United Nations ECE/CES/ GE.41/2016/7

Economic and Social Council

Distr.: General
14 July 2016

Original: English

Economic Commission for Europe

Conference of European Statisticians

Group of Experts on Population and Housing Censuses

Eighteenth Meeting
Geneva, 28 - 30 September 2016
Item 5 of the provisional agenda
Methods for assessing quality and usability of registers and administrative sources

Assessing the usability of a statistical population register for the Census of Population in Canada

Note by the Statistics Canada[1]

Summary

Canada has conducted a traditional census since 1871. As in many countries, there is a move to investigate alternative enumeration strategies, such as the use of long standing population registers in Northern Europe. Statistics Canada has put in place a research project to assess the possibility to complement or supplement the Canadian Census with a statistical population register. This register would be built using a variety of data sources available to the statistical agency. It could eventually reduce programme costs, respondent burden, and improve the timeliness of population data. A key element of this project, the Canadian Statistical Demographic Database (CSDD), aims to conceptually replicate the census population universe with basic socio-demographic variables (age, sex and geography).

This document provides a short description of the CSDD, including the data sources and methods used to build it and presents the data quality indicators developed to assess its fitness and comparability with the 2011 Census and official Censal Population Estimates. The last section consists of an evaluation of the CSDD using these indicators and visualization tools. The document concludes by outlining the CSDD's strengths and weaknesses with a proposal for the improvement of the next iterations of the register to be produced at the same time as the 2016 Canadian Census and beyond .

[1] Prepared by André Lebel and Johanne Denis. The views expressed in this document are those of the authors and do not necessarily reflect those of StatisticDOs Canada.

I. Introduction

1. Canada has held traditional censuses every ten years since 1871, and every five years since 1956. As in many countries, there is a move to investigate alternative enumeration strategies, such as the use of long standing population registers in Northern Europe[2,3,4]. 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). More and more countries still holding a traditional census[5] have recently started reviewing the potential of using administrative and other alternative data sources to replace the direct enumeration of the population (e.g., United Kingdom[6]and New Zealand[7]).

2. Statistics Canada has put in place a research project to assess the possibility to complement or supplement the Canadian Census with a statistical population register. This register would be built using a variety of administrative data sources[8] (e.g., tax, foreign entries and vital statistics) available to the statistical agency as part of the "Beyond 2016 Census Program" project. It could eventually reduce program costs, respondent burden, and improve the timeliness of population data. The first attempt towards a statistical population register, the Canadian Statistical Demographic Database (CSDD), aims to conceptually replicate the census population universe with basic socio-demographic variables (age, sex and geography).

3. This document provides a short description of the CSDD, including the data sources and methods used to build it. It will also present the data quality indicators developed to assess its fitness and comparability with the 2011 Census and official Censal Population Estimates. The last section will consist of an evaluation of the CSDD using these quality indicators with graphical representations from data visualization tools. The document will conclude by outlining the CSDD's strengths and weaknesses with a proposal for the improvement of the next iterations of the statistical population register, to be produced at the same time as the 2016 Canadian Census and beyond.

II. Construction of a statistical population register in Canada

4. A vintage of a CSDD with a reference year of 2011 was created in 2013 by linking multiple data sources using record linkage tools developed within Statistics Canada. The starting point for the CSDD was 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 on children in the

[2] UNECE (2007). *Register-based statistics in the Nordic countries – review of best practices with focus on population and social statistics.*

[3] 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

[4] Prins, Kees (2016) *Population register data, basis for the Netherlands' Population Statistics*, Statistics Netherlands.

[5] UNECE (2014). *Measuring population and housing, Practices of UNECE countries in the 2010 round of censuses*, p.230

[6] Office for National Statistics (2013). *Beyond 2011: Producing Population Estimates Using Administrative Data*, p. 25

[7] Statistics New Zealand (2011). *A register-based census: what is the potential for New Zealand?*, p. 17

[8] Statistics Canada (2016). *Corporate Business Plan, Statistics Canada 2016-17 to 2018-19*, http://www.statcan.gc.ca/eng/about/bp

CSDD came from the CRA Canada Child Tax Benefit (CCTB) file that includes the Universal Child Care Benefit (UCCB) recipients, and the Canadian Birth Database (CBDB). Using files from Immigration, Refugees and Citizenship Canada (IRCC), the CSDD included immigrants whose landing date was between 1980 and the 2011 Census day and non-permanent residents who had a valid temporary resident, work or student permit on the 2011 Census day, or refugee claimants. 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 provincial and territorial vital statistics, with deaths declared in CRA taxation files.

5. Individuals who appeared "inactive" in Canada were removed from the CSDD using a derived auxiliary file called 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).

6. The last step in the creation of the CSDD was to assign individuals their most likely address on Census day, and to use the Statistics Canada's Address Register (AR) to code it to the dwelling level. Addresses that could not be coded because they were not precise enough, were incomplete or incorrect were assigned to the next most precise geographical location in the hierarchy (i.e., the block face).

III. Data sources used in the evaluation of the CSDD

7. Population counts derived from the CSDD on the 2011 Census day (10 May) were assessed and evaluated against the published Census counts and the official Censal Population Estimates (PEs) Statistics Canada produces through the Population Estimates Program[9]. PEs are a key input for the calculation of revenue transfers and contributions that are allocated to provinces and territories[10] from the federal government under the Canadian *Fiscal Arrangement Act*. Every year, more than $60 billion Canadian dollars, mainly on a per capita basis, are transferred through that process. For this reason, PEs are considered as benchmark when comparing results at aggregated levels. However, when looking at more disaggregated levels, for example below the municipality level or even at the unit level, the CSDD was assessed against the Census Response Database or the published Census counts. This part of the assessment is not covered in this document.

8. Coverage studies are carried out after each census to estimate how many individuals were missed (undercoverage) or counted more than once (overcoverage). Census net undercoverage estimates (CNUC), that is undercoverage minus overcoverage, are derived by sex, age group and province or territory. Statistics Canada then produces official PEs by sex, age and geographic areas down to Census subdivisions[11] from the published Census counts further adjusted for CNUC. The CSDD fitness to PEs at various levels of geography, age group and sex will thus be compared in the following through a variety of data quality indicators and graphical visualization tools.

