Assessing the Usability of a Statistical Population Register for the Census of Population in Canada

UNECE-Eurostat Expert Meeting on Census
Geneva, Switzerland, September 2016

Outline

- Context
- Research Project
- Data Quality Assessment: Objectives and indicators
- Results

Context

- Traditional censuses held in Canada every ten years since 1871, and every five years since 1956
- Pressures on current census methodology
- Investigation of alternative enumeration strategies
- Research into a Canadian Statistical Demographic Database (CSDD)

CSDD Research Project

- Comprehensive review of the potential for administrative and other alternative data sources to complement or supplement the agency's census and surveys
- Feasibility of developing a virtual population register by using/linking multiple administrative data sources

Data Quality Assessment:
Objectives

- Identify and develop data quality indicators to be able to assess the CSDD against two population sources, the Census and the Censal Population Estimates (PEs) Program, at various levels of geography and by age and sex
- Target strengths and weaknesses (i.e. age, sex, geography) of the CSDD and propose areas of improvement for the next iteration
- Ongoing evaluation of CSDD data quality for each iteration and period of reference (2011 and soon 2016) using same quality indicators and data visualization tools

Data Quality Indicators (DQI)

- **Measurement of Average Errors (Differences)**
  - Absolute Percentage Error (APE)

$$APE^i = \left|\frac{F^i - A^i}{A^i} * 100\right|$$

  - Weighted Mean Absolute Percentage Error (WMAPE).

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

Where F denotes an alternative population (Census or CSDD) and A is the PEs for geographic area i.

- **Detection of outlier values and percentage errors using data visualization support**
  - Treemap produced using SAP Lumira

Overall Quality

The CSDD contains demographic information for 34,4 millions individuals, a count slightly higher (+0.3%) and closer to PEs than the Census (-2.3%). The sex ratio is also closer to PE (98.2) for the CSDD (98.6) than for the Census (96.2)

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

	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

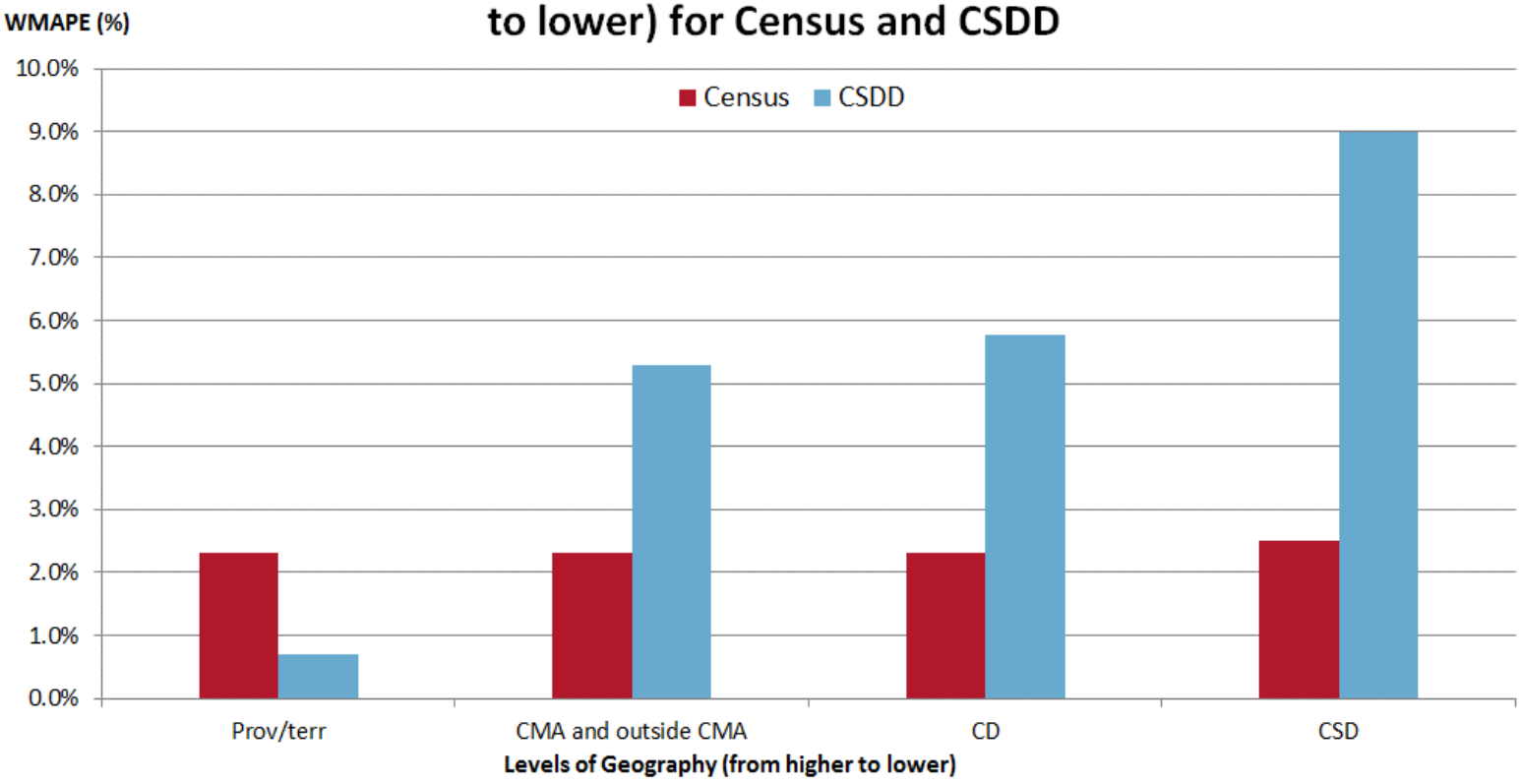
Note: CSDD does not include missing sex, provinces and/or age

