Catalogue no. 11-522-X
ISSN 1709-8211

Proceedings of Statistics Canada Symposium 2022:
Data Disaggregation: building a more representative data portrait of society

Integration of existing data to develop an ethnicity indicator in the LSDDP

by Aziz Farah, Bassirou Diagne and Abdelnasser Saïdi

Release date: March 25, 2024

Statistics Canada Statistique Canada

Canada

Proceedings of Statistics Canada's Symposium 2022
Data disaggregation: Building a more representative data portrait of society

Integration of existing data to develop an ethnicity indicator in the LSDDP

Aziz Farah, Bassirou Diagne and Abdelnasser Saïdi[1]

Abstract

The Longitudinal Social Data Development Program (LSDDP) is a social data integration approach aimed at providing longitudinal analytical opportunities without imposing additional burden on respondents. The LSDDP uses a multitude of signals from different data sources for the same individual, which helps to better understand their interactions and track changes over time. This article looks at how the ethnicity status of people in Canada can be estimated at the most detailed disaggregated level possible using the results from a variety of business rules applied to linked data and to the LSDDP denominator. It will then show how improvements were obtained using machine learning methods, such as decision trees and random forest techniques.

Keywords: Data integration; machine learning; ethnicity indicator.

1. Introduction

In the context of the Disaggregated Data Action Plan (DDAP), the Longitudinal Social Data Development Program (LSDDP) proposes to develop an algorithm with the objective of developing an ethnicity indicator for each individual in the Canadian population. The algorithm is based on the 2016 reference year and can be generalized to any subsequent year. Assignment will be at the individual level (the most disaggregated level possible), for all individuals in the Canadian population. This indicator will provide additional analytical opportunities when used as a covariate. It can also be used to develop more inclusive sampling frames when targeting specific ethnic groups for sample allocation or stratification.

The LSDDP integrates existing, linked administrative data based on a single, anonymous key developed in the Social Data Linkage Environment (SDLE). The SDLE is a secure environment that consistently adheres to the strictest standards of privacy and data security. Furthermore, the LSDDP is a set of algorithms and processes for analyzing social data trajectories. It is not a large integrated database, nor an evolving environment, and does not contain any personal identifiers.

2. Context

2.1 Ethnic categories

The concept of ethnicity measured here is similar to that of "ethnic identity," as defined in the question on population group in the 2016 Canadian Census of Population (question 19) (please refer to Statistics Canada (2016)). In addition to the 12 visible minority groups, two other categories not classified as visible minorities —the "White" group and the "Indigenous" group— are added to create the following 14 ethnic categories:

[1]Aziz Farah, Statistics Canada, Canada, aziz.farah@statcan.gc.ca; Bassirou Daigne, Statistics Canada, Canada, bassirou.diagne@statcan.gc.ca; Abdelnasser Saïdi, Statistics Canada, Canada, abdelnasser.saidi@statcan.gc.ca.

Table 2.1-1
Ethnic groups

Code	English name	
1	South Asian	Visible minority
2	Chinese	
3	Black	
4	Filipino	
5	Latin American	
6	Arab	
7	Southeast Asian	
8	West Asian	
9	Korean	
10	Japanese	
11	Visible minorities, n.i.e. (not included elsewhere)	
12	Multiple visible minorities	
13	Other, not a visible minority [White]	Not a visible minority
14	Indigenous	

2.2 LSDDP denominator (Canadian population)

This is a portrait of the Canadian population of interest for a given reference date. It is produced on demand using an algorithm that combines multiple linked data sources based on the SDLE. This denominator is central to the LSDDP's data disaggregation and development programs (please refer to Aubin, P. (2021)). The ultimate goal of this work is to assign one of the 14 ethnic categories in Table 2.1-1 to each individual in the LSDDP denominator, while maintaining high levels of quality.