[9] Population and Family Estimation Methods at Statistics Canada (91-528-X)

[10] Portion of Canada's land area governed by a political authority. Canada is divided into 10 provinces and 3 territories.

[11] A Census subdivision is a municipality or an area that is deemed to be equivalent to a municipality for statistical reporting purposes. For more information on Canada's geography, please refer to http://www.statcan.gc.ca/pub/92-143-g/2011001/app-ann/app-annb-eng.htm

IV. Methods for data quality assessment

9. The quality of the CSDD population counts at various levels of geography is first assessed by looking at a selection of data quality indicators (DQI). DQIs are used to look at the importance of missing values, and the proportion of small, moderate or large differences between the population counts derived from the CSDD or the published Census counts with the PEs. Overall average measures of differences for key demographic or geographic variables are also considered in the validation. Various data visualization tools are used in the second part of this evaluation to confirm results from the DQI, but mainly to identify outliers, in particular extreme ones.

A. DQI: Measures of Average Error

10. The three measures of average differences with PEs are: the Mean Absolute Percentage Error (MAPE), the Median Absolute Percentage Error (MedAPE) and the Weighted Mean Absolute Percentage Error (WMAPE). These measures are based on the percentage error of alternative population counts (CSDD, published Census) with PEs at a certain level of geography (e.g., across the thirteen provinces and territories). The use of MAPE in population accuracy studies[12, 13] (forecast or estimates) 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 (APE), can be used to complement MAPE since it is less affected by outliers. It is calculated as:

$$MedAPE = \ sorted\ APE\ of\ \frac{n}{2}\ (if\ n = pair) * 100$$

$$MedAPE = \ sorted\ APE\ of \frac{n+1}{2}\ (if\ n = unpaired) * 100$$

11. An additional weakness of MAPE is that it gives an equal weight to each observation 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 per cent error in a population of 1,000,000 has more impact on the WMAPE than a 25 per cent error in a population of 1,000. This is critical for Canada as population size varies considerably between provinces and territories, ranging from more than 13 million for Ontario, to less than 150,000 for Prince Edward Island, and less than 50,000 for each of the territories. A good fit with the benchmark is assumed when the indicator is close to 0 per cent.

$$WMAPE = \sum_i \left(\left| \frac{F^i - A^i}{A^i} * 100 \right| * \frac{A^i}{\sum_i A^i} \right), \text{ where } \left(\frac{A^i}{\sum_i A^i} = \text{population weight}\right)$$

where F denotes the alternative population (CSDD or Census) and *A* the actual value of PEs for geographic area i.

[12] 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

[13] 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.

B. DQI: Measures of Distribution Error

12. Alternative population distribution errors are derived with Absolute Percentage Errors (APEs) at various levels of geography (see table 2) from larger (provinces/territories) to smaller geographies (Census subdivisions or municipalities), or by the population size of Census Divisions[14] such as:

(a) Small - where APEs are within 2.5 per cent

(b) Moderate - where APEs extend from 2.5 to 4.9 per cent

(c) Large - where APEs is equal or greater than 5 per cent

(d) Missing - where population from alternative source is null.

C. Visualization tools

13. Data quality indicators provide an overall assessment of quality within a given level of geography. However, they do not allow for the identification of outlier differences between alternative populations, either the CSDD or the Census, and the PEs for specific areas. Innovative graphical visualization tools that are less commonly used at Statistics Canada are therefore being explored to provide alternative ways to assess and evaluate data quality. Statistics Canada has recently started to use circular plots[15,16] to present interprovincial migration flows in Canada, which are generally presented in migration matrices, in addition to commonly used tools for assessing data quality and performing analyses such as scatter plots, bar charts, histograms and tables. However, some of the limitation of these tools is that only a restricted number of observations can be presented before becoming unreadable or presenting visual overlap of multiple points. The number of variables that can be presented in these tools is usually limited to two, as multiple aggregations (e.g., higher and lower levels of geography) cannot be shown effectively, and no relationship with population size can be efficiently displayed.

14. In Section V, two graphical data visualization tools will be used to assess and compare differences between the CSDD or Census of population counts with the PEs for the 5,253 Census Divisions (CD) of Canada's 10 provinces and 3 territories. These visuals have been produced using SAP Lumira[17] version 1.29.3. For example, the first type of figures are displayed in figures 2 and 4, in the annex, which are enhanced scatter plots called Lumira bubble charts. These are simply scatter plots with dot sizes enhanced relative to the population size. In figure 2, for example, provincial and territorial population differences with PEs are plotted for the CSDD on the vertical axis, and for the Census on the horizontal axis.

15. The second type of visualization tool is called treemaps, which is a space-filling data representation method for hierarchical datasets. Treemaps were developed in the 1990s to

[14] Group of neighbouring municipalities joined together for the purposes of regional planning and managing common services (such as police or ambulance services).

[15] For more details and information on the origin and interpretation of these graphs programmed in R, readers are encouraged to consult the articles from the Vienna Institute of Demography: Sander et al. (2014), "Visualizing Migration Flow Data" and Abel (2015), "Estimates of Global Bilateral Migration Flows by Gender between 1960 and 2010".

[16] http://www.statcan.gc.ca/pub/91-215-x/2015000/ct013-eng.htm

[17] http://go.sap.com/product/analytics/lumira.html

display computer hard disk contents[18] and have proven useful for data visualization[19]. For example, figure 5a in the annex presents the differences between the Census and PEs in percentage by Census Division (CD), separated by province and territory. The treemap displays two main characteristics: the area of rectangles representing the population size (i.e. from PEs) and the colour of rectangles representing differences between the Census and the PEs. The different shades of red mean that the Census counts are lower than the PEs (it is known that the Census is mainly affected by under coverage). The few shades of blue mean the opposite, and the darker the colours are, the larger the differences.