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

Weighted Mean Absolute Percentage Error by Level of Geography

Overall fitness at the provincial and territorial levels is better for the CSDD with lower WMAPE; below that level of geography (CMA, CD and CSD), Census counts are closer to Population Estimates

Figure 1 WMAPEs by Various Level of Geographies (from higher to lower) for Census and CSDD

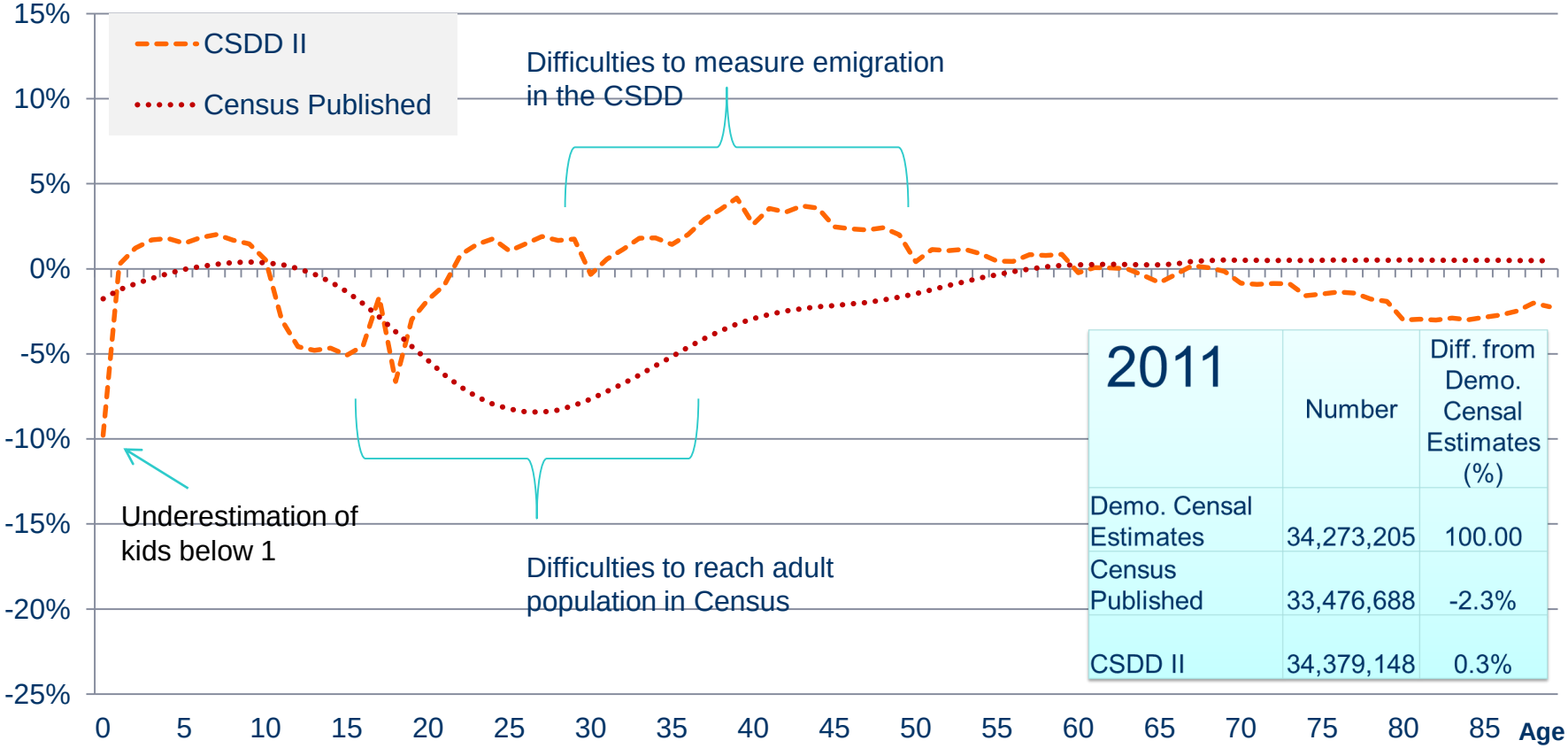


Sources: 2011 Canadian Statistical Demographic Database, 2011 Census and 2011 Censal Population Estimates, Statistics Canada

Percentage Errors (0-89)

The 2011 CSDD is within 5% (over or under) of the Population Estimates for all ages with the exception of populations aged 0 and 18. The CSDD systematically overestimates the population aged 23 to 63, while it underestimates the population aged 64 to 89.

