Data Quality and Confidentiality Standards and Guidelines (Public)

2011 Census Dissemination

Table of contents

Page

1. Introduction

Data disseminated by the census are subjected to a variety of automated and manual processes to determine whether the data needs to be suppressed. This is done primarily for two reasons: (1) to ensure non-disclosure of individual respondent identity and characteristics (which will subsequently be referred to as **confidentiality**) and (2) to limit the dissemination of data of unacceptable quality (which will subsequently be referred to as **data quality**).

Additionally, suppression of data may be applied for product specific reasons due, typically, to formatting issues. The term **product** refers, primarily, to tabular output. Data may be either modified in the product or removed from the product altogether to reflect the suppression rules required. This document summarizes the data quality and confidentiality standards and guidelines to be applied for the 2011 Census Dissemination Project.

The following new or updated rules have been adopted for 2011:

- Adjustment of population and dwelling counts at the federal electoral district level (see Section 3.1)
- Update to the rule for disseminating population aged 100 and older data (see Section 4.4)
- Adopting Statistics Canada standard table symbols in 2011 Census products (see Section 6.1).

2. Confidentiality (non-disclosure) rules

The following describes the various rules used to ensure confidentiality (or non-disclosure) of individual respondent identity and characteristics. All census data are subject to confidentiality (non-disclosure) rules.

2.1 Area suppression for standard[1] and non-standard geographic areas

Area suppression is used to remove all characteristic data for geographic areas below a specified population size.

The specified population size for all standard areas[1] or aggregations of standard areas is 40, except for blocks, block-faces or postal codes. Consequently, no characteristics or tabulated data are to be released if the total population of the area is less than 40.

The specified population size for six-character postal codes (forward sortation area – local distribution unit [FSA-LDU]), geocoded areas and custom areas built from the block, block-face or LDU levels is 100. Consequently, no characteristics or tabulated data are to be released if the

1. Refer to the Census Dictionary for more information on standard areas.

total population of the area is less than 100. Generally, blocks and individual urban block-faces (one side of the street between two intersections) will be too small to meet the above-specified population size thresholds. Where an aggregation of blocks or block-faces fall above the threshold specified by the population size, data can be retrieved through a custom tabulation.

These specified population size thresholds are applied to 2011 Census data as well as all previous census data.

Please refer to section 2.2 Postal code minimum aggregation rules for additional rules applicable to postal code data.

2.2 Postal code minimum aggregation rules

In addition to the confidentiality rules on disseminating Census data with the postal codes, the following rules are applied to postal codes. These rules fall under clause 03.01 (n) of the Commercial Non-Mailing licence between Statistics Canada and Canada Post Corporation.

- All requests must include batches of two or more postal codes; the only exception being for postal codes which have a zero as the second digit (rural postal codes).
- Groups of postal codes are to be assigned a unique classification/number (e.g. K1A 0T6, 0T7, 0T8 = Custom Area 1); under the terms of the contract listed above, clients cannot be provided with lists of postal codes, only the name specified in the client's request can be used.
- All other confidentiality rules for custom extractions still apply as per Section 2.1.

Also, the following disclaimer is applicable to all postal code custom requests:

> Postal code validation disclaimer: Statistics Canada makes no representation or warranty as to, or validation of the accuracy of any postal code^{OM} data submitted to Statistics Canada.

Please note these rules are applicable to historical postal code requests as well.

2.3 Random rounding

All counts in census tabulations are subjected to random rounding. Random rounding transforms all raw counts to random rounded counts. This reduces the possibility of identifying individuals within the tabulations.

All counts are rounded to a base of 5, meaning they will end in either 0 or 5. The random rounding algorithm employed controls the results and rounds the unit value of the count according

to a predetermined frequency. Table 2.1 below shows those frequencies. Note that counts ending in 0 or 5 are not changed and remain as 0 or 5.

Table 2.1 Random rounding frequency

Unit values of	Will round to count ending in 0	Will round to count ending in 5
1	4 times out of 5	1 time out of 5
2	3 times out of 5	2 times out of 5
3	2 times out of 5	3 times out of 5
4	1 time out of 5	4 times out of 5
5	Never	Always
6	1 time out of 5	4 times out of 5
7	2 times out of 5	3 times out of 5
8	3 times out of 5	2 times out of 5
9	4 times out of 5	1 time out of 5
0	Always	Never

The random rounding algorithm uses a random seed value to initiate the rounding pattern for tables. In these routines, the method used to seed the pattern can result in the same count in the same table being rounded up in one execution and rounded down in the next.