16. Canada has a very large landscape (close to 9 million km^2) and a small population, just over 34 million according to the 2011 PEs, which leads to an overall relatively small population density (3.8 inhabitants per km^2). Standard mapping techniques (e.g. using ARCGIS) make it difficult to graphically represent the same information as in the ~~tree map~~treemaps with the same levels of geography, since population density increases substantially in a few cities (Toronto, Mississauga, Montreal, Vancouver). For example, Maps 1 and 2 in the annex display population differences between the CSDD and PEs by CD for Western and Eastern Canada (same data as in figure 5b). Although ~~tree map~~treemaps do not allow for the geographical representation of a CD's location on a map, they do classify it in the right unit (e.g. provinces), provide insight on its population size ranked in descending order, and display information for all units. Areas in which the CSDD overestimates the population, represented by the blue areas in maps 1 and 2, are barely visible on regular maps. This encompasses areas that take up a small part of Canada's land, but which have very large population sizes, and thus high population density (Toronto (2.7 M), Vancouver (2.4 M), Montreal (1.9 M), and Mississauga (1.3 M)). ~~Tree map~~Treemaps are also helpful for detecting different patterns, errors in a dataset, or different trends in data series through the comparison of two ~~tree map~~treemaps.

V. Data quality evaluation

A. Missing information

17. The CSDD file contains partial or complete information for 35,178,827 individuals. For the provincial and territorial analysis, 799,679 records were excluded if at least one of the three main variables (age, sex or province/territory) was missing. The resulting file thus contains 34,379,148 individuals with complete or plausible information for age, sex and province or territory. At the subprovincial level, an additional 1.6 million records (4.6%) of the CSDD could not be assigned a valid geography code, mainly due to data sources that only provide a correspondence address, as opposed to a residential address.

B. Overall data quality

18. Table 1 presents the overall population counts for the CSDD, the Census and PEs. The CSDD contains valid demographic information at the provincial/territorial level for 34,379,148 individuals, a count slightly higher (+0.3 per cent) than PEs (34,273,205). On the other hand, population counts from the Census are below PEs (-2.3 per cent).

[18] Shneiderman, B. (1992). Tree visualization with treemaps. A 2d space-filling approach. *ACM Trans*. Graph., 11(1):92-99.

[19] Tennekes, M. and de Jonge, E. (2011). Top-down data analysis with treemaps. *In Proceedings of the International Conference on Information Visualization Theory and Applications, IVAPP 2011*

Table 1
CSDD and Census population counts compared to Population Estimates, Total population, Canada, 2011

CANADA		Number			% of Population Estimates			Sex ratio
		Total	Male	Female	Total	Male	Female	% Male for 100 Females
	Pop. Estimates	34,273,205	16,977,217	17,295,988				98.2
	Census	33,476,688	16,414,229	17,062,459	-2.3%	-3.3%	-1.4%	96.2
	CSDD	34,379,148	17,067,560	17,311,588	0.3%	0.5%	0.1%	98.6

Note: CSDD does not include missing sex, provinces and/or age

Sources: 2011 Canadian Statistical Demographic Database, 2011 Census and 2011 Censal Population Estimates, Statistics Canada

19. The CSDD counts for both males and females are slightly above PEs by 0.5 and 0.1 per cent, respectively. In that matter, they perform better than the Census, which underestimates PEs for both male (-3.3 per cent) and female (-1.4 per cent) populations. The sex ratio, calculated as the number of males per 100 females, is also closer for the CSDD (98.6) when compared to PEs (98.2) than for the Census (96.2). The difference between the CSDD and PEs of less than half a percent is promising, but it masks overrepresentation in certain groups and underrepresentation in others. This issue will be presented in the following sections.

C. Overall data quality by age

20. Figure 1 (in the annex) shows the age distribution for the CSDD, the Census and the PEs by single year of age, for ages 0 to 99 and for centenarians (100+). For the three data series, the general shape of the curves is similar, with some discrepancies at specific ages. Overall, the population size increases with age, until the mid 40s to early 50s, after which population size rapidly decreases due to the rise of mortality hazards at those ages, leading to population counts below 100,000 after age 85. This figure also presents differences in percentage (right scale) between the alternative populations and the PEs. Compared to PEs, the Census underestimates population counts overall, and the differences in absolute value for most ages are smaller with the CSDD. The Census underestimates population for most ages (14 to 56) with larger differences, more than 5% below, among young adults aged 21 to 36. The CSDD counts are within 5% (over or under) of PEs for ages 0 to 95, with the exception of age 0 (-9.8%)20 and ages 18 to 20. The CSDD systematically overestimates the population aged 23 to 63, while it underestimates the population aged 64 to 89. Above age 95, both the Census and the CSDD greatly overestimate the PEs, which benefit from additional old-age adjustment recently developed and based on the extinct cohort method to estimate these populations only using mortality data.

D. Overall data quality for provinces and territories

21. Figure 2 (in the annex) presents differences in percentage between the PEs, the CSDD and the Census for all provinces and territories in a bubble chart where the size of

[20] The source of the difference has been identified: too many births were removed from the initial CSDD because they did not display any fiscal activity, a rule that was useful in identifying potential emigrants and unknown deaths for the majority of the population but not for the recent births.

the dot represents the size of the population provided by the PEs. The CSDD is within 1.0% for all provinces, except for British Columbia (+1.9%) and Nunavut (-1.2%). The Census counts systematically underestimate the PEs for all provinces and territories, ranging from -0.5% for New Brunswick to -6.2% for Nunavut. For the two most populated provinces, Ontario (38.6% of total Canadian population) and Quebec (23.3% of total Canadian population), the CSDD counts are closer to the PEs than the Census counts are. For the third most populated province, British Columbia (13.1% of total Canadian population), the CSDD is closer (+1.9%) to the PEs, but in the opposite direction of the Census (-2.0%). For that level of geography, the CSDD (0.7%) has on average a lower WMAPEs that is closer to the PEs than the Census is (2.3%).

E. Measures of average error by age group

22. WMAPEs by age groups are presented in Figure 3 (in the annex). Below age 18, differences are small between the CSDD and the Census when compared to PEs. However, for the age groups 18 to 24 and 25 to 39, the CSDD performs better on average for all provinces and territories with lower WMAPEs than the Census. The Census performs better for all age groups 40 years and over, in particular for the population aged 80 to 99.