Figure 2 Differences with Population Estimates by Age, 2011

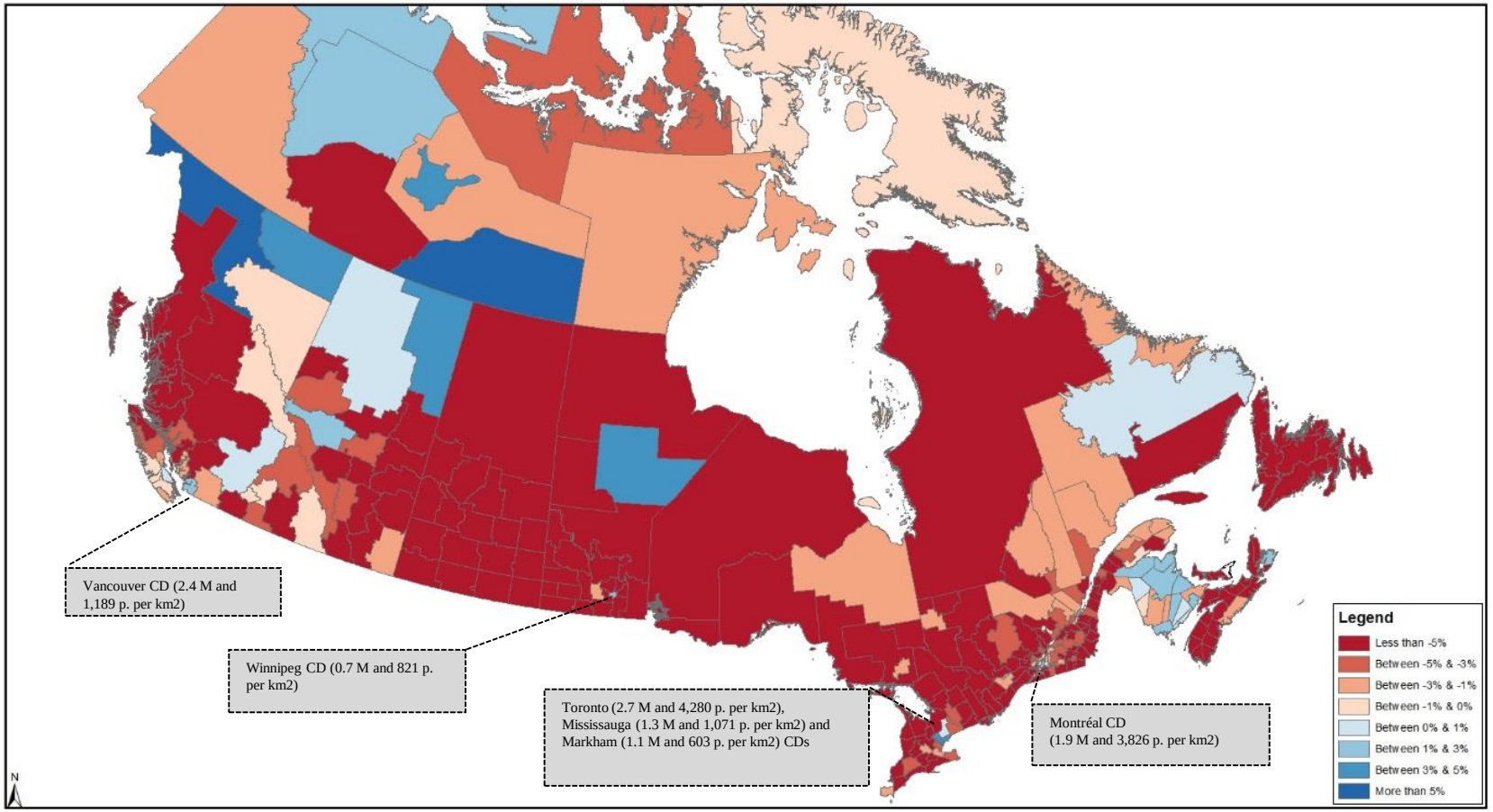


2011	Number	Diff. from Demo. Censal Estimates (%)
Demo. Censal Estimates	34,273,205	100.00
Census Published	33,476,688	-2.3%
CSDD II	34,379,148	0.3%

Data visualization

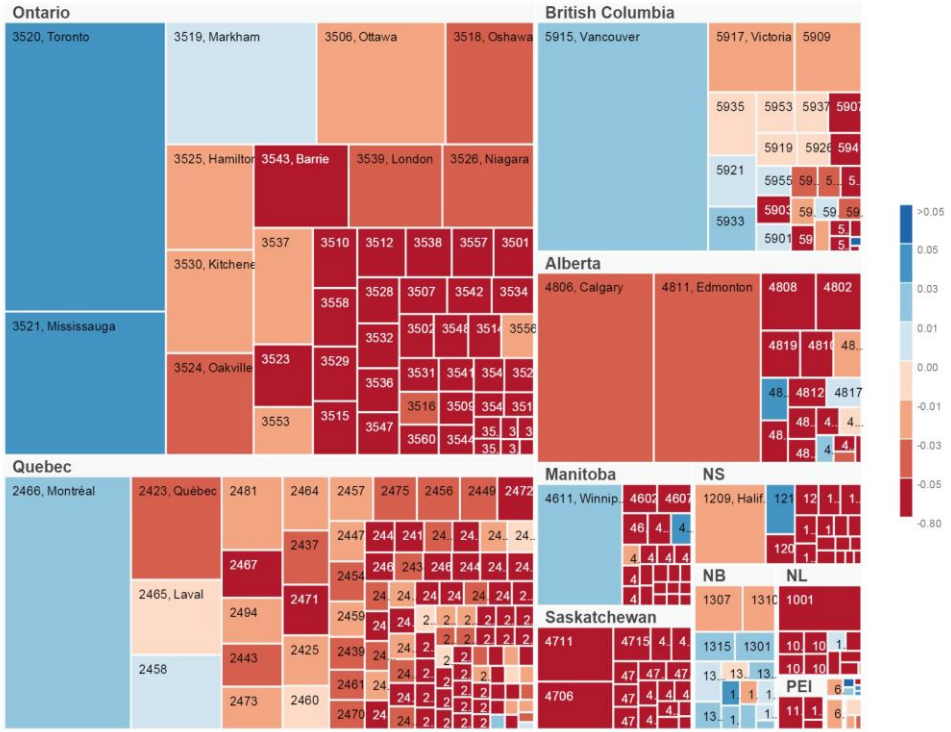
Canada has a very large landscape (close to 9 million square km) with a relative small population (34 million in 2011) for an overall low population density (3.8 inhabitants per square km).

Map 1 Differences between the CSDD and Population Estimates by Province and Census Division, 2011



Data visualization

- The visualization tool presented in the following is called treemap: a space-filling data representation method for hierarchical datasets where the area of rectangles represents the population size (i.e. from PEs) and the colour of rectangles represents differences between either the Census/CSDD and the PEs.
- The different shades of red refer to level of undercoverage while shades of blue refer to level of overcoverage.
- The darker the colours are, the larger the differences between the two populations.

