European Population Conference 2016, Mainz, Germany
August 31st to September 3rd, 2016

Title: Use of administrative data for the Census of Population: overview of a research project and data quality assessment tools developed by Statistics Canada

André Lebel, demographer
(Presenter and contact author)

Demography Division
Statistics Canada
150 Tunney's pasture driveway
Ottawa, Canada
K1A 0T6
Tel : 613-286-3781
Email : andre.lebel@canada.ca

Short Abstract:

Canada has held censuses every ten-years since 1871, and every five-years since 1956. In industrialized countries, the traditional form of census is being questioned, and some countries, mostly from Northern Europe, have just recently moved away from this model by taking advantage of their long existing population register. The term traditional census refers to the collection of information on individuals through the use of a variety of collection methods (full field enumeration, self-completion paper questionnaire, telephone or internet). Even for countries still holding a traditional census, many have recently been reviewing the potential of using administrative and other alternative data sources to replace the direct enumeration of the population (e.g. United Kingdom, Australia and New Zealand).

Statistics Canada has also put in place a research project to assess the extent to which it is possible to complement or supplement the Canadian Census with a "*virtual population register"* using a variety of administrative databases (tax and economic activity, foreign entries, vital statistics, etc.) already available to the statistical agency. A key element of this project is referred to as the Canadian Statistical Demographic Database (CSDD) which contains basic socio-demographic variables (age, sex, geographic location) which replicate parts of the census universe. The first objective of this paper is to present the CSDD and briefly explain data sources and methods used to build it. In the second part, data quality indicators developed to assess its fitness and comparability with the current Census will be presented. The third part will contain an assessment of the quality of the coverage achieved by the CSDD for different levels of geography (Provinces, Census Metropolitan Areas (CMA) and Census Divisions (CD) and also, by age groups and sex. Finally, this paper will conclude by outlining CSDD's strengths and weaknesses and propose ways to improve it in 2016.

Long Abstract:

Canada has held censuses every ten-years since 1871, and, every five-years since 1956. In industrialized countries, the traditional form of census is being questioned, and some countries, mostly from Northern Europe[1,2], have just recently moved away from this model by taking advantage of their long existing population register. The term traditional census refers to the collection of information on individuals through the use of a variety of collection methods (full field enumeration, self-completion paper questionnaire, telephone or internet). Even for countries still holding a traditional census[3], many have recently been reviewing the potential of using administrative and other alternative data sources to replace the direct enumeration of the population (e.g. United Kingdom[4], Australia and New Zealand[5]).

Statistics Canada has also put in place a research project to assess the extent to which it is possible to complement or supplement the Canadian Census with a "*virtual population register"* using a variety of administrative databases[6] (tax and economic activity, foreign entries, vital statistics, etc.) already available to the statistical agency. A key element of this project is referred to as the Canadian Statistical Demographic Database (CSDD) which contains basic socio-demographic variables (age, sex, geographic location) which replicate parts of the Canadian census universe. The first objective of this paper is to present the CSDD and briefly explain data sources and methods used to build it. In the second part, data quality indicators developed to assess its fitness and comparability with the current Census will be presented. The third part will contain an assessment of the quality of the coverage achieved by the CSDD with the Demographic Censal Estimates for different levels of geography (Provinces, Census Metropolitan Areas (CMAs) and Census Divisions (CDs) and also, by age groups and sex. Finally, this paper will conclude by outlining the strengths and weaknesses of the CSDD and proposes ways to improve it for the next iteration in 2016.

The CSDD was created by linking multiple data sources using record linkage tools developed within Statistics Canada. The starting point for the CSDD is a file created by the Canada Revenue Agency (CRA), the Ident file, with a vintage of December 31, 2011. This file contains all historical tax filers since 1984, along with their latest contact information. Information for children on the CSDD comes from the CRA Canada Child Tax Benefit (CCTB) file that also includes the Universal Child Care Benefit (UCCB) recipients, and the Canadian Birth Database (CBDB). Using files from Citizenship and Immigration Canada (CIC), the CSDD included immigrants whose landing date was between 1980 and the 2011 Census day, as well as non-permanent residents that had a valid temporary resident, work or student permit on the 2011 Census day or were refugee claimants. Throughout the process, deaths were removed from the CSDD using an auxiliary file called the Improved Canadian Mortality Database which combined the Canadian Mortality Database (CMDB) developed from

[1] UNECE (2007). *Register-based statistics in the Nordic countries – review of best practices with focus on population and social statistics*.
[2] Statistics Norway (2008). *Topics difficult to measure in a register-based census (Norway).* Joint UNECE/Eurostat meeting on population and housing censuses, Geneva. www.unece.org/stats/documents/2008.05.census.htm
[3] UNECE (2014). *Measuring population and housing, Practices of UNECE countries in the 2010 round of censuses*, p.230
[4] Office for National Statistics (2013). *Beyond 2011: Producing Population Estimates Using Administrative Data*, p. 25
[5] Statistics New Zealand (2011). *A register-based census: what is the potential for New Zealand?*, p. 17
[6] Statistics Canada (2015). *Corporate Business Plan Statistics Canada 2015-16 to 2017-18,* http://www.statcan.gc.ca/eng/about/bp

provincial and territorial vital statistics and deaths declared in CRA taxation files.

