Distr.: General
05 October 2017

English

United Nations Economic Commission for Europe

Conference of European Statisticians

Work Session on Migration Statistics
Geneva, Switzerland
30-31 October 2017
Item 3 of the provisional agenda
Data integration and administrative data

2016 Census of Population of Canada: Integration of immigration administrative data

Note by Statistics Canada*

Abstract

An important data source for demographic, social, and economic characteristics of immigrants in Canada, the Census of Population includes questions such as year of immigration, place of birth, and citizenship. After these variables are collected from respondents, Statistics Canada treats the data for missing and inconsistent responses. The 2016 Census of Population took advantage of a record linkage to administrative immigration data in order to improve the quality of the traditional variables and to add new variables associated with admission category (e.g., economic immigrants, refugees). This presentation will provide an overview of how linked administrative data were used to enhance processing (edit and imputation) methods for responses to immigration questions. In addition, the presentation will cover the methods and quality assessment associated with the newly integrated admission category variables.

*Prepared by Scott McLeish, Section Chief, Social and Aboriginal Statistics Division, and Kathryn Spence, Analyst, Social and Aboriginal Statistics Division

I. Background: Immigration questions on the Canadian Census of Population

1. The Canadian Census of Population is one of the main data sources on the socio-economic outcomes of immigrants in Canada. Conducted every five years, the census asks respondents several questions related to immigration including immigrant status, year of immigration, place of birth, place of birth of parents and citizenship. Users can connect these variables with a broad range of topics, including language, education, labour, income, and housing, to meet their analytical needs. Most recently, immigration variables were collected on the 2016 Census of Population with the results planned to be released October 25th, 2017. While immigration questions were not included on the 2011 Census, they were asked as part of the 2011 National Household Survey (NHS), a voluntary large-sample survey under the census program.

A. Processing methods for immigration variables

2. For every census, the questionnaires undergo thorough review and testing prior to collection in order to ensure that the questions are well understood and respondent burden is minimized (Statistics Canada, 2017). Despite these measures, as is the case with any survey, questions are not always answered by respondents and answers provided are not always accurate. Consistency edits are applied to improve coherence between responses and donor imputation methods are used to address item non-response. There are several possible reasons for responses to be missing, inaccurate or inconsistent for the immigration questions. These could include the following examples:

i. Respondent does not understand the question (e.g. year of immigration vs. year of arrival)
ii. Respondent does not remember exact year of immigration
iii. Proxy error (the person filling out the questionnaire does not know the true values for everyone in household)
iv. Capture error (e.g. not scanning year of immigration accurately)
v. Ambiguous responses which could refer to multiple values (e.g. ‘Ireland’)
vi. Changing geographies (e.g. ‘U.S.S.R.’ in 1990 vs. ‘Russia’ in 2011)

The processing methods used for the 2016 Census of Population were not new; most of them had been applied for several censuses. Nearest-neighbour donor imputation was performed using the Canadian Census Edit and Imputation System (CANCEIS) in order to impute these cases with a valid answer. Table 1 provides the imputation rates observed for the 2011 NHS for the immigration variables. Overall, these rates ranged between 2.0% (place of birth) to 12.5% (year of immigration).

Table 1: Imputation rates in the 2011 National Household Survey by immigration variable

Variable	Imputation rate (%)
Citizenship	2.3
Place of birth	2.0
Place of birth of mother	5.7
Place of birth of father	6.0
Immigrant status	1.3
Year of immigration	12.5

Source: Statistics Canada, Place of Birth, Generation Status, Citizenship and Immigration Reference Guide, 2011 National Household Survey

3. Throughout processing, the consistency edits and donor imputation results are monitored closely to ensure that they are applied as intended. The final processed results are then confronted, evaluated and certified using other data sources such as administrative data, previous census cycles and projections.