2.3 Data sources and direct and indirect signals of ethnicity

Direct signals (by self-response) or indirect signals (by inference) from the following data sources were used to develop the indicator:

- Data from the last five censuses (the 2016, 2011, 2006, 2001 and 1996 censuses) that contain responses to the question on population group in the long-form questionnaire (direct signal)
- The parent-centric file (for the years 1993 to 2017) which describes the link between parents and their children (indirect signal). That file is built in the SDLE environment (please refer to Cascagnette, P. (2020, version 35), Cascagnette, P. (2020, version 30), Labrecque-Synnott, F. (2020), Gissler, G. (2020)) .
- The historical birth and immigration files, which represent the main source of the LSDDP denominator (indirect signal)
- Registered Apprenticeship Information System (RAIS, 2008 and later) (direct signal)
- Postsecondary Student Information System (PSIS, 2009 and later) (direct signal)
- Ontario Mental Health Reporting System (OMHRS, 2005 and later) (direct signal)

T1 Family File (T1FF, 2005 and later) (indirect signal)

3. Methodology

3.1 Direct and indirect methods

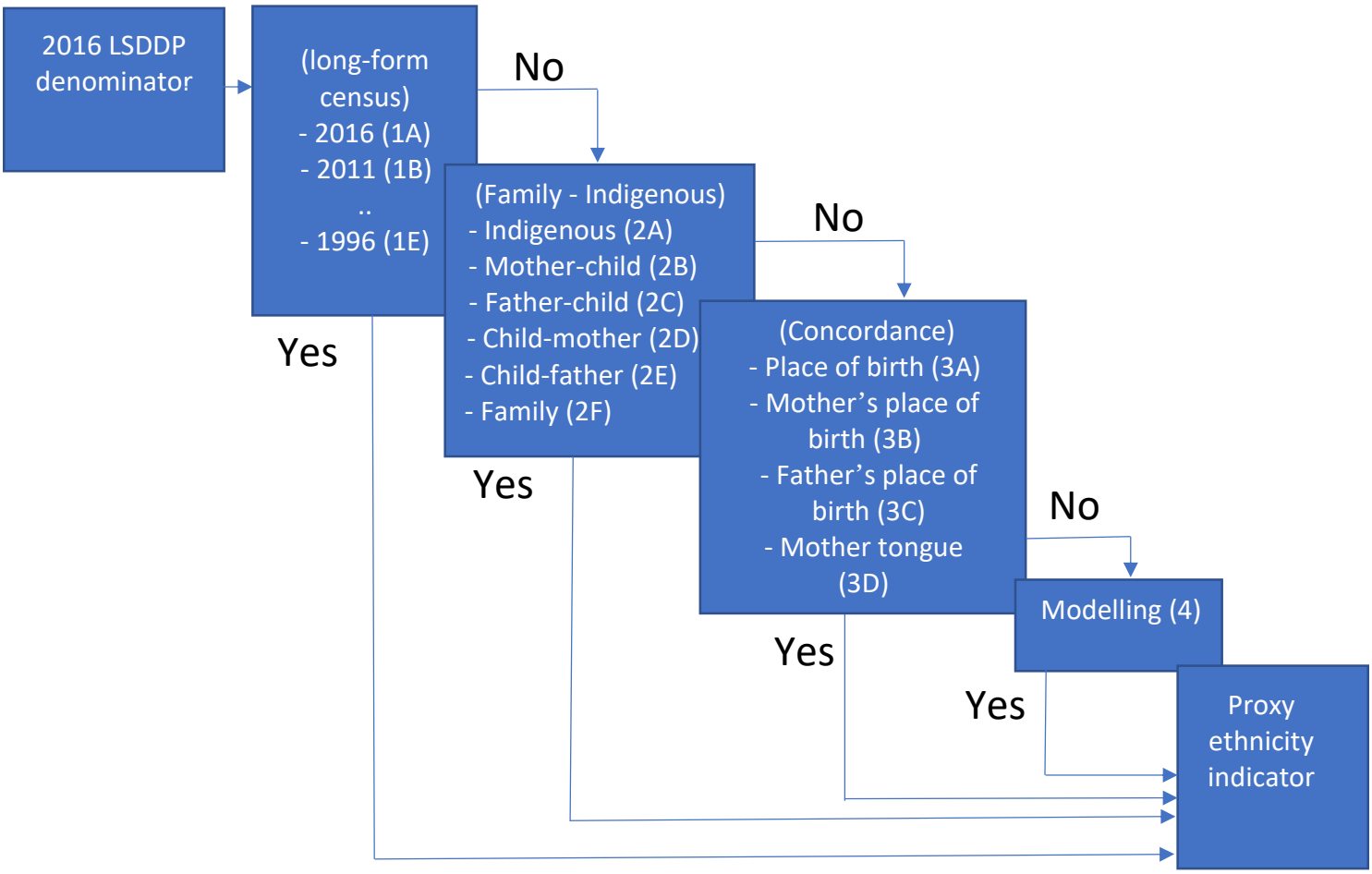
All the data sources used are already linked in the SDLE environment. As a result, more than one ethnicity-related signal (direct or indirect) can be associated with the same individual. Such an association can be longitudinal (based