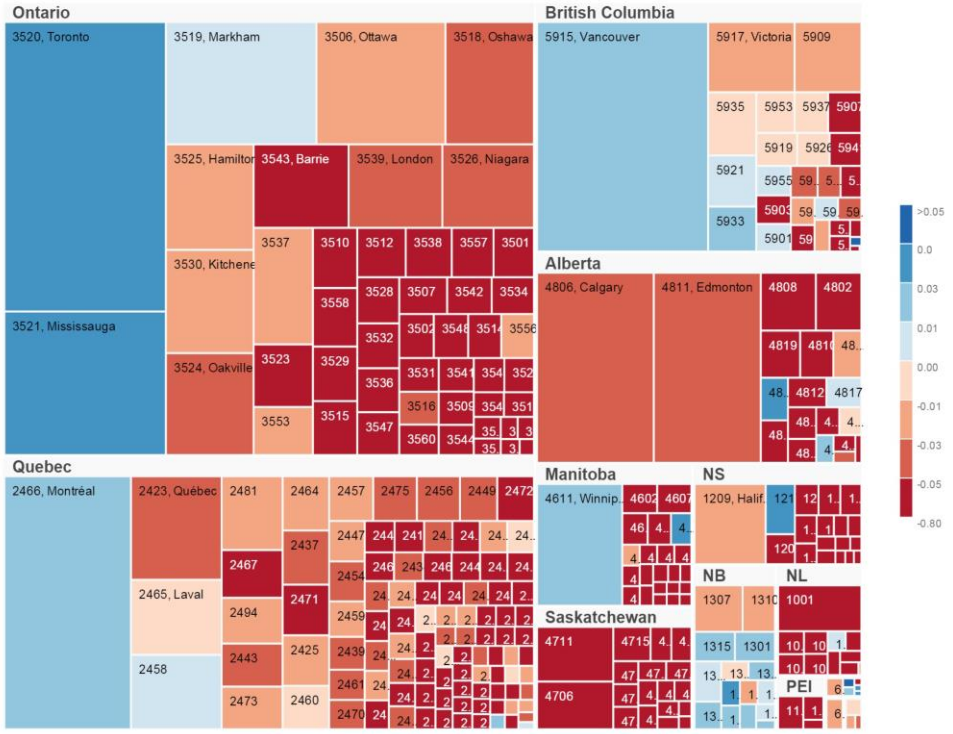


Data Visualization: Treemaps by Province and CD

For most CDs differences between the Census and PE are below 3% (Figure 3a). The CSDD treemap (Figure 3b) is more colourful, displaying both positive (blue) and negative (red) differences. As population size decreases, larger underestimation (darker red) are seen in the CSDD for most provinces which gets partially compensated at the national level by an overestimation (darker shades of blue) in larger CDs.

Figure 3a Differences between the CENSUS and PEs by Province and Census Division, 2011

Figure 3b Differences between the CSDD and PEs by Province and Census Division, 2011



Data Visualization: Treemaps by Province and Age

The Census tends to underestimate PEs by at least 5% for age groups 18 to 24 and 25 to 39 for most provinces. Differences for the CSDD are not only lower (in absolute term) than for the Census, but 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.

Figure 4a Differences between the CENSUS and the PE by Age Group and Province, 2011

