Data Quality and Confidentiality Standards and Guidelines (Public)

Modified on May 1, 2008

2006 Census Dissemination

Statistics Canada Statistique Canada

Canada

Table of contents

Page

1. Introduction

1.1. Background

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 2006 Census Dissemination Project.

1.2. Executive summary

The data quality and confidentiality standards were developed for application in the 2006 Census Dissemination Project. The summary below includes revisions made to the rules and practices that were carried over from 2001, along with new rules and practices adopted for 2006.

Revisions to 2001 rules and guidelines

- An explanation of blocks and individual urban block-faces for area suppression for standard and non-standard