F. Measures of distribution errors

23. Table 2 provides a comparison of measures of distribution errors between the PEs and the CSDD and the Census at different levels of geography, from the higher (provinces and territories) to the lower (Census subdivisions or CSD) levels. As presented previously, the overall fitness at the provincial and territorial levels is better for the CSDD than the Census, with lower WMAPE and all 13 entities having small APEs (less than 2.5%). This is compared to seven entities having small APEs for the Census. At lower levels of geography, these indicators favour the Census more than the CSDD. As the levels of geography decrease, WMAPEs and proportions of larger differences (i.e. APE equals or greater than 5%) increase for the CSDD but are relatively stable for the Census. For example, the proportion of large differences for the CSDD ranges between 29.8% for Census Metropolitan Areas (CMA)[21] and areas outside of CMA (i.e., 13 non-CMAs, one for each province or territory), and 59.0% for Census Divisions (CD). For the Census, these values range from 2.1% at the CSD and CMA levels to 7.7% at the provincial/ territorial level. For the CSDD, WMAPEs fluctuate from 0.7% for provinces or territories to 9.0% at the CSD level, whereas these indicators are more stable and much lower (around 2.3%) for the Census.

[21] A Census Metropolitan Area (CMA) is an area consisting of one or more neighbouring municipalities situated around a core. A CMA must have a total population of at least 100,000 of which 50,000 or more live in the core. In 2011, there were 34 CMAs defined in the Standard Geography Classification (SGC 2011). For more detail, please refer to http://www12.statcan.gc.ca/census-recensement/2011/ref/dict/geo009-eng.cfm

Table 2

Proportion of APEs: small (below 2.5%), moderate (from 2.5% to 4.9%) or large (equal or greater than 5%), missing value and WMAPE of alternative populations (CSDD, Census) with Population Estimates at various levels of geography

		Census	CSDD	Census	CSDD	Census	CSDD
		Number		Proportion		WMAPE	
Prov/terr	APE < 2.5%	7	13	53.8%	100.0%	2.3%	0.7%
	2.5% <= APE < 5.0%	5	0	38.5%	0.0%		
Total = 13	APE >= 5.0%	1	0	7.7%	0.0%		
	Empty or error	0	0	0.0%	0.0%		
CMA and outside CMA	APE < 2.5%	21	21	44.7%	44.7%	2.3%	5.3%
	2.5% <= APE < 5.0%	25	12	53.2%	25.5%		
Total = 47	APE >= 5.0%	1	14	2.1%	29.8%		
	Empty or error	0	0	0.0%	0.0%		
CD	APE < 2.5%	211	61	72.0%	20.8%	2.3%	5.8%
	2.5% <= APE < 5.0%	74	59	25.3%	20.1%		
Total = 293	APE >= 5.0%	8	173	2.7%	59.0%		
	Empty or error	0	0	0.0%	0.0%		
CSD	APE < 2.5%	3,466	646	66.0%	12.3%	2.5%	9.0%
	2.5% <= APE < 5.0%	985	460	18.8%	8.8%		
Total = 5,253	APE >= 5.0%	108	2,940	2.1%	56.0%		
	Empty or error	694	1,207	13.2%	23.0%		

Sources: 2011 Canadian Statistical Demographic Database, 2011 Census and 2011 Censal Population Estimates, Statistics Canada

24. Quality indicators can also be calculated by population size to assess comparability of either the CSDD or the Census with the PEs (Table 3). Similar conclusions can be drawn from this table: CSDD data quality indicators (DQI) increase as population size decreases. For example, the CSDD WMAPE is equal to 3.9% for CDs with PE of 85,000 or more, 10.5% for PE between 40,000 and 85,000, and goes up to 18.8% for CDs with PE of less than 20,000. For the Census, DQIs do not show this pattern. The WMAPE is more than 10 times lower than the CSDD for a CD with PE less than 20,000, and 60% lower for a CD with PE of 85,000 or more. A similar conclusion can be drawn from Figure 4 (in the annex) which presents percent differences between the PEs with the CSDD and with the Census in a bubble chart for 293 CDs, where the size of the dot represents the size of the population provided by the PEs, and the colour refers to each province or territory. For the CSDD, larger CDs (in terms of population) are closer to PEs than smaller ones, and the range of the differences are a lot greater, from -71.9% to 8.6%, compared to a range of -17.3% to 0.8% for the Census. For a majority of CDs (90%), differences between the Census and the PEs are contained between -0.5% and -3.2%, compared to differences of 0.2% to -31.1% for the CSDD.

Table 3
MAPE, MedAPE and WMAPE by Census Division population size (quartile) of Population Estimates with alternative populations (CSDD, Census)

	MAPE		MedAPE		WMAPE		
	Census	**CSDD**	**Census**	**CSDD**	**Census**	**CSDD**	n
Population < 20,000	1.8%	18.2%	1.5%	9.9%	1.6%	18.8%	68
20,000 <= *Population* < 40,000	1.7%	14.0%	1.2%	9.8%	1.7%	13.7%	80
40,000 <= Population < 85,000	2.0%	10.9%	1.6%	6.4%	2.1%	10.5%	72
Population >= 85,000	2.2%	6.3%	2.5%	3.4%	2.4%	3.9%	73
Total	1.9%	12.3%	1.7%	7.1%	2.3%	5.8%	293

Sources: 2011 Canadian Statistical Demographic Database, 2011 Census and 2011 Censal Population Estimates, Statistics Canada

G. Data quality assessment using treemaps

25. The analysis of WMAPEs discussed above provides insight into the quality of the CSDD and information about its general strengths and weaknesses for specific age groups or levels of geography. For example, as the level of geography and/or the population size decrease, the differences of the CSDD with the PEs increase significantly and are larger than those observed between the Census and PEs. Treemaps, which are a space-filling data visualization method, can also be used to locate areas (specific geographies or age groups) where differences are larger than in others. It can also take into account the sign of these differences which are not shown in absolute value when WMAPEs are calculated. Figures 5a and 5b present, respectively, differences in percentage with the PEs of either the Census (5a) or the CSDD (5b) by Census Division (CD) and province or territory. The size of the box represents the population size from PEs for CDs (the name of the largest city within a CD is displayed for larger CD), and the colour represents differences between the two data series. Shades of red indicate when either the Census or CSDD counts are lower than the PEs, and shades of blue indicate when they are larger, with darker shades indicating increases in differences.