2.4 Disclosure avoidance for statistics

Statistics (such as mean, standard error, sum, median, percentile, ratio or percentage) are not subject to random rounding. However, when shown in tabulations accompanying the counts used to calculate the statistic, their presence can result in disclosure of individuals. To prevent this, we use statistic suppression methods, or special statistic calculations.

2.4.1 Statistic suppression

For all quantitative variables, a statistic is suppressed if the number of actual records used in the calculation is less than 4.

2.4.2 Special statistic calculations

(1) The statistic value is never rounded, except for frequencies.

(2) All statistics based on ranks (medians, percentiles) are calculated the usual way and they are never rounded. We never release the minimum or the maximum of a statistic.

(3) When a sum is specified for age, then the program multiplies the unrounded average of the group in question by the rounded frequency. Otherwise, if a sum is specified for a variable other than age, the program rounds the actual sum.

When a division is specified (averages, percentages, ratios, etc.), the program must apply point (3) to both numerator and denominator before it proceeds with the division.

Note: Statistics based on ranks like median and percentiles are always calculated via linear interpolations. That means that, for low count cells, these statistics are not reliable. That is the reason no additional confidentiality measures are applied to them.

Note: The average of an age is not altered by the rounding, because the numerator is the product of the true average by the rounded frequencies and the denominator is the rounded frequencies. The two frequencies cancel each other, leaving the true average untouched.

3. Confidentiality practices

3.1 Confidentiality adjustment for population and dwelling counts

The population counts of small dissemination blocks with low population counts may be adjusted to reinforce the confidential nature of the data. In fact, all dissemination block population counts less than 15 will be rounded to a base of 5. This adjustment, however, will be controlled. That is, aggregates (totals) of the adjusted population counts for dissemination areas (DA) will always be within 5 of the actual values. The control will be even tighter at the census subdivision (CSD) and federal electoral district level (FED). In fact, while always being within 5 of the actual values, the adjusted population counts and the actual values agree for a maximum number of census subdivisions and federal electoral districts. Finally, all census division adjusted population counts and actual values agree, which means that the population counts for all census divisions (CD) remain unchanged.

3.2 Preventing disclosure

Prevention of direct or residual disclosure must also be addressed when determining product content. When assessing the potential for disclosure, a number of factors must be considered. The detail of individual variables, cross-classification of variables and the geographic level of the data will all contribute to the risk. For example, there may be no risk in producing households by number of persons in the dwelling and detailed groupings of age showing various characteristics

of the household members for large geographic areas. However, the risk of disclosure would increase for the lower levels of geography.

The most common method used for preventing disclosure is defining content that is appropriate for a given geographic level. Increasing population thresholds or applying manual suppression, as needed, are other methods that can be employed. Since these are typically product-specific requirements, they are not part of the automated suppression systems.

3.3 Census of Agriculture tabulations

Census of Agriculture and Census of Population data are matched using geographic information, name, age and sex of farm operators. Imputation is performed to account for non-matches. Data are available for all members of households where a farm operator resides.

Census of Agriculture data include farm type, farm sales, area of crops and numbers of livestock, while the Census of Population provides data including age, sex, family structure, marital status and language of family members and household members. Both pre-planned standard products and custom tabulations are produced at the province level only. Multidimensional tables are common and usually include two or three dimensions.

Confidentiality measures include random rounding of 'counts' (e.g., number of operators, number of census families) and in some specific cases, low-bound suppression. Averages are provided unaltered. All verification of tabulations is done internally by Census of Agriculture staff before release.

For a summary of 2006 and previous censuses confidentiality rules, please refer to the 2006 Data Quality and Confidentiality Standards and Guidelines (Public).

4. Data quality practices

The following section describes the methods used to restrict the dissemination of census data of unacceptable quality.

4.1 Data quality measures

4.1.1 Data quality indicators

Data quality indicators (commonly referred to as data quality flags) are attached to each standard geographic area disseminated. In the census database environment, the data quality indicators consist of a five-digit numeric field. In electronic products browsed via Beyond 20/20, these flags are displayed as a five-digit numeric code (example: 0 2 1 0 0). On the census website, flagging to end users of partially enumerated areas is done through the use of symbols.

