ENVIRONICS
ANALYTICS

CensusPlus

2016 Technical Documentation

October 2018

Table of Contents

A. Product Description

What it is

CensusPlus is a database that offers a snapshot of Canadians in 2016. The database is enhanced by our modellers to fill in missing values where data are suppressed by Statistics Canada and to correct for random rounding, while maintaining a close relationship to the original census. This work ensures that there are no missing values in CensusPlus and the variables add up within thematic categories and across all levels of geography. CensusPlus includes all of the popular variables that analysts and marketers rely on from census profiles.

How It's Used

CensusPlus provides a complete, detailed set of 2016 demographic variables for all of Canada, down to the smallest geographic areas, including custom client areas. Analysts can use CensusPlus to analyze demographic and economic themes like household income, mother-tongue language, or visible-minority status to understand their customers and trades areas without worrying about effects of random rounding and suppression in the raw census. These data hierarchy add up to parent geographies and variables within themes. As an example, when the male and female population are added together they always equal the total population.

B. Product Purposes

The purpose of the product is to act as a surrogate for the equivalent themes and data in the 2016 Census from Statistics Canada at all levels of geography. This product is simpler to use and less likely to be viewed as 'wrong' by non-technical users who can be troubled by things like percentage values not adding up to 100 percent. This data product has been engineered to provide solutions or answers, rather than information.

The data in this product are adjusted for random rounding and data suppression applied by Statistics Canada, which was applied to ensure privacy of census respondents. Additionally, CensusPlus uses a consistent set of universes across themes, regardless if data are from the short- or long-form census. A detailed discussion of random rounding and suppression can be found in the Appendix.

Random rounding and suppression cause a variety of challenges when working with raw census data:

1. Data cannot be added up in any natural way within the census profile data.
2. Users must assume all values reported by the census are plus or minus five, unless the value is zero or 10. For values of 10 users should assume the actual value is in the range of one to 15. Values of zero can be assumed most commonly to be between zero and 10, but it could be any value in some cases.
3. There are missing data.
4. In many circumstances, values of zero should not be assumed to be truly zero.

5. Households may be suppressed to ensure confidentiality or to limit the dissemination of data deemed unacceptable with respect to quality by Statistics Canada.

Due to this anonymization of the census data, business analysts will spend more time justifying why data do not add up properly on a report or engineering the data to add up, rather than spending their time analyzing the real business problem at hand.

Below we highlight a typical example of the challenges with census data. Note that the geographies used are census subdivisions (CSDs) within a single census division (CD). The variable theme is Total Dwellings by Age of Primary Household Maintainers. Throughout this document, the word 'children' refers to any component parts of a large entity or 'parent'. In the example bellow, the CSDs are children geographies of the CD 3539 and the age cohorts reported are children variables of the universe total households.

Table 1 - Census Data Sample for Census Division 3539

Code	Total Households	15-24 Years	25-34 Years	35-44 Years	45-54 Years	55-64 Years	65-74 Years	75-84 Years	85 Years and over	Sum of Age Groups
3539	**190,045**	8,815	29,045	30,320	37,550	36,705	26,405	15,325	5,885	190,050
3539002	180	0	30	25	50	20	35	20	0	180
3539005	2,350	30	265	340	440	580	430	135	120	2,340
3539015	8,295	155	995	1,220	1,745	1,640	1,375	810	350	8,290
3539017	0	0	0	0	0	0	0	0	0	0
3539018	0	0	0	0	0	0	0	0	0	0
3539021	0	0	0	0	0	0	0	0	0	0
3539027	4,920	50	375	810	1,175	1,140	785	460	125	4,920
3539033	5,985	25	460	1,060	1,430	1,370	960	520	150	5,975
3539036	163,140	8,465	26,250	25,985	31,615	30,835	21,995	12,960	5,030	163,135
3539041	2,335	30	315	370	470	520	345	220	65	2,335
3539047	990	15	120	175	220	230	170	55	10	995
3539060	1,785	35	225	320	390	360	295	135	30	1,790
CHILD SUM	**189,980**	8,805	29,035	30,305	37,535	36,695	26,390	15,315	5,880	189,960

1. For the parent CD 3539, the sum of the age groups of 190,050 (in the last column) does not equal the total household value of 190,045 as published by Statistics Canada . This is a function of randomly rounding to the nearest five. These values should be the same.
2. The sum of the children CSDs does not equal the respective parent values for any of the age groups (highlighted in yellow).
3. In most cases, the sum of the age groups for each CSD row is not equal to the total dwellings.
4. Due to record suppression, the sum of total households for the CSDs is 65 less than the reported household total for the parent CD.