26. In Figure 5a, Census counts for almost all CDs are lower (shades of red) than the PEs. For most CDs in Ontario and British Columbia differences are below -1%, and even between -3% and -5% for Toronto and most CDs in Alberta. A few small CDs (in terms of population size) display larger negative (< -5%) differences that are explained partly by the fact that they include Indian Reserves that were not enumerated in the Census, but which were estimated in the PEs. On the other hand, the treemap for the CSDD (Figure 5b) is more colourful, displaying both positive (blue) and negative (red) differences with the PEs. As presented previously, as the population size of a CD decreases, WMAPEs increase and the differences in absolute value are significantly larger for the CSDD. At the national level, these underestimations are partially compensated by a systematic overestimation (darker shades of blue) of populations for larger CDs in Canada, such as in Toronto, Montreal, Vancouver, Markham, Mississauga and Winnipeg.

27. The treemaps presented in Figures 6a and 6b now display differences in percentage with the PEs of either the Census (6a) or the CSDD (6b) by age group and province or

territory. The conclusion on the Census treemap is that it also tends to underestimate, by at least 3% the population for age groups 18 to 24 and 25 to 39. However, with this type of graphical data representation, we can see that this is the case for all provinces and territories. A similar chart for the CSDD now displays the fact that the differences, as presented in Figure 3 with WMAPES, are not only lower than for the Census, but that they are positive, showing a mark overestimation of populations aged 25 to 39 and 40 to 64 that get compensated at the national level by the underestimation of other age groups. These overestimations are even more pronounced in British Columbia. This could be related to difficulties encountered by the CSDD in identifying, and therefore excluding, emigrants or people living abroad.

VI. Conclusion

28. Overall, the CSDD population counts are closer to the PEs than the Census counts are to the PEs. This is also true for each sex. Data quality indicators presented in this paper demonstrated that the CSDD is, on average for all provinces and territories, closer to the PEs than the Census is. However, as levels of geography and/or population sizes decrease, the population counts from the Census are a better fit to the PEs. Similar conclusions can be drawn from data analysis using treemaps or bubble charts. Treemaps provide a strong graphical representation that contrasts overestimation in larger geographical units that gets partially compensated by strong underestimation in smaller units. Much work remains to be done before the CSDD can be used for purposes other than research, as the main goal of the Census is to provide data at a finely disaggregated level of geography.

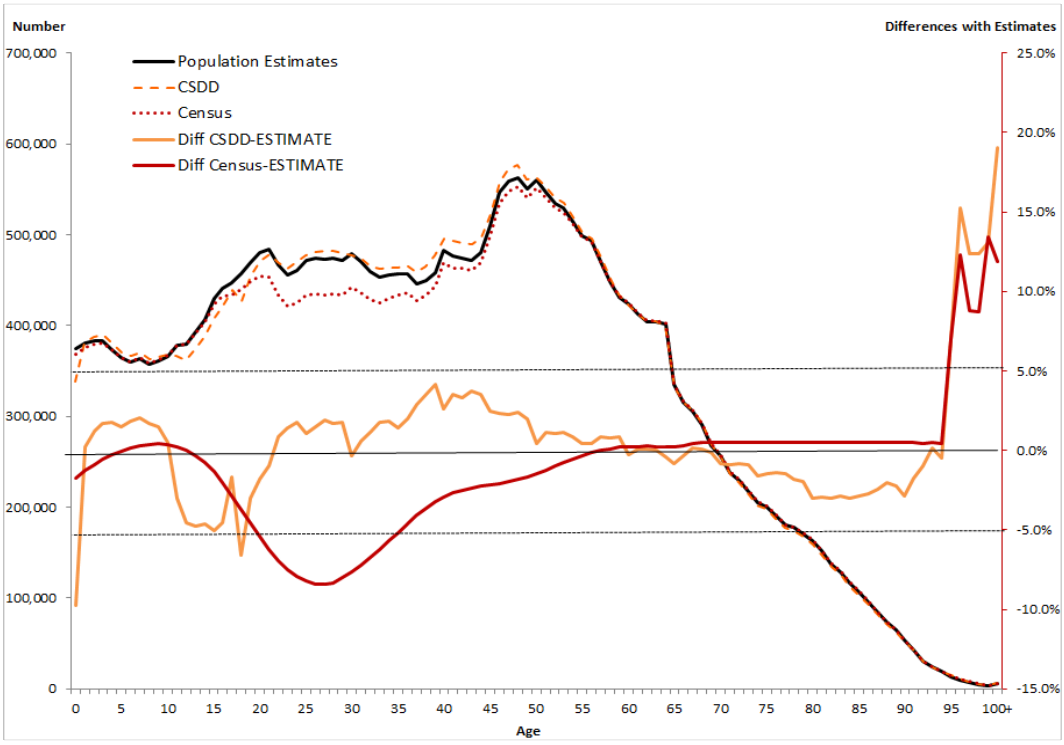
29. For the next CSDD iteration in 2016 and beyond, a few areas of improvement will be explored based on the research provided in this paper. Many of these improvements rely on the access to additional administrative data sources. In order to reduce the overcoverage of young adults, the CSDD will keep working on methods to better estimate emigration with the Fiscal Activity Indicator, specifically in larger municipalities such as Toronto and Vancouver. Efforts will also be made to investigate and reduce the overcoverage of people above the age of 90 using new data sources, which have recently become available. To improve fitness at lower levels of geography, Statistics Canada will continue to work on gaining access to provincial and territorial administrative data and utility data that are more likely to include a residential address (versus a mailing address), and to improve the CSDD's ability to put people in the right place.

30. Quality indicators displayed in this paper, either aggregated (WMAPE or APE) or graphical (bubble charts or treemaps), provide a comparison of data quality for key demographic variables and various levels of geography. The preferred data visualization tool for our analysis is treemaps, as they have effectively provided ways to identify important population differences for alternative populations, which are often hidden, with commonly used aggregated quality indicators such as WMAPEs or APEs.