4.1.1.1 Incompletely enumerated Indian reserves and Indian settlements

In the 2011 and previous censuses, enumeration was not completed on some Indian reserves and settlements, due to non-participation. Data quality rules require these incompletely enumerated Indian reserves and settlements be identified and removed from products. As well, higher-level geographic areas containing these areas must be identified in the products. These higher-level indicators are automatically included in output tabulations in Beyond 20/20, CSV and 'flat file' formats. For a list of 2011 incompletely enumerated Indian reserves and settlements, users can go to the reference materials section of the census website.

Although census data are not available for incompletely enumerated Indian reserves and settlements, the areas themselves are included as part of the standard geographic hierarchies on the census database.

4.1.1.2 Global response rates

Global response rates are determined for each of the census geographic areas. These areas are flagged on the database according to the non-response rate. Geographic areas with a non-response rate higher than or equal to 25% are suppressed from tabulations. Geographic areas with a global non-response rate higher than or equal to 5% and lower than 25% are broken into two categories and are flagged according to the following ranges: falling between 5% and 10% and falling between 10% and 25%. These geographic areas are identified in tabulations, but not suppressed. In electronic products, a numeric flag is provided with the area identifier indicating low data quality.

4.1.1.3 Population and dwelling counts error flag

After the release of the population and dwelling counts, errors are occasionally uncovered in the data. It is not possible to make changes to the census data presented. Users can, however, obtain the population and dwelling count amendments, listed by census subdivisions and other levels of geography, by visiting the 2011, 2006 or 2001 census portion of the Statistics Canada website at www.statcan.gc.ca.

4.1.1.4 Not applicable data quality indicator (20% sample data quality flag)

The fourth numeric code of the five-digit data quality indicator on the database is not applicable for the 2011 Census, and it is automatically set to zero for each geographic area. The value that resides on the database is a place holder for historical reasons; in 2006 and previous censuses, all five digits were applicable, the fourth digit was the 20% sample data quality flag.

4.1.1.5 2006 adjusted population flag

Users wishing to compare 2011 Census data with those of other censuses should take into account that the boundaries of geographic areas may change from one census to another. In order to facilitate comparison, the 2006 Census counts are adjusted as needed to take into account boundary changes between the 2006 and 2011 censuses. The flag is also used to refer to corrections to the 2006 counts and to identify areas that have been created since 2006, such as newly incorporated municipalities (census subdivisions) and new designated places. However, most of these flags are the result of boundary changes.

Table 4.1 below describes the data quality indicator field and its contents. Note that a zero in any of the five digits is the default for the respective indicator and means that no data quality action is required.

Table 4.1 Data quality indicators – 2011 Census

Digit	Description	Flag	Flag description
1st (0XXXX)	Incomplete enumeration flag	0	Default.
		1	Incompletely enumerated Indian reserve or Indian settlement (suppressed).
		2	Excludes census data for one or more incompletely enumerated Indian reserves or Indian settlements.
2nd (X0XXX)	Data quality flag	0	Default.
		1	Data quality index showing, for the short census questionnaire (100% data), a global non-response rate higher than or equal to 5% but lower than 10%.
		2	Data quality index showing, for the short census questionnaire (100% data), a global non-response rate higher than or equal to 10% but lower than 25%.
		3	Data quality index showing, for the short census questionnaire (100% data), a global non-response rate higher than or equal to 25% (suppressed).
3rd (XX0XX)	Population and dwelling counts error flag	0	Default.
		1	An error exists in the 2011 population and dwelling counts for this area. For further details, please refer to the population and dwelling counts data section of the 'Notes' file.
		2	In 2006, the population and/or dwelling counts for this census subdivision were found to be incorrect. Since it is not possible to make changes to the 2006 Census data presented in these tables, the 2006 data should be used with caution. For further details, please refer to the population and dwelling counts data section of the 'Notes' file.
		3	Both the 2011 and 2006 population and/or dwelling counts for this area were found to be incorrect. Since it is not possible to make changes to the census data presented in these tables, these counts should be used with caution. For further details, please refer to the population and dwelling counts data section of the 'Notes' file.

Table 4.1 Data quality indicators – 2011 Census (continued)

Digit	Description	Flag	Flag description
4th (XXX0X)	Not applicable	0	Default.
5th (XXX0)	2006 adjusted population flag	0	Default.
		1	2006 adjusted count; most of these are the result of boundary changes.

Note: The data quality flag does not apply to the population and dwelling counts.