Throughout the process, individuals who appear "inactive" were removed from the CSDD using another auxiliary file, which is the Fiscal Activity Indicator File. This file includes flags indicating the presence or absence of a given individual on fourteen different CRA files from 2000 to 2011 (e.g., employment revenue reports (T4 slips), Canada Pension Plan, social assistance and other provincial supplements). The last step in the creation of the CSDD was to assign individuals a most likely Census-day address and to use the Statistics Canada's Address Register (AR) to code it to an AR_UID, which is the AR's unique identifier that corresponds to a dwelling. Addresses that could not be coded because they were not precise enough, incomplete or incorrect were assigned to the most precise geographic location next in the hierarchy (i.e., the block face).

Population counts from the CSDD on Census day (May 10th, 2011) are assessed and evaluated against the estimates of the Population Estimates Program (PEP), referred to as the Demographic Censal Estimates. Its fitness to the latter is then compared with the official population counts from the 2011 Canadian Census. Although these final or published census counts are never revised, coverage studies are carried out after the census on samples of individuals taken from the previous census and current administrative data to estimate how many individuals were missed (undercoverage) or counted more than once (overcoverage). From this, census net undercoverage estimates (CNUC), i.e. overcoverage minus undercoverage, are produced by sex, age group and broad geographic areas (province, territory and a few large CMAs). Statistics Canada produces official population estimates using this CNUC adjusted census, resulting in Censal population estimates on Census day.

The quality of the CSDD population counts is being assessed in this paper by looking at a variety of data quality indicators (DQI). DQI range from looking at the level of missing information or differences distribution and to overall measures of average differences for key demographic and geographic variables. Mainly, three measures of average difference with Demographic Censal Estimates are used in this paper: the Mean Absolute Percentage Error (MAPE), the Median Absolute Percentage Error (MedAPE) and the Weighted Mean Absolute Percentage Error (WMAPE). They are based on the Percentage Error (PE) of alternative population counts (CSDD, Census Published) with Demographic Censal Estimates, defined for any geographic area. The use of MAPE in population accuracy studies[7, 8] (forecast or estimate) have been criticized since the distribution of absolute error values is frequently right-skewed, thus 'overstating' error. The MedAPE, the middle value of the ranked set of Absolute Percentage Errors, can be used to complement MAPE since it is less affected by outlier values. The weakness of the MAPE is that it gives for each observation, equal weighting in the calculation of average error. The third measure, the Weighted Mean Absolute Percentage Error (WMAPE), takes into consideration the population size to relatively weigh the error. By doing so, a 5% error in a population of 1,000,000 has more impact on the WMAPE than a 25% error for a population of 1,000. A good fit with the benchmark is assumed when the indicator is close to 0%.

[7] Tom Wilson (2011). *The forecast accuracy of local government area population projections: a case study of Queensland*, Australasian Journal of Regional Studies, Vol. 17 No. 2, 2011

[8] Tayman, J., Swanson, D. A. and Barr, C. F. (1999). *In search of the ideal measure of accuracy for subnational demographic forecasts*. Population Research and Policy Review, 18, pp. 387–409.

The results of this analysis have not yet been published by Statistics Canada, and thus for the moment, cannot be divulgated in a lot of detail. Overall, results show that the CSDD does very well at the national, provincial and territorial level, but displays weaknesses at lower levels of geographies, as population size decreases and in rural area. By age, the CSDD performs also better than the Census, although there is still room for improvements for middle-aged and among the oldest-old.

Conclusion:

Much work remains to be done before the CSDD is ready for use for purposes other than research. For the next iteration of the CSDD in 2016, a few areas of improvement will be explored based on the analysis provided in this paper.

In order to reduce overcoverage of young adults, the CSDD should continue to enhance methods to estimate emigration using an improved version of the Fiscal Activity Indicator database. Adding new data sources could also be beneficial in some ways but this is not necessarily the only solution to resolve overcoverage for young adults. Effort should be made also to investigate and reduce overcoverage above age 90. To improve fitness at lower geography levels and non-CMA areas (rural), Statistics Canada should continue to work on the quality of the addresses at the subprovincial level in order to reduce the exclusion of records. The CSDD project is deemed successful so far and will continue because it has changed, to a certain extent, the perception on how administrative data in Canada could be used in an environment as important as the Census of Population and could even help streamline some Census operations (e.g. non-response follow-up, processing, quality assessment, edit and imputation).