VII. Annex

Figures and Maps

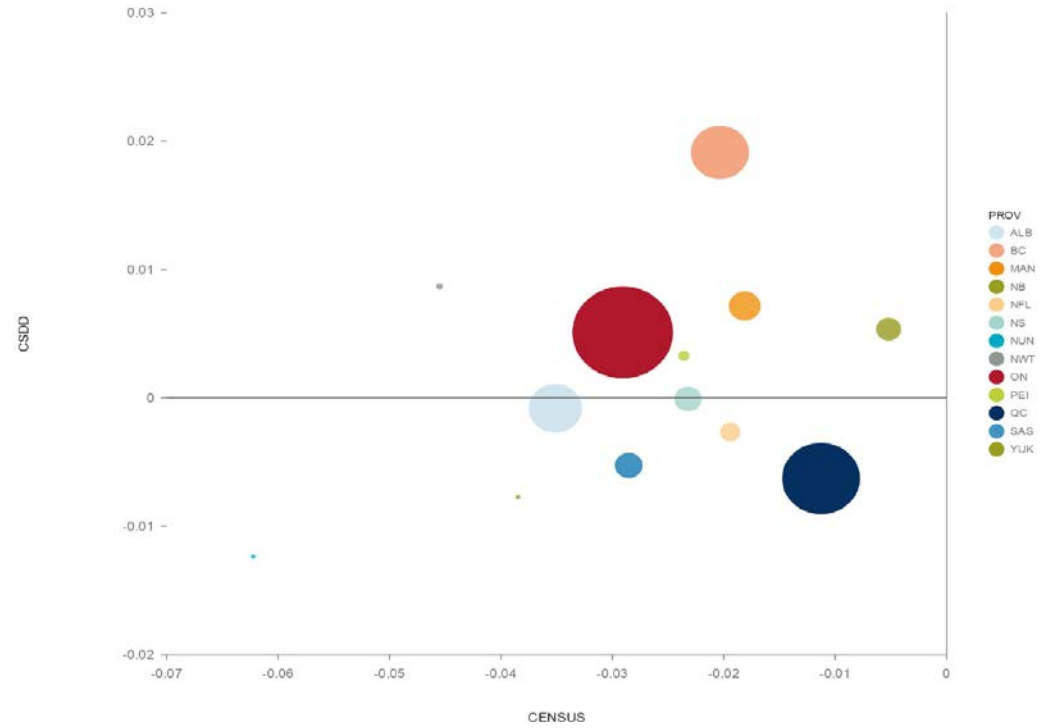
Figure 1
Age distribution of CSDD, Population Estimates and Census (left scale) and difference with Population Estimates (right scale), 2011



Sources: 2011 Canadian Statistical Demographic Database, 2011 Census and 2011 Censal Population Estimates, Statistics Canada

Figure 2

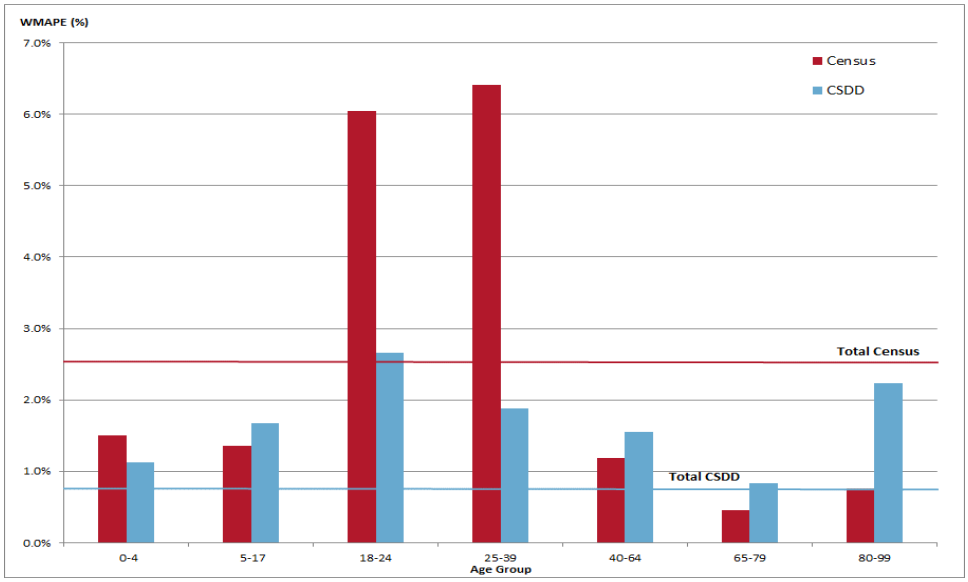
Difference with Population Estimates by Province and Territory for the CSDD and the Census



Sources: 2011 Canadian Statistical Demographic Database, 2011 Census and 2011 Censal Population Estimates, Statistics Canada

Figure 3

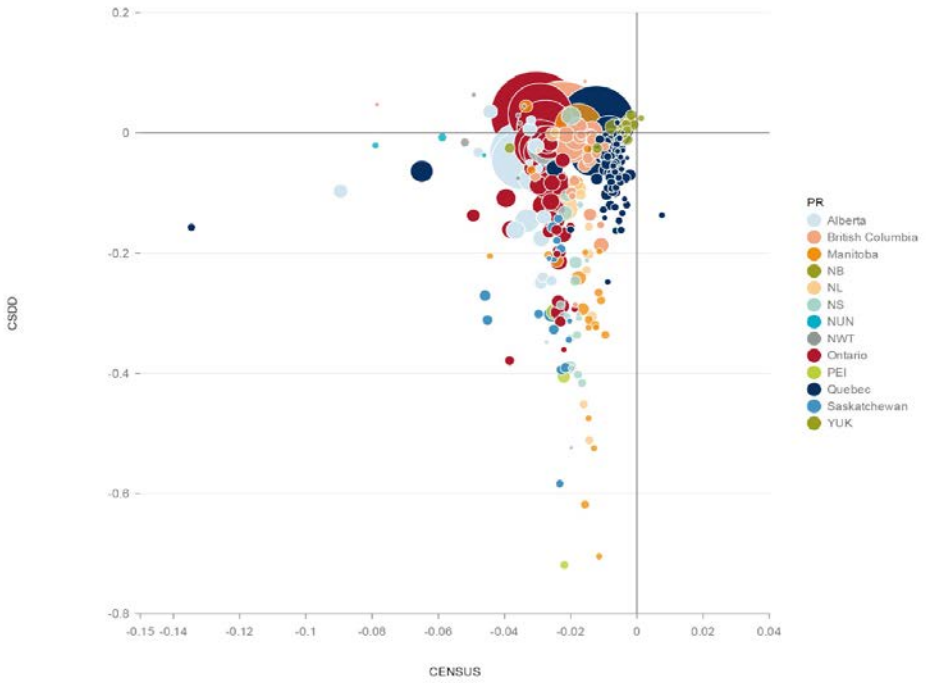
Provinces and Territories WMAPEs by age groups for Census and CSDD