on multiple years) and/or based on the different data sources. In addition, this association does not use any personal identifiers, but instead a common, anonymous key from the SDLE, which makes the procedure completely confidential.

The same individual can appear in the responses of more than one census at the same time, with identical or sometimes conflicting ethnic statuses. The same individual can also appear in an administrative file (e.g., historical immigration file), sending an indirect signal that infers their ethnic status. For example, place of birth or mother tongue can be considered indirect signals when they strongly refer to a particular ethnicity.

Chart 3.1-1 below shows the hierarchy of steps used to develop the indicator (step 1, step 2, ...step 4). Within each step, the results of the classification of sub step A are selected first, then sub step B, then C, etc.

Chart 3.1-1
Hierarchy of the ethnicity indicator classification steps



The methodology used assumes that the census is the most reliable source of information, and therefore self-response from the long form will be prioritized among all other direct or indirect signals.

Table 3.1-1 below, from a study of fluidity between responses in the 2011 and 2016 censuses, illustrates the great stability of responses for the main ethnic groups from one census cycle to another (please refer to Farah, A. (2022)):

Table 3.1-1
Stability rate of responses between the 2011 and 2016 censuses

Group	Stability rate	Group	Stability rate
South Asian	94.1%	Southeast Asian	73.5%
Chinese	95.8%	West Asian	78.8%
Black	91.3%	Korean	98.0%
Filipino	94.7%	Japanese	91.3%
Latin American	84.9%	White	97.7%
Arab	80.5%	Indigenous	95.9%

The first step in the development process involves using as much direct information as possible from the available censuses. In this case, when the same individual responds to multiple censuses at the same time, the response in the most recent census is selected.

Cases not classified by any of the last five available censuses move to step 2. In this step, assuming that ethnicity is relatively static within a given family, the ethnic status of one family member, when available, can be imputed to another family member who is not yet classified (parent, child, brother, sister). The parent-centric file helps to describe the link between parents and children, and therefore identify membership in the same family. The imputation approach used can be criticized and, for that reason, family imputation was placed in the second step of the hierarchy.

Then, unclassified cases move to step 3, where it is suggested that the link between place of birth, mother tongue and ethnic status be used for a few groups in particular. In this case, very strict business rules were developed to select only a few places of birth (country of birth of the individual or their parents) and a few mother tongues whose rate of correct classification with a given ethnicity exceeds 80%. More concretely, the ethnic status responses from the 2016 Census were cross-tabulated with the classification from step 3 to select subsets of countries or mother tongues for which more than 80% of people answered that they belong to the ethnicity in the 2016 Census.

3.2 Modelling and machine learning methods

After applying direct and indirect methods, roughly 30% of the LSDDP denominator remains unclassified. Compared with the distribution of ethnic status in the 2016 long-form census, it is inferred that most of these cases belong to the "White," "Black," or "Indigenous" groups.

Using machine learning modelling proves useful after step 3. The models used were developed with up to 70% training data in order to optimize the settings, while predictive quality was measured using test data.

Since most of the unclassified cases fall into the "White" category, the first classification model was devoted to exclusively classifying this category using the decision tree method (please refer to Breiman, L., J. Friedman, R. A. Olshen, and C. J. Stone (1984)) with the HPSPLIT option in SAS EG (please refer to SAS Institute Inc. (2020)). The second model was applied to the characterization of the other categories ("Black," "Filipino," "Southeast Asian" and "Japanese") using the XGBoost algorithm, i.e., extreme gradient boosting (please refer to Chen, T., and C. Guestrin (2016)). This learning model processes large, complex data quickly and effectively through a combination of decision trees.