Please refer to the 2006 Data Quality and Confidentiality Guidelines document (Public) for the flag legends for historical census years.

4.2 Other methods of data quality suppression

The methods of suppression mentioned to this point provide sufficient data quality suppression and identification for most census data products. However, in some products, the specifying area or production area may require that additional data quality suppression be performed. Examples of additional suppression could include increasing population thresholds or applying distribution or cell suppression. These are typically product-specific requirements and therefore are not part of the automated suppression systems. In all cases, some form of manual process is required.

4.3 Calculation of order statistics

For variables which have integer values, a median (or other quantile) is calculated using linear interpolations to give the variable a decimal, even if the variable is an integer. This is done to provide a sense of the relative position of the median record among those records that have the same value (the median). Therefore, a value of 23.46 means that the record in the middle (the median) has Age = 23 and 46% of all records with Age = 23 lie to the left of the middle. The following set {23, 23, 23, 23, **23**, 23, 23, 23, 23} (the median is bolded) yields 23.11 as the calculated/reported median, for example.

4.4 Data quality rule for disseminating data for population aged 100 and older

Data for the population aged 100 years and older cannot be disseminated in single years of age. For custom requests that require a more detailed breakdown than provided in standard data products, in which the population aged 100 years and older is grouped together, the most detailed age breakdown which can be provided is as follows, and it can only be provided for 'Canada':

Total population 100 years and older
100 years to 104 years
105 years to 109 years
110 years and older

4.5 Data quality rule for disseminating data on same-sex and opposite-sex couples

The questionnaires of the 2011 Census of Population and the 2011 National Household Survey introduced for the first time a specific response on household relationships to determine the number of same-sex married couples. Analysis of the data on same-sex married couples has shown that there may be an overestimation of this family type and marital status. The 2011 Census shows a total of 64,575 same-sex couples in Canada, of which 21,015 are married couples. The range of overestimation of both these counts, at the national level, is between 0 and 4,500.

For levels of geography such as Canada, provinces, territories and census metropolitan areas (CMAs), counts are generally higher, so the potential overestimation is expected to be small in relative terms; however, the data should still be interpreted with caution.

At lower levels of geography, the same potential overestimation could be relatively large, and not only should the data be interpreted with caution, but certain suppression rules restrict their publication. These rules apply to both the 2011 Census and the 2011 National Household Survey.

First, the breakdown of same-sex couples or opposite-sex couples by conjugal status, that is, whether they are married or living common law, cannot be disseminated for geographic areas other than Canada, provinces, territories and CMAs.

Second, data cannot be disseminated that identify either same-sex or opposite-sex couples (in total, married or living common law) of any area with a population of less than 5,000 (as measured in the 2011 Census for private households).

In summary,

- All data may be disseminated for same-sex or opposite-sex couples for Canada, provinces, territories, census metropolitan areas (CMAs), although they should still be interpreted with caution.

- Data on same-sex couples and opposite-sex couples may be disseminated for other geographic areas if they have a population of 5,000 or more, provided that the breakdown by conjugal status (married, living common law) is not included.

- No data may be disseminated that identify any same-sex or opposite-sex couples for areas of population less than 5,000.

5. Data suppression – Other

Suppression of data may be applied for product-specific reasons due, typically, to the size of the product and/or the constraints of the media on which the product is being disseminated.

5.1 Incidence reporting

Incidence reporting is a process used to order or rank characteristic data by size within products. It can be used as a method to select only the 'n' highest categories of a characteristic for inclusion in a product.

5.2 Zero suppress

Zero suppress refers to the removal of records in which all of the counts are equal to zero. This method is used to reduce the size of an output product by removing any rows of the output matrix where all data are equal to zero.

6. Best practices

6.1 Data quality and confidentiality table symbols

New for 2011, census standard and custom products will contain three Statistics Canada standard table symbols. The table below shows each new symbol being adopted for 2011 and its description, as well as the symbol that was used in 2006 Census products (HTML and B20/20 format).

New symbol	Description	Symbol used in 2006 census products (HTML)	Symbol used in 2006 census products (B20/20)
..	not available for a specific reference period	N	- (dash)
...	not applicable	N	- (dash)
x	suppressed to meet the confidentiality requirements of the *Statistics Act*	0 (zero)	0 (zero)

It is important to note that these symbols are being adopted for 2011 Census products only. All historical products (2006 Census and earlier) will not contain any of these new symbols nor will any historical data in 2011 Census products.