B. Consistency edits

4. For the immigration questions, consistency edits have been developed and updated over multiple cycles of the census. They are largely based on immigration and citizenship policy in Canada. For example, the immigrant status question asks "Is this person now, or has this person ever been, a landed immigrant?" If a respondent answers "No" to this question but when responding to the question on citizenship replies that they are a Canadian citizen **by naturalization** rather than a Canadian citizen **by birth**, this would reflect an inconsistency; in order to become a citizen by naturalization, one must necessarily first be a landed immigrant. However, based on these two responses alone, it is not evident which question was answered incorrectly. Responses to other questions such as place of birth are used to determine which response must be corrected.

5. Another example of an inconsistency would include individuals reporting having immigrated to Canada in a year prior to their birth year or reporting Canadian citizenship by birth and responding "Yes" to the immigrant status question. For all of these cases, the identified response in error is either assigned to the only possible consistent value or blanked and sent to imputation.

6. The most significant consistency edit applied to the immigration questions is for respondents identifying as Canadian citizens by naturalization, not immigrants, and born abroad. Most of these records are amended to be categorised as immigrants. Since the year of immigration question is only asked of respondents who initially answer "Yes" to the immigrant status question, this means that a valid value of year of immigration will have to be imputed for all these edited cases. The consequence of this effect is that the imputation rate for year of immigration question is much higher than the ones observed for other immigration questions as shown in Table 1.

C. Donor Imputation

7. For records that have missing values (failed records), either as a result of non-response or as a result of editing, nearest-neighbour donor imputation is used. This process uses available and correlated variables (matching fields) to select suitable donors for the missing values. Ideally, a selected donor matches the failed record on every field. However, as this is often not possible, the 'nearest-neighbour' is found by matching on fields as closely as possible. Some matching fields are given more importance than others. For example, when imputing place of birth, as mother tongue is more correlated, it would be more important to match on this than age.

8. Traditionally, when imputing immigration variables, the matching fields have included age, sex, place of residence, mother tongue, ethnocultural characteristics and any completed immigration variables. Because of the important family dynamic of immigration, imputation is stratified to account for these within-household correlations. Couples are imputed together, where possible, and children are imputed taking advantage of information from their siblings and their parents. Decisions regarding which variables to use as matching fields are largely based on past observed associations and subject-matter knowledge.

II. Immigration administrative data in Canada

9. Immigration, Refugees and Citizenship Canada (IRCC), is the federal department responsible for immigration in Canada. They collect a wide range of administrative data related to immigrants landing in Canada. Principally, IRCC collects information at the time immigrants are admitted to Canada from their permanent residency visa application. This information includes the year of immigration, place of birth, and the category under which the immigrants were granted permanent residency (e.g., skilled workers, sponsored family members, refugees, etc.). While this IRCC administrative data provides detailed information on immigrants when they land in Canada, it cannot be used to estimate the population of immigrants living in Canada at a specific point in time as there is no follow-up data collection. However, this information can be linked with other data sets in order to connect the admission characteristics with short or long-term outcomes after landing.

10. Statistics Canada, in collaboration with IRCC, produces the Longitudinal Immigration Database (IMDB), which links an administrative census of immigrants who have landed since 1980 (using the data from IRCC) to annual tax records since 1982 (Evra, 2017). This database permits analysis of socio-economic outcomes of immigrants to Canada on a longitudinal basis, with over 30 years of follow-up for earlier cohorts. In addition to providing a wealth of information related to income, the tax files provide an indication of where immigrants reside within Canada and internal migration patterns. On the other hand, the IMDB does not provide direct measures for other topics included in the Census of Population, such as education in Canada, labour, language and housing. In addition, since the outcomes are restricted to those found in tax files, children and other populations who do not have any fiscal activity are not currently covered.