Sources: 2011 Canadian Statistical Demographic Database, 2011 Census and 2011 Censal Population Estimates, Statistics Canada

Figure 4

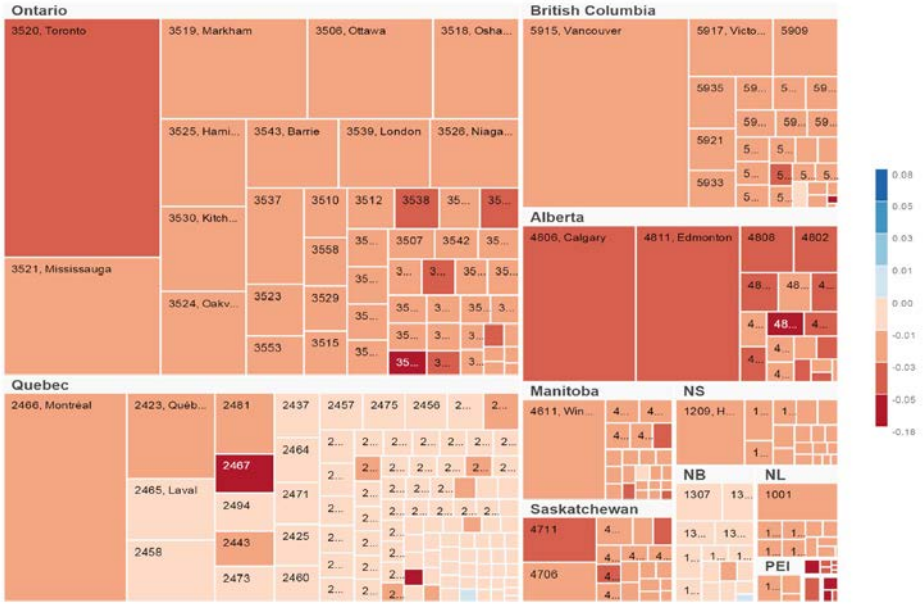
Differences with Population Estimates by Census Division (CD) for the CSDD and the Census.



Sources: 2011 Canadian Statistical Demographic Database, 2011 Census and 2011 Censal Population Estimates, Statistics Canada

Figure 5a

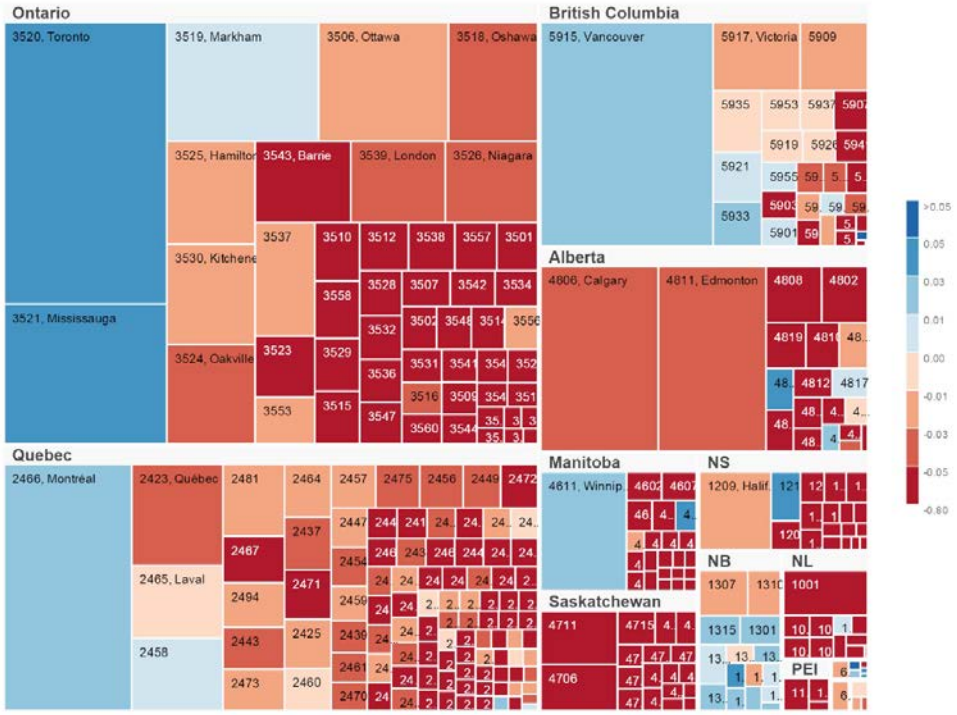
Differences between the Census and the Population Estimates by Province and Census Division, 2011