At a smaller geographic scale, the mismatch problems are more numerous and can be far more problematic because of increased occurrence of records and cell suppression.

C. Background and Description

This product includes the most popular themes from the standard census cumulative profile. We have also added additional themes based on client needs and our input requirements for DemoStats, our demographic estimates and projections product. Universes from the short-form census are used as our control totals for all census themes, including those themes from the long-form census. For more information regarding the unification of the two census universes, see section F3 below. There are several core census universes included in this product, such as:

- Households
- Population
- Household population
- Census families
- Children

D. Key Features

The data in CensusPlus feature:

1. The unrounded population and household counts that are part of the first release of the 2016 Census are considered the correct numbers; they are not modelled or adjusted in any way by our processes. These key universes are the starting point and benchmark for all other variables in the product.
2. All count variables are reported as integer values.
3. The numbers add up from small geographies to larger parent geographies.

4. All differences between raw census and CensusPlus are consistent with underlying variability in the raw census data because of random rounding, suppression and weighting differences between short- and long-form census data.
5. The percentages always add up to 100 percent over the children belonging to a geographic or variable parent.
6. For CensusPlus 2016, the long-form data are adjusted to match the short-form census household and household population universes.
7. For the first time in 2016, income information was gathered by Statistics Canada from administrative data sources (such as tax data), which not only reduced response burden, but also increased the quality and quantity of income data available.

E. Themes

In general, the themes that are included in CensusPlus are consistent with the standard Statistics Canadas cumulative profile themes and variables. CensusPlus 2016 also includes four data sets that are not part of the standard profile release:

1. Family households by family structure
2. Family households by family size
3. Household population by age
4. Children at home by five-year age groups

Note that the families by family structure and families by family size both have census families as a universe in the 2016 Census. *Family Households* themes in CensusPlus have been transformed to a household universe. Both versions are included as themes in CensusPlus.

The majority of the themes in 2016 are the same as what was provided in the 2011 product, with the addition of the total and after-tax household income themes, which were not included in 2011 due to the poor data quality from the National Household Survey (NHS).

E1. New Themes for 2016

1. Household total income groups for private households
2. Household after-tax income groups for private households
3. Total children at home by five-year age groups. In 2011, there were fewer age groups and the groups were based on different aggregations.

E2. Excluded Themes for 2016

A number of themes that were part of the 2011 CensusPlus were excluded from the 2016 version of this product. These include the following:

1. Total, female and male population by single year of age. Only five-year age groups are included.
2. Religion. The religion question is asked by Statistics Canada every 10 years. This theme was included for 2011 and will be available again for the 2021 Census.

F. Methodology

The methodologies used to produce CensusPlus are more complex than one might guess. Our goal was to produce something that resembledand matched the census or at least the pattern of the census as much as possible, but solved all of the above-noted issues that the published census possessed. The following methodology describes the structure of our CensusPlus methodology.

Note that the data structure of the census data is in two important hierarchies:

1) The variable theme hierarchy
2) The geographic hierarchy

Figure 1 shows an example of a variable hierarchy. There are two sub-groups under the marital status theme: couples (married or living with a common-law partner) and non-couples (not married and not living with a common-law partner). Within the couples, there are two groups married (and not separated) and living common-law. Within non-couples there are four groups: single (never legally married), separated, divorced and widowed. There are three levels of this variable hierarchy and the children from each level should added up to the parent total reported at the next level in the hierarchy.