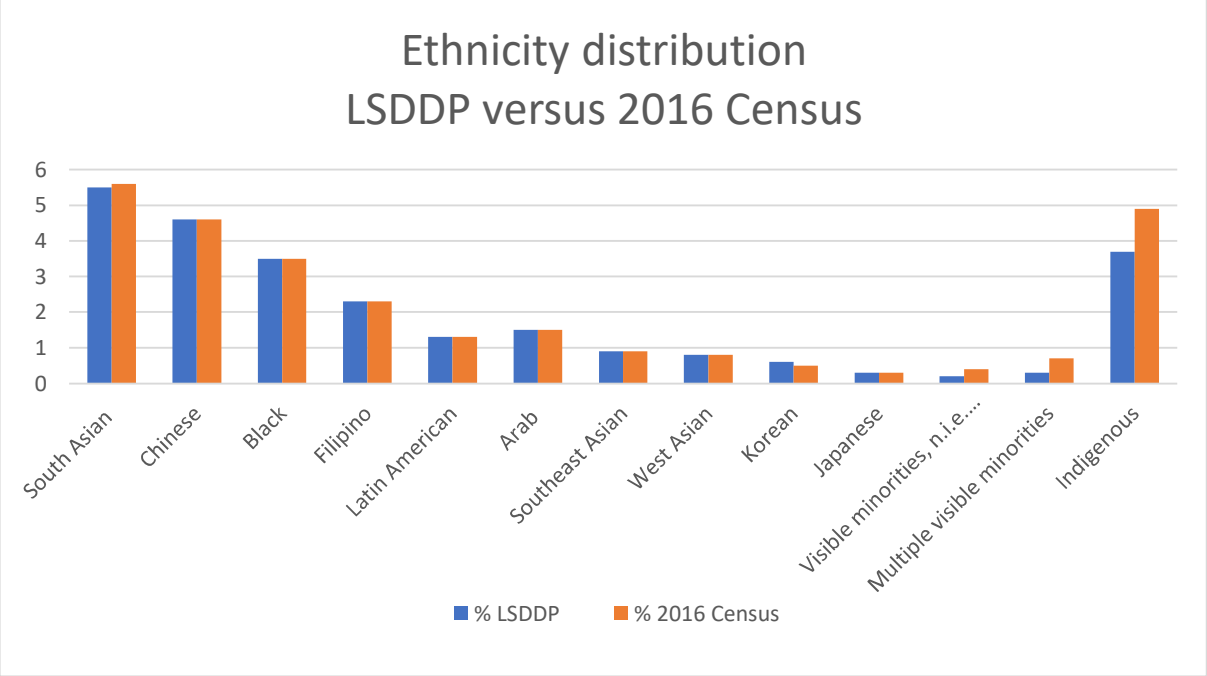
4. Empirical results

4.1 Comparison with the 2016 Census

After all these steps, a comparison of the distribution of ethnicity counts between the LSDDP and the 2016 Census shows a great similarity for most categories, except the "Indigenous" group, to which the modelling methods were not applied, along with two difficult groups: "Multiple visible minorities" and "Visible minorities not included

elsewhere." Chart 4.1-1 below shows the results of this comparison. Note that the "White" group, the largest category among all other categories, is not presented in this figure for visibility purposes.

Chart 4.1-1
Comparison of the ethnicity distribution between the LSDDP estimates and the 2016 Census estimates