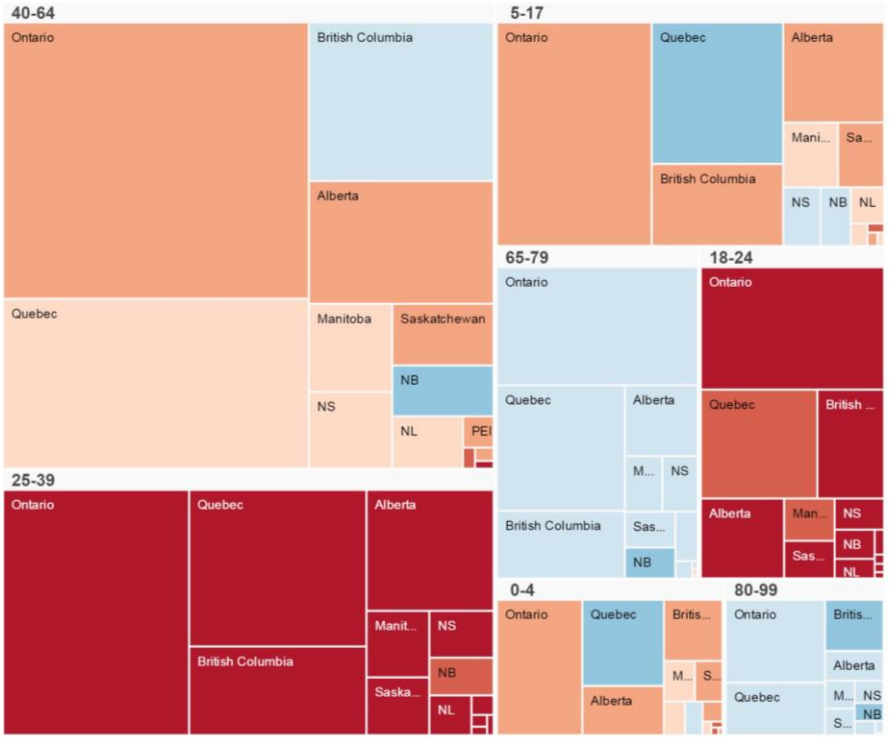
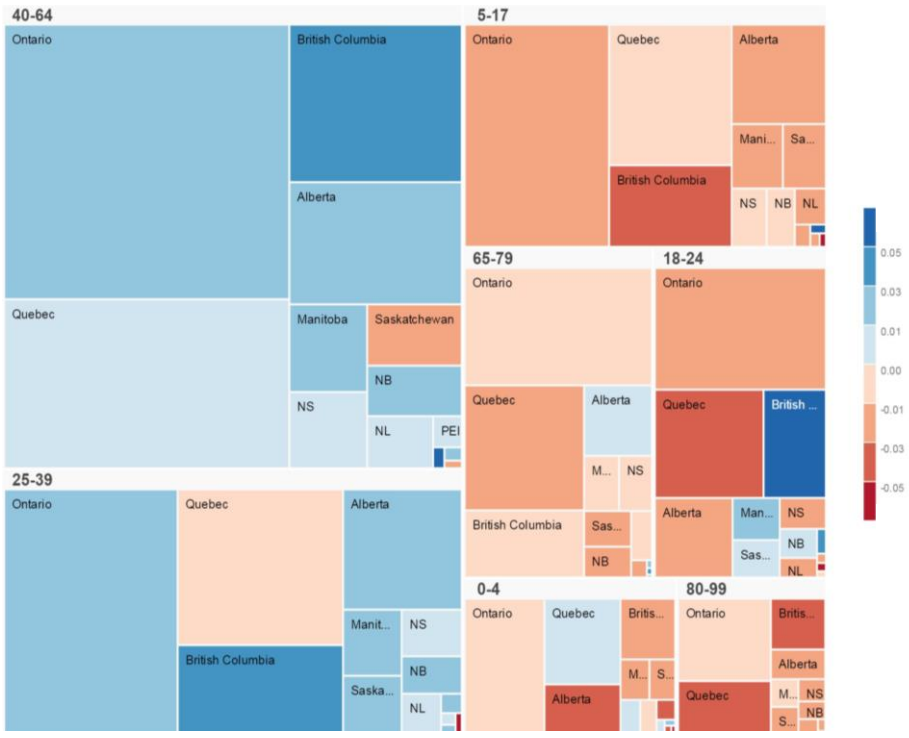


Figure 4b Differences between the CSDD and the PE by Age Group and Province, 2011



Conclusion

- Data quality indicators presented demonstrated that the CSDD is, on average for all provinces and territories, closer to the PEs than the Census is.
- As levels of geography and/or population sizes decrease, the population counts from the Census are a better fit to the PEs.
- Treemaps provide a strong graphical representation that contrasts overestimation in larger geographical units that gets partially compensated by strong underestimation in smaller units for the CSDD.
- 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.
- Statistics Canada will work on gaining access to local administrative data to improve the CSDD's ability to put people in the right place. It will also work in parallel on an exhaustive statistical register on buildings and dwellings.
- Looking forward to 2021, the CSSD could also help streamline some Census operations (e.g. non-response follow-up, processing, quality assessment, etc.).