Figure 1 - Variable Hierarchy

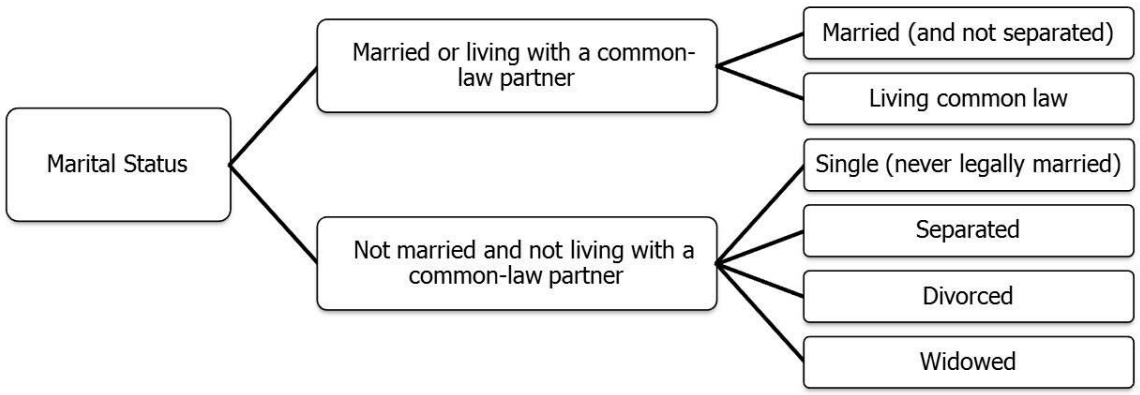
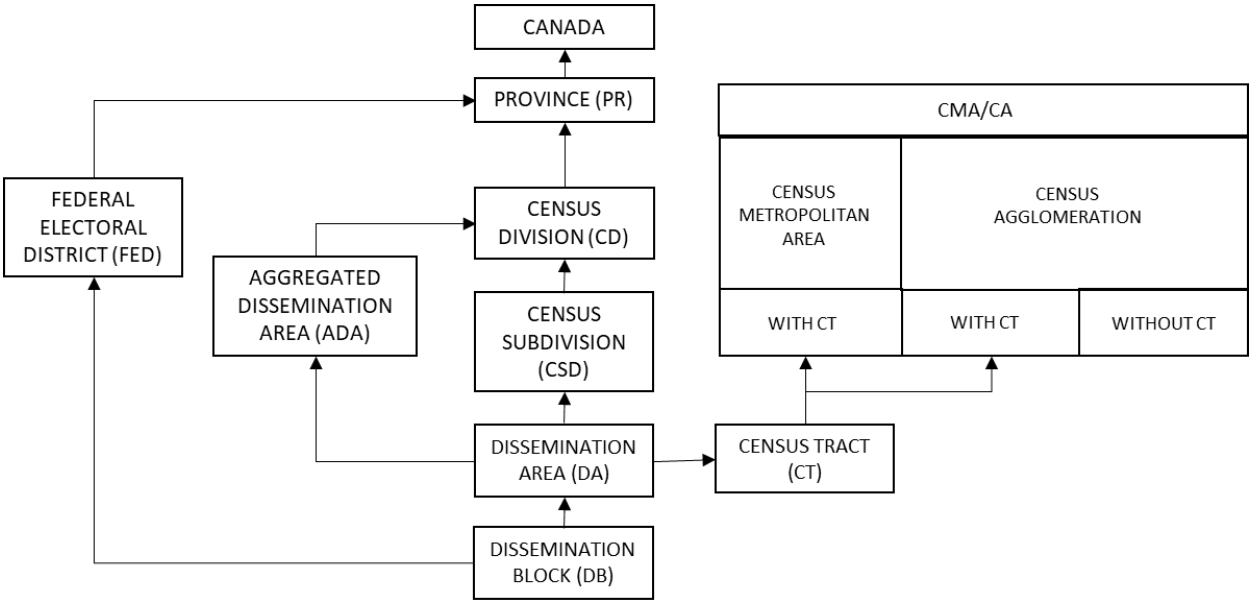


Figure 2 shows census geography hierarchy. Census geography has smaller areas nested within larger areas. For example, dissemination areas (DAs) nest within census subdivisions (CSDs), which nest within census divisions (CDs). Like the variable hierarchy, values for child geographies should sum up to the values reported by the parent geography. In 2016, Statistics Canada created a new census dissemination geographic area, called the aggregated dissemination area (ADA) to ensure the availability of census data, where possible, across all regions of Canada.