Sources: 2011 Census and 2011 Censal Population Estimates, Statistics Canada

Figure 5b

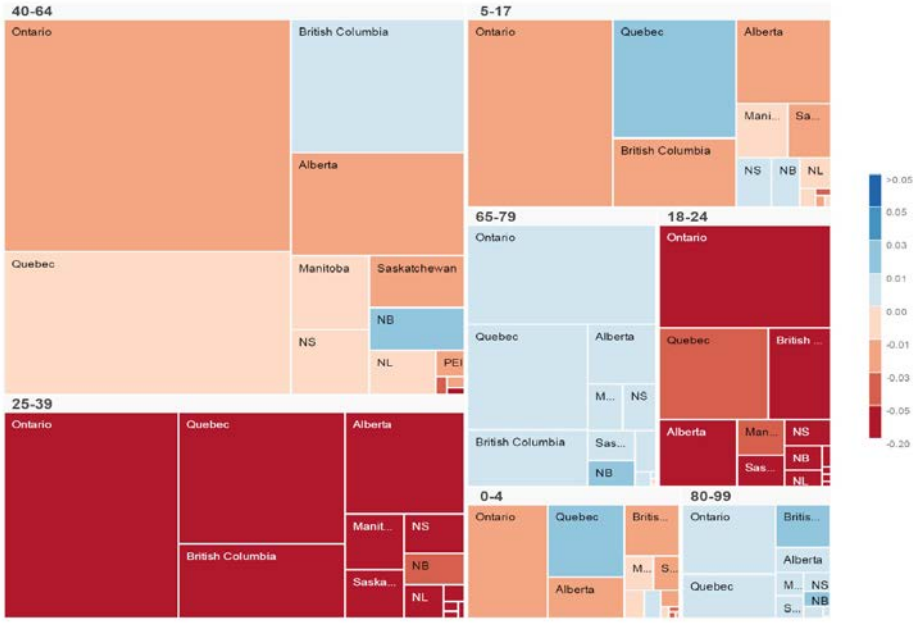
Differences between the CSDD and the Population Estimates by Province and Census Division, 2011



Sources: 2011 Canadian Statistical Demographic Database and 2011 Censal Population Estimates, Statistics Canada

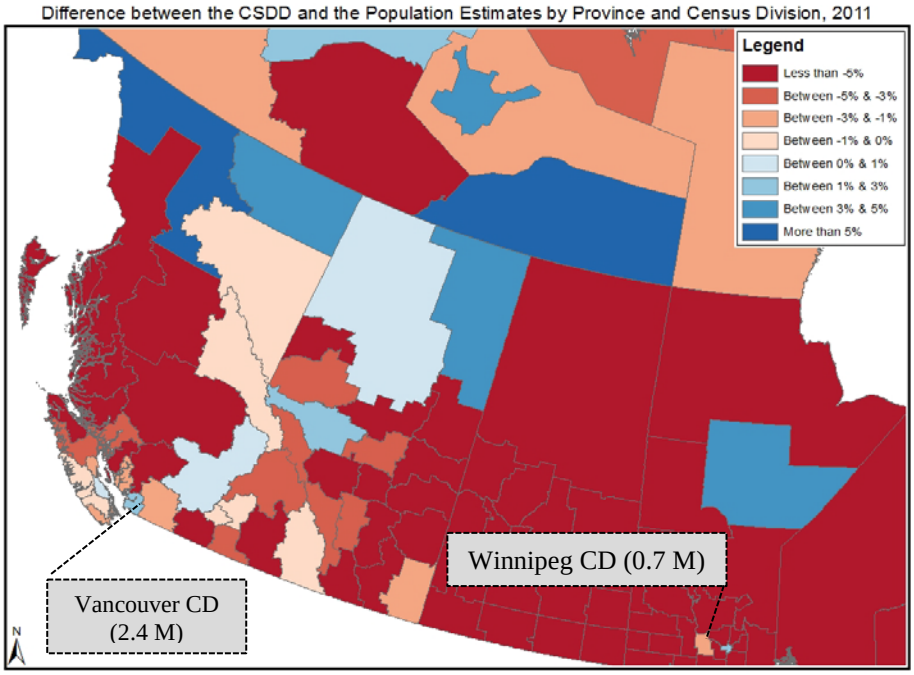
Figure 6a

Differences between the Census and the Population Estimates by Province and Age Group, 2011



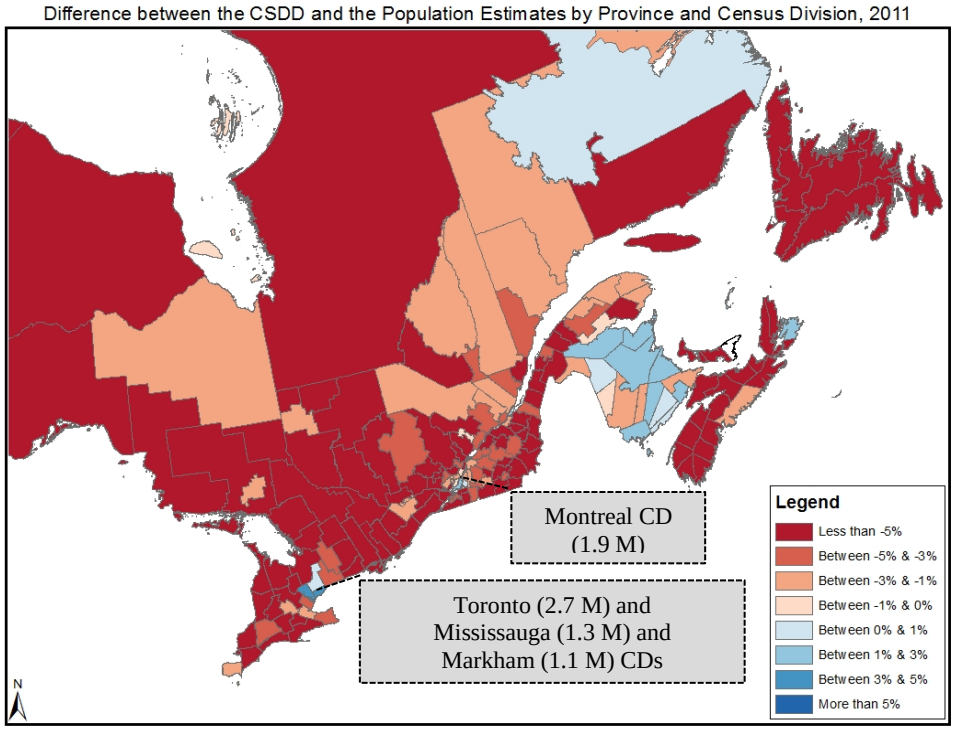
Sources: 2011 Census and 2011 Censal Population Estimates, Statistics Canada

Map 1: Western Canada



Sources: 2011 Canadian Statistical Demographic Database and 2011 Censal Population Estimates, Statistics Canada

Map 2: Eastern Canada



Sources: 2011 Canadian Statistical Demographic Database and 2011 Censal Population Estimates, Statistics Canada