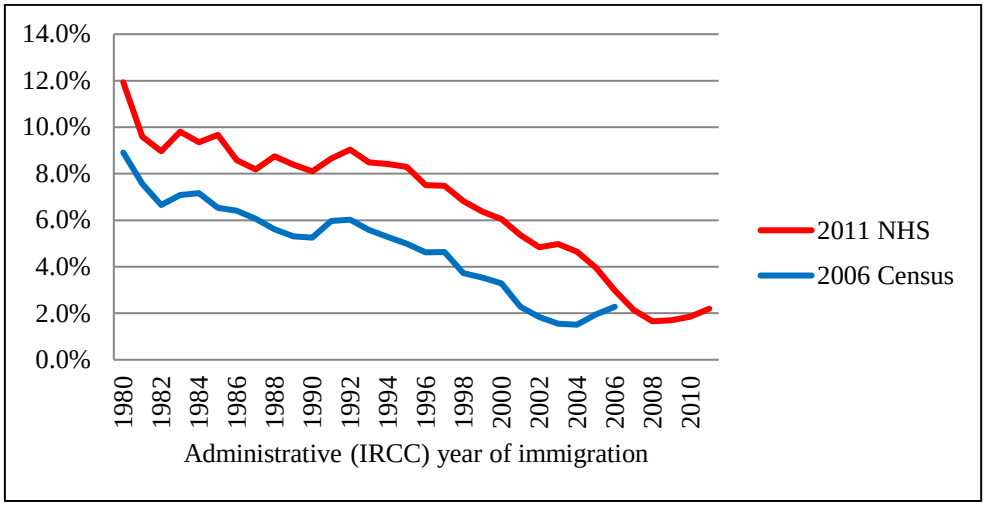
D. Record linkage with census results for data quality evaluation

11. Following a feasibility study involving the 2006 Census, the 2011 NHS was linked with IRCC immigration data for those who landed in Canada since 1980. This was done to assess the data quality of the NHS which employed similar immigration content as previous censuses (Brennan, 2011; Brennan, 2013; McLeish, 2012; McLeish, 2014). While much of the resulting analysis focussed on the effects of global non-response to the voluntary NHS, the linkage also provided an unprecedented examination of measurement errors associated with responses to immigration questions and the effectiveness of existing processing methods including donor imputation.

12. This analysis resulted in some modifications to consistency edits and approaches to imputation, including modifying the imputation of imprecise responses such as "Ireland". Previously, these were mostly imputed using donors with precise responses in the same class (e.g. potential donors for "Ireland" would be restricted to the Republic of Ireland and the United Kingdom). However, the linkage demonstrated that this approach ignored the fact that the probability of providing the imprecise response was not necessarily similar among respondents born in the precise places of birth, even after controlling for matching fields. For example, respondents born in the Republic of Ireland are much more likely to answer "Ireland" than those born in the United Kingdom. As a result of this analysis, linked proportions were used for this imputation instead (Dobson, 2014).

13. One of the important findings of this analysis included the fact that year of immigration was not missing at random. The longer it has been since an immigrant landed in Canada, the more likely they were to answer "No" to the immigrant status question. Chart 1 illustrates this trend using results from the 2006 Census and 2011 NHS linkages. As stated in section B, while this response error is deterministically corrected based on their responses to other questions, a year of immigration still needs to be imputed. When selecting a donor for this missing value, some variables, such as age and place of birth, would be correlated with year of immigration. However, the pool of qualified donors would be restricted to those who answered the question. This could result in a bias towards those who have landed more recently.

Chart 1: Percentage of respondents linked to the administrative immigration data responding "No" to the immigrant status question by linked administrative year of immigration, 2006 Census and 2011 NHS



Source: Statistics Canada, 2006 Census linked to IRCC administrative immigration data and 2011 NHS linked to IRCC administrative immigration data

14. Table 2 demonstrates the coherence between linked administrative values and NHS responses for year of immigration before and after edits and imputation. Among those linked to IRCC data, the consistency edits appear to be correcting the majority of false negatives (immigrants who respond that they are not immigrants). However, while over 70% of non-imputed year of immigration responses were equal to the linked administrative values, the imputed values were often more than 5 years different.