Figure 2 - Geographic Hierarchy



F1. General Rules that Guided Our Adjustments

1. **Preserve the most reliable numbers in the dataset.** For example, data reported for census divisions are more reliable than dissemination areas, or total population or household counts are more than detailed demographics.

2. **Use a top-down approach.** Start with provincial-level data and work down the geographic hierarchy to smaller levels of geography. Similarly, we start by processing the broadest variables and progressively process more detailed variables in the hierarchy.

3. **Prevent data loss**. Never set non-zero values to zero.

4. **Impute data only where necessary**. Replace zeroes with non-zero values only where needed to match geographic or variable control totals (see imputation methodology later in this document).

5. **Keep adjustments to a minimum**. Start with the raw data provided by Statistics Canada and make the minimal adjustment possible to the known values, in order to make the data conform to geographic and variable controls.

6. **Use all available information**. Use all information available to inform adjustments to the raw census data.

7. **Leverage geographic proximity to inform imputed values.** This leverages the core principle of quantitative geography that things that are near each other tend to be more alike than those that are far away.

F2. Row and Column Structure

For the purpose of the CensusPlus product, geographic areas are represented as records in the dataset and demographic attributes are the columns. The geographic and variable structures are critical to the CensusPlus dataset. All data must add up to the control totals based on child-parent geographic relationships. Similarly, all variables must add up via the defined variable hierarchy.

Geographic Hierarchies:

For geographic names, abbreviations and hierarchy, see

Figure 2.

The schema for the main geographic hierarchy is:

CAN, PR, CD, CSD, DA

The secondary geographic hierarchy schema is:

CAN, PR, CMA/CA, CT, DA

The tertiary geographic hierarchy schema is an alternative to the primary hierarchy above:

CAN, PR, CD, ADA, DA

Details regarding idiosyncrasies in the geographic hierarchy are described in Section H below.

Variable Hierarchies:

Each census theme or category has a distinct hierarchy to account for the specific collection of demographic attributes reported. Each theme has a different logic and a slightly different framework. Understanding, the hierarchy of variables is important for the processing and usage of CensusPlus. The ordering of the processing of each theme in the adjustment process is important in achieving the best results. For details on demographic attributes or variables see the variable list supplied with the CensusPlus product.

F3. Census Short-form Universe versus Long-form Universe

The census is split into two components: the short- and long-form census. The short-form census is a 100 percent sample of the Canadian population and includes all persons living in Canada, including populations living in collective or institutional dwellings. The short-form census captures basic demographic information about people and households: sex, age, family status, language spoken, dwelling structural type and, for the first time in 2016, income. The long-form 2016 Census is a sample of 25 percent of households. This sample excludes people living in collective and institutional populations. The long-form census captures detailed demographic attributes about cultural diversity, family structures and dwellings. The long-form census is reweighted to approximate the official population and household counts reported by the short form census.

The nature of weighting the long-form census to the short form is such that there can be discrepancies between the population and households counts reported between the two components of the census. Smaller geographic areas tend to deviate more when comparing values between the short- and long-forms. Significant differences between estimates occur at the census subdivision level and below. The short-form census data being 100 percent sample are more reliable than the long-form data. Therefore, we use population and households counts and other key universe values from the short-form census as our control totals for all census themes including long-form census. This has the added benefit of create an extra degree of comparability between short- and long-form themes.

For more info on the differences between the short- and long-forms from Statistics Canada, see the link provided.

http://www12.statcan.gc.ca/Census-recensement/2016/ref/98-304/chap9-eng.cfm

F4. Random Rounding, Data Suppression and Imputation

Random rounding is applied to every count in census tabulations. Some values are randomly rounded to zero or cell suppression. In other cases, entire records have all values suppressed to zero or records suppression. See the notes in the Appendix to better understand random rounding and data suppression. Our procedures attempt to identify false zero values in the dataset and provide an “impute” for those cases. Generating imputed values helps to support our objective of minimally disrupting the “good data values” reported in the census.

