model approach

- Basis of the household model is database of administrative persons, created for the sole purpose of the Census research.
- This database includes a variable predicting if an administrative person is in-scope for the Census, the person's age and sex at birth.
- As well, contains auxiliary data from a variety of administrative data sources such as tax files, immigration files and vital statistics.
- Some sources include detailed address information.
- List of unique person-address pairs which includes all possible addresses was created.

Person-place model

- Predicts the probability that an administrative person is observed at the correct dwelling using logistic regression model:

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

- For each person-address pair, we obtain a person-level estimated probability of coherence $\hat{p}_{ih} = P\left(y_{ih}^{PP} = 1\right)$.
- For a person with administrative records at more than one dwelling, we assign the address with the highest probability $\max_h \hat{p}_{ih}$.
- Administrative households defined as all persons assigned to a given dwelling
- For each dwelling *h*, we define the dwelling-level estimated probability of coherence:

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

Household composition model

- Predicts the probability that an administrative household matches the household observed in the Census of Population.
- Categorized into four levels of coherence:
  1. Perfect match
  2. Partial match type 1 – at least one administrative person matches, admin count is greater or equal to census count and composition matches
  3. Partial match type 2 – at least one administrative person matches, admin count is less than census and/or composition does not match
  4. Non-match
- Model probabilities of coherence levels using multinomial logistic regression.

Distance Metric

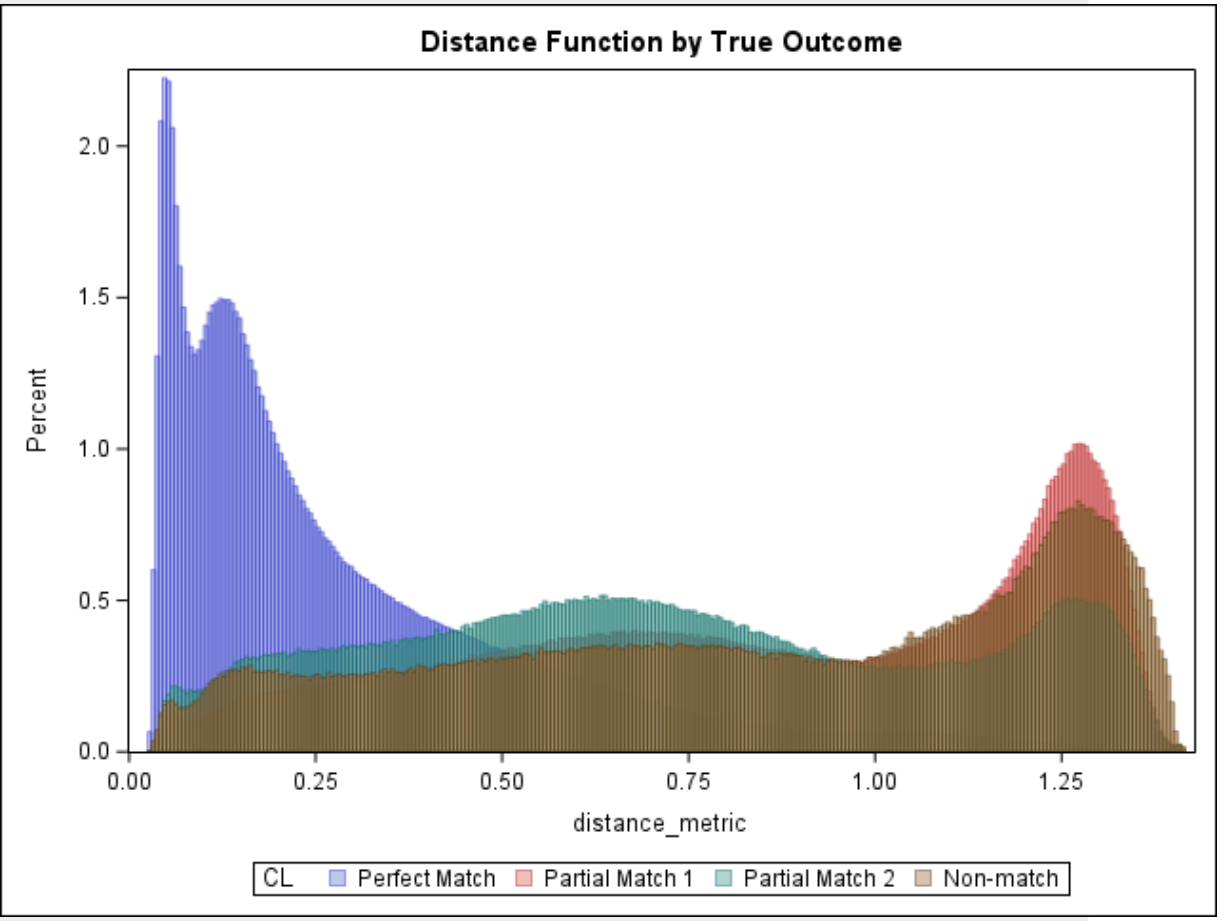
- Incorporate dwelling-level estimated probability of coherence and probability of perfect match into one measure of dwelling-level quality.
- Use extension of Euclidean distance-based function (Keller et al, 2018) with penalty term for administrative household of size 1:

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

$\hat{p}_h^{PP}$ 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.

Penalty term $e_h = 1$ for households with $n_h = 1$ and $e_h = {}^1/_2$ otherwise.



- Evaluated household model approach using data from the 2016 Census.
- Models fit using auxiliary data that reflects the vintages available prior to 2016 Census.
- Majority of dwellings with a low distance metric value are perfect matches.
- Distribution for partial matches and non-matches are left skewed.
- Skewness most pronounced for partial match type 1.

Threshold determination

- All dwellings below specified threshold(s) deemed to be good quality.
- Based on key measures of quality:
  1. Proportion of true perfect matches
  2. Proportion of near matches (count within 1 and composition match)
  3. Sensitivity
  4. Specificity
- Specified thresholds based on percentiles for each geography region and by minimum age of administrative household members.
- Used 75th percentile for households with minimum age 0-64 years and 40th percentile for households with minimum age 65+ years.
- Resulted in 74.3% perfect match, 91.3% near match, 91.6% sensitivity and 56.2% specificity.

Assessing fit for use

- Not operationally feasible to conduct comprehensive simulation study.
- Developed alternative methodology for this evaluation based on age distribution.
- Simulated non-response scenario in which late respondents to the 2016 Census were considered non-respondents.
- Compared age distributions for:
  - Eligible dwellings who were late respondents using Census data
  - Eligible dwellings who were late respondents using admin data
  - Early respondents using Census data (potential donors)
- Summarized differences using chi-square difference measure:

$$D = \sum_l \frac{(q_l - \hat{q}_l)^2}{q_l}$$

	Late respondents in eligible dwellings		Early respondents
	Reported data Census %	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

Adaptive implementation in the 2021 Census

- Household model performance evaluated with most recent administrative data during the collection period using a preliminary version the 2021 Census database.
- Notable decrease in proportion of perfect matches and specificity when models fit using 2016 data applied to preliminary 2021 data.
- Decrease more pronounced for younger households.
- Not feasible to refit statistical models using 2021 preliminary data during the collection period.
- However, can easily change the threshold specifications.
- Lowered threshold from 75^{th} to 65^{th} percentile for households with a minimum age of 0-64 years.

Adaptive implementation in the 2021 Census

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

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

Adaptive implementation in the 2021 Census

- Of the 15.40 million dwellings with administrative data available, 9.23 million dwellings were below the final threshold.
- In the 2021 Canadian Census of Population direct imputation using administrative data was implemented in geographical areas with response rates less than 90%.
- Approximately 12,000 non-respondent dwellings imputed with administrative data.

Future work

- Extension of person-place model to a higher level of geography to incorporate administrative persons not linked to an exact address.
- Continued research to assess additional uses of administrative data within the Census.
- Possibility of combined census where administrative data would be used more extensively and earlier in the Census collection.
- Evaluation of the impact of use of administrative data on the coverage and demographic estimates.

IOS-ISI 2024 MEXICO CONFERENCE

Thank you Merci

For more information, please contact

Erin Lundy
erin.lundy@statcan.gc.ca

Pour plus d'information, veuillez contacter

Erin Lundy
erin.lundy@statcan.gc.ca

INEGI International Statistical Institute IAOS IMPROVING OFFICIAL STATISTICS