Table 2: 2011 NHS year of immigration consistency with linked administrative values for year of immigration before and after consistency edits and donor imputation

2011 NHS year of immigration		
Consistency with linked administrative value	**Before**	**After**
Same	70.9%	71.5%
1 year diff.	8.8%	9.9%
2-5 Year diff.	8.6%	11.4%
More than 5	2.3%	5.8%
Non-immigrant	5.8%	1.3%
Missing	3.7%	

Source: Statistics Canada, 2011 NHS linked to IRCC administrative immigration data

E. Record linkage with census results for analytical purposes

15. In addition to its use for data quality evaluation, the record linkage between the IRCC administrative immigration data and the 2011 NHS was also used to perform analysis connecting admission characteristics (such as category of admission) with outcomes from the NHS (McLeish, 2016). After processing and weighting, about 79% of immigrants who landed since 1980 according to the NHS

were linked to an administrative record. The resulting database was used to perform custom tabulations without any adjustment for missing links or inconsistencies between administrative values and NHS responses.

III. 2016 Census of Population development

16. Following the data quality assessment performed for the 2011 NHS, it was decided to use the linked administrative data directly in processing the 2016 Census immigration responses (Crowe, 2017). While no values would be deterministically changed as a result of disagreeing with a linked administrative value, (as the record linkage process itself is not without error), the linkage would inform consistency edits and linked administrative variables would be used as matching fields when selecting a donor. The latter use is particularly important, as this would provide auxiliary versions of two important immigration variables that may need imputation: place of birth and year of immigration. These would be more powerful matching constraints than less correlated variables such as mother tongue or place of residence.

F. Development of edit and imputation methods using administrative data

17. In advance of the 2016 Census of Population, a testing environment was created using data from the 2011 NHS including the relevant linked administrative immigration variables. As the edit and imputation processes were revised and developed, they were also tested using this environment. This allowed a better understanding of the effects of introducing administrative variables into the donor selection process or revising any consistency edits.

18. When assessing the consistency edits affecting whether individuals should be classified as immigrants or not, the percentage of records linked provided a proxy for the number of immigrants in a given situation. For example, if a low percentage of respondents in a certain inconsistency are linked, this may indicate that the large majority of those cases are not immigrants. While this was informative and in some cases shed additional details which warranted changes, consistency edits remained an interpretation of Canadian immigration and citizenship policy and were not applied simply because of the results of the record linkage.

19. For the imputation of immigration variables, the linked values provided powerful matching fields, even for respondents who failed to answer multiple questions. In particular, the use of these values improved the effectiveness of treating the non-ignorable missing values for year of immigration. On the other hand, since the administrative data is being used in this way, it can no longer be used as an independent certification data source.

G. Addition of admission category variables

20. Following the 2011 NHS linkage, IRCC funded Statistics Canada to add new variables to the 2016 Census related to the admission category of immigrants (McLeish, 2016). These variables would come directly from the IRCC administrative data linked to the 2016 Census for those immigrants who had landed since 1980 (according to their census record). The variables included were admission category and applicant type. The categories were determined to permit detailed

analysis of specific immigration programs (i.e., Canadian experience class, provincial nominees, and blended visa office-referred refugees) while being able to classify these into broader groupings (i.e., economic immigrants, immigrants sponsored by family, refugees, and other immigrants). These categories were developed in collaboration with partners from IRCC.

1. Development of consistency edits for new variables

21. While the new linked variables do not suffer from the same types of response errors as the variables resulting from the census immigration questions, there are errors in the record linkage process itself, resulting in both false negatives and false positives. In addition, the linked values need to be consistent with the traditional content of the census in order to complete their integration as new census variables. For these reasons, consistency edits needed to be developed and applied.

22. Unlike the traditional immigration variables found on the census, there was no precedent for consistency edits for the new admission category variables. Since the variables are policy-driven, and following the approach used for the other variables, consistency edits were established based on immigration policy in Canada. Under guidance from partners at IRCC, consistency rules ensured that program timelines and requirements were applied to the linked data. For example, a respondent who responded that they landed in 1994 (or had this value imputed) could not be permitted to have a linked admission category value for a program which did not exist in that year. These consistency edits were tested using the linked 2011 NHS results.

2. Development of imputation methods for new variables

23. In part due to false negatives in the record linkage process and in part due to census response and imputation errors, not all immigrants who landed since 1980, according to the census, are linked to an administrative record. In addition, any case for which a consistency edit determined that a linked admission category value was invalid required a new (and valid) value for admission category. These were addressed using the same donor imputation methods employed for the traditional immigration variables. Given that the 2011 NHS linkage would have required a 21% imputation rate, this process was expected to determine the final data quality for these new variables (McLeish, 2016).

24. As was the case for consistency edits, determining the imputation strategy, the matching fields, and their relative importance, was without precedent. Using an approach similar to the one that was used for the traditional immigration content, the imputation was stratified to account for as much familial information as possible (individuals living in couples were imputed using their spouse's information and children imputed based on their siblings' and parents' characteristics). This was driven by the 2011 NHS linkage results, which demonstrated the strong intra-family correlations for admission category (Martin, 2017).

25. Within each stratum, in order to determine which matching fields should be used, and their relative importance to the donor selection process, the ReliefF algorithm was employed (Kononenko, 1994; Kira and Rendell, 1992). ReliefF works under similar assumptions as the nearest-neighbour donor imputation methods

employed for the Canadian Census of Population. The algorithm assumes that records with similar matching fields should come from the same class and records with different matching fields should come from different classes. Matching fields which tend to be similar within classes and different between classes are given more weight. Table 3 demonstrates the results from the 2011 NHS test environment for individuals living in a couple who shared the same year of immigration as their spouse. The results of the algorithm align with subject-matter knowledge in that the variables most important to predict admission category for this stratum are the spouse's admission category and applicant type, as well as the respondent's own place of birth, and year of immigration.

Table 3: ReliefF resulting weights for highest weighted matching fields to impute admission category for individuals living in a couple with the same year of immigration as their spouse, 2011 NHS test environment

Matching fields	Weight
Spouse's Admission Category	0.674
Spouse's Applicant Type	0.329
Place of Birth	0.284
Year of Immigration	0.18
Age at Immigration	0.127
Level of Schooling	0.0999
Job Type	0.0832
Province of Residence	0.0713
Official Languages	0.04
Sex	0.0387

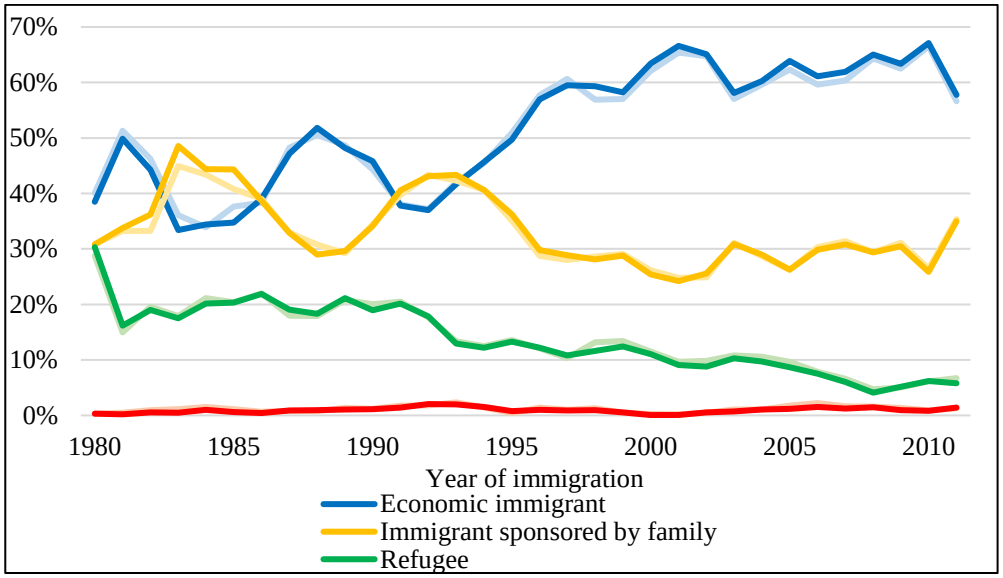
Source: Statistics Canada, Edit and imputation test data, 2011 National Household Survey and IRCC administrative immigration data linkage

H. Testing and certification of developed imputation methods

26. In order to assess the quality of the imputation methods, random subsets of complete records from the test environment were blanked and then sent to imputation. Their imputed values could then be compared with their known responses. While this provides evidence of how imputation can address missing data under different assumptions, it is still based on responses (or non-missing values).

27. For the newly developed donor imputation methods for admission category, this analysis showed that imputation not only maintained the marginal distribution but also maintained important cross-tabular correlations (e.g. for year of immigration). Chart 2 demonstrates the distribution for admission category by year of immigration based on original (lighter tones) and imputed values (darker tones) among adults.

Chart 2: Original and imputed distribution of admission category by single year of immigration, 2011 NHS test environment

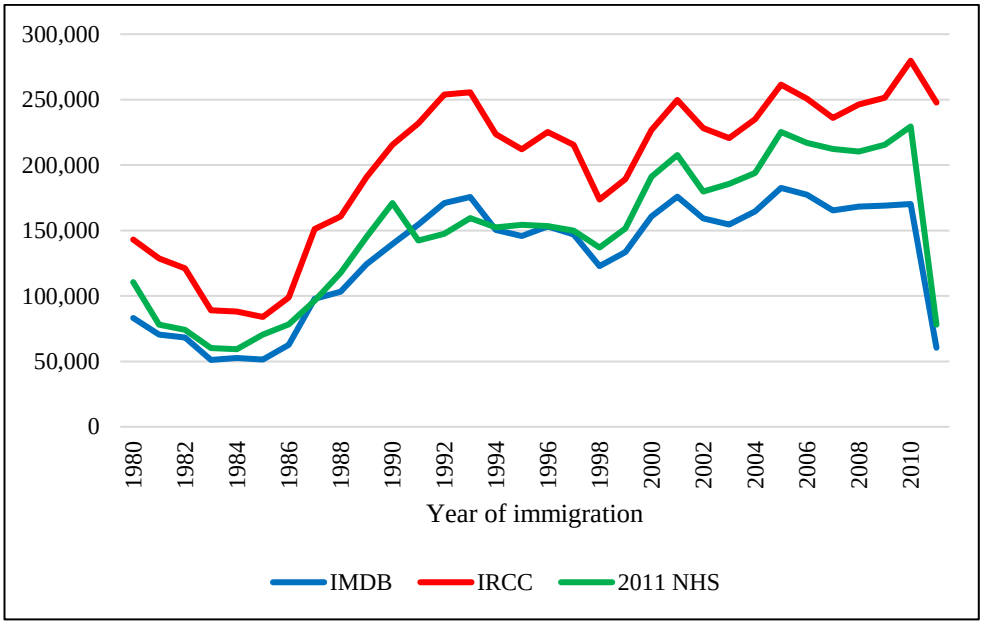


Source: Statistics Canada, Edit and imputation test data, 2011 National Household Survey and IRCC administrative immigration data linkage