In cases of cell suppression, mathematical techniques leveraging the report values for sibling variables, sibling geographies, parent geography and children geographies are used to generate imputation values consistent with all known information. For record suppression, there is typically a lot less directly available information. We use a variety of techniques—including geographic nearest neighbours, statistical models, conditional-probability models and machine-learning algorithms—to generate initial impute values. In either cell or record suppression, we design our imputation methodology to conform to a few basic rules:

1) Imputation values will minimize disruption to good census values;
2) Imputation values will maximally use the informational content in the raw census data;

3) Imputation values will be consistent with the broader demographic context of the geographic area.

I. Notes on Geographic Idiosyncrasies

Census metropolitan areas are areas that include the larger cities of the country and the commutersheds around them. Some of these, like Ottawa-Gatineau, cover multiple provinces. Census agglomerations are simply smaller CMA type areas for cities like Kingston, ON. As illustrated in

Figure 2, CMAs and CAs are made up of CSDs and CSDs are made up of DAs. There are two types of CMAs and CAs, those that are "tracted"—meaning containing census tracts—and those that are "un-tracted". CMAs and CAs combined do not provide full coverage of Canada.

On average, census tracts (CTs) include seven DAs and are child geographies of CMAs and CAs. Like their parent geography, CTs do not cover the country. Generally, CTs are smaller than CSDs, but they do not nest within CSD geography. There are more than 100 CTs that cross CSD boundaries. For processing purposes, we create a custom geography called CT splits. CT splits create the opportunity to use CT like geographies as a child of CSD geography.

Aggregated dissemination areas (ADAs) are the most recent geography introduced with the 2016 Census. ADAs included on average 11 DAs, so slightly larger than CTs. The main benefit ADAs compared to CTs is that they provide full coverage of Canada. As a general rule, ADAs provide more detailed coverage in urban areas than CSDs and less detailed coverage rural areas. Similar to the CT, we create an ADA split geography so that we can process ADA-like geographies as children of the CSD.

For processing CensusPlus, we have most commonly used the geography hierarchy: PR – CD – CSD – ADA split – DA. Depending on circumstances, we deviated from this standard processing hierarchy and used: PR – CD – ADA – DA. In all cases, PR, CD, CSD, ADA, DA, CT and CMA/CA raw Census data are used to assess results of the CensusPlus process and make sure that CensusPlus values have not overly deviated from the official Census profile data.

Appendix

Random Rounding and Suppression

Random Rounding

For the census data, all counts are rounded to a base of five. This means that all counts will end in either zero or five. In cases where there are small counts of people or household attributes, (less than 10), they are randomly rounded to 10 or zero. Random rounding is applied independently to every value in the data set, meaning that there is no relationship between how one value in the Census is rounded to any other value. Further, randomly rounding, as the name suggests, means that values are not systematically rounded up or down based on standard rounding rules. With random rounding, 23 does not always round up to 25, and 52 does not always round down to 50. Depending on the value reported, there is a certain probability that the value will round up or down. For example, 23 will round up to 25 approximately 60 percent of the time and down to 20 approximately 40 percent of the time.

Suppression

The following three conditions will result in the suppression of statistics in the Canadian census released statistics:

1) Small counts – in any cases where there is a small number of individuals reported for a specific demographic attribute, values will either be randomly rounded to zero or actively suppressed to zero. In certain circumstances, where a geography has a very small number of people or households, the values for the entire records might be supressed to zero.
2) Data quality – where the sample for a particular geography is determined to be poor, all attributes for that geography will be supressed to zero.
3) Residual disclosure – in order to make it so that small counts cannot be directly reverse–engineered, complimentary data points will be suppressed to zero, even when there are sufficient population available in the cell.

For further technical details about census see http://www12.statcan.gc.ca/Census-recensement/2016/ref/98-304/98-304-x2016001-eng.pdf.