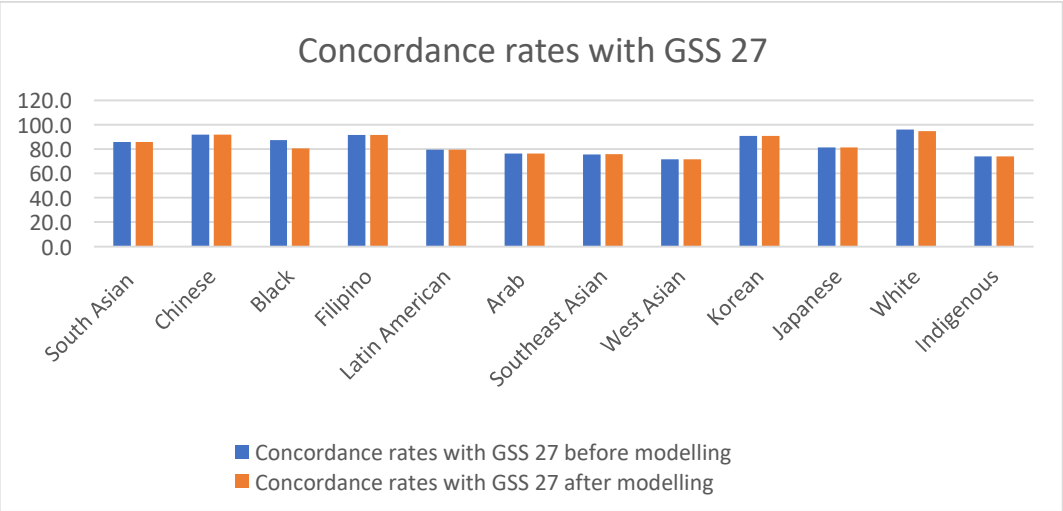


4.2 Micro comparison with the GSS 27

Cycle 27 of the General Social Survey (GSS) also looked at ethnicity. Its data were not used to develop the ethnicity indicator, but to validate the results of our classification.

Chart 4.2-1 below presents the concordance rates (correct classification) between the LSDDP estimates and the GSS 27 estimates pre- and post-modelling.

Chart 4.2-1
Validation of the estimates with GSS 27



5. Conclusion and looking ahead

Using existing linked data shows that it is possible to build static indicators such as ethnicity using direct and indirect methods and machine learning modelling. By simply using existing census data, about 50% of the Canadian population (LSDDP denominator) was classified by ethnicity. Indirect methods were able to add about 20% more, and modelling helped to reach almost the entire target population, while maintaining good quality levels in our estimates.

However, challenges remain for the most difficult groups to classify, such as the "Indigenous" group and the two groups "Multiple visible minorities" and "Visible minorities not included elsewhere." In addition, it would be relevant to more closely examine the source of overestimation and underestimation of certain groups and the potential for applying calibration methods. Validating with census data and with an external survey (GSS) certainly helps to justify the quality level of the estimates in general, but consideration must also be given to developing a quality indicator at the individual level.

6. References

Aubin, P. (2021), "Methodology of the Longitudinal Social Data Development Program (LSDDP) Phase III", Internal report, Statistics Canada.

Breiman, L., J. Friedman, R. A. Olshen, and C. J. Stone (1984), *Classification and Regression Trees*, Belmont, CA: Wadsworth.

Cascagnette, P. (2020), "Linkage between Birth Mothers (1993 to 2017) and the SDLE Derived Record Depository", version 30, Internal report, Statistics Canada.

Cascagnette, P. (2020), "Linkage between Birth Fathers (1993 to 2017) and the SDLE Derived Record Depository", version 35, Internal report, Statistics Canada.

Chen, T., and C. Guestrin (2016), "XGBoost: A Scalable Tree Boosting System", *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794. New York, NY, USA: ACM. https://doi.org/10.1145/2939672.2939785

Farah, A. (2022), « Développement d'un indicateur sur les minorités visibles », Internal document, Statistics Canada.

Gissler, G. (2020), "Linkage between Stillbirths and_Mothers (1993 to 2017) and the SDLE Derived Record Depository", version 30, Internal report, Statistics Canada.

Labrecque-Synnott, F. (2020), "Linkage between Stillbirths and Fathers (1993 to 2017) and the SDLE Derived Record Depository", version 35, Internal report, Statistics Canada.

SAS Institute Inc. (2020), SAS/STAT® 15.2, *User's Guide*, Cary, NC: SAS Institute Inc.

Statistics Canada (2016), "Census of Population 2016", Visible Minority and Population Group Reference Guide, Census of Population, 2016 (statcan.gc.ca).