28. One disadvantage of employing the linked administrative data in the imputation process is its loss as an independent data source to certify the results. However, IRCC administrative data (without any record linkage) and the IMDB can still be used to provide upper and lower bounds, respectively, to the final census estimates. The IRCC data provides a necessary upper bound, as it is an administrative census of all immigrants landing in Canada, but it does not account for any deaths or emigration. The IMDB can be used to get an estimate of the number of resident tax filers in the most recent tax year. This should be lower than the census estimate, as it would exclude children and others who do not file taxes. These two data sources are useful as they contain similar information such as year of immigration, place of birth, and, now, admission category. While differences resulting from response and imputation errors could affect comparisons for year of immigration and place of birth, the admission category data comes from the same source for all three data sources, with differences possibly caused by linkage error, coverage differences, and imputation for the census.

29. Chart 3 demonstrates how the IRCC data and the IMDB can be used to certify the final census results. This plot shows a comparison of the distribution of immigrants by single year of immigration between these data sources. As expected, the IRCC data provide the upper bound and the 2010 tax filers for the IMDB provide the lower bound. In general, the trends associated with the ebbs and flows of immigration to Canada are reflected in a consistent manner between the three data sources. This analysis can be replicated to focus on specific cohorts, places of birth, age groups, sex, and admission category.

Chart 3: Estimates by single year of immigration for immigrants who have landed in Canada since 1980, IMDB, IRCC administrative records, 2011 National Household Survey



Source: Statistics Canada, 2014 IMDB – immigrants who filed taxes in 2010, 2011 National Household Survey; Immigration Refugees and Citizenship Canada, administrative immigration data

IV. Conclusion and next steps

I. Use of administrative data has improved edit and imputation methods

30. At the time of writing this paper, the final results of the 2016 Census have not yet been released (the immigration variables will be disseminated on October 25th, 2017). However, in testing and development, the integration of administrative data was found to have enhanced the quality of consistency edits and donor imputation for traditional immigration variables. Consistency edits have been more directly confronted with the use of the linkage as a proxy for immigrant status and donor imputation has improved from the use of more directly correlated auxiliary variables, such as linked administrative year of immigration and linked administrative place of birth.

31. The record linkage undertaken for the 2016 Census was expanded to include immigration records back to 1952, as well as records for asylum claimants and temporary residents. This will enhance the processing beyond what was tested using the 2011 NHS results (Biernot, 2017).

32. For future censuses, other content from linked administrative data could be used in similar approaches to ensure consistency edits are sound and that the donor selection process of imputation takes advantage of the most relevant information available.

J. Edit and imputation methods for admission category variables were developed

33. The collaboration between Statistics Canada and IRCC has led to more thorough consistency edits for the new admission category variables. These were tested using the 2011 NHS linkage results and then re-tested using the 2016 Census linkage results.

34. While a primary concern leading up to the 2016 Census was the reliance on donor imputation for the admission category variables, testing showed that the imputation methods developed led to reliable results that were in line with the IRCC administrative data and the IMDB. The use of ReliefF to determine the matching fields and their relative importance in the donor selection process has provided a statistical foundation for this process instead of relying solely on subject-matter knowledge. In addition, the 2016 Census linkage rates were much higher than what was observed for the 2011 NHS – this will lead to lower imputation rates than initially expected for these variables (Biernot, 2017).

35. Overall, the introduction of these two new variables heralds new analytical possibilities for the socio-economic outcomes of immigrants to Canada, connecting the wealth of information traditionally collected in the Census of Population with admission program information from IRCC. The methods used to produce the final estimates were developed and tested to ensure this analysis will benefit from good data quality.

V. Acknowledgments

36. This project would not be possible without support and collaboration from many individuals representing different areas of expertise including subject-matter analysts, methodologists and database managers. In particular, Statistics Canada would like to acknowledge the ongoing contributions of Lorna Jantzen and her colleagues at IRCC and their partnership for the inclusion of admission category variables on the Census of Population, as well as their ongoing support in interpreting their administrative data. Within Statistics Canada, the authors would like to specifically recognize the ongoing efforts of Eric Mongrain, Laetitia Martin, Hélène Maheux, Mireille Vézina, Chantal Poirier, Hyunji Lee, and Jarod Dobson from Social and Aboriginal Statistics Division; Andrew Stelmack, Sean Crowe, and Lyne Guertin from Social Survey Methods Division; Piotr Biernot, Caroline Pelletier, Paul Cascagnette, and Colin Babyak from Household Survey Methods Division, and many others who have contributed to the development and production of immigration content for the 2016 Census of Population.

VI. References

Biernot, P. 2017. “External linkage between the Immigration File (1952-2016) and the 2016 Census Response Database.” Unpublished report. Ottawa: Statistics Canada.

Brennan, J. 2011. “CIC Landing File to Census 2006 Linkage.” Unpublished report. Ottawa: Statistics Canada.

Brennan, J. 2013. "CIC Landing File to Census 2011/NHS Linkage." Unpublished report. Ottawa: Statistics Canada.

Crowe, S. and Janes, D. 2017. "Edit and Imputation Report: Ethnocultural Process." Unpublished report. Ottawa: Statistics Canada.

Dobson, J., Costa, R., and Martin, L. 2014. "Narrative of Editing and Imputation Procedures: EC Variables." Unpublished report. Ottawa: Statistics Canada.

Evra, R. and Prokopenko, E. 2017. "Longitudinal Immigration Database (IMDB) Technical Report, 2014." *Analytical Studies: Methods and References.* Catalogue no. 11-633-x2017007. Ottawa: Statistics Canada.

Guertin, L., Bureau, M., and Morel, J. 2014. "Editing the 2011 Census data with CANCEIS and options considered for 2016." *2014 Work Session on Statistical Data Editing*. Conference of European Statisticians. United Nations Economic Commission for Europe.

Kira, K., Rendell, L. 1992. "The Feature Selection Problem: Traditional Methods and a New Algorithm." *AAAI-92 Proceedings*, 129-134.

Kononenko, I. 1994. "Estimation Attributes: Analysis and Extensions of RELIEF." *Proceedings of the 1994 European Conference on Machine Learning*. Catania, 6-8 April 1994, 171-182.

Martin, L. 2016. "Compte rendu des procédures de contrôle et imputation pour le processus catégorie d'admission (AC)." Unpublished report. Ottawa: Statistics Canada.

Martin, L. 2017. "From the source to dissemination: processing, accessibility and possible uses of the Census admission category of immigrants" Presentation delivered to the 19th Canadian National Metropolis Conference. Ottawa: Statistics Canada.

McLeish, S. 2012. "Feasibility Study: IMDB for Statistical Purposes." Unpublished report. Ottawa: Statistics Canada.

McLeish, S. 2014. "Using administrative data to evaluate data quality: Immigrants in the 2011 National Household Survey", *Proceedings of Statistics Canada Symposium 2013,* Catalogue no. 11-522-X. Ottawa: Statistics Canada.

McLeish, S. 2016 "Adding Immigrant Admission Category to the Canadian Census of Population" *2016 Work Session on Migration Statistics*. Conference of European Statisticians. United Nations Economic Commission for Europe.

Statistics Canada. 2013. "Place of Birth, Generation Status, Citizenship and Immigration Reference Guide." 2011 National Household Survey. Catalogue no. 99-010-XIE2011008. Ottawa: Statistics Canada.

Statistics Canada. 2014. "CANCEIS, version 5.2. Basic User Guide." Ottawa: Statistics Canada.

Statistics Canada. 2016. "Dictionary, Census of Population, 2016" 2016 Census of Population. Ottawa: Statistics Canada.

Statistics Canada. 2017. "Guide to the Census of Population, 2016" 2016 Census of Population. Catalogue no. 98-304-X. Ottawa: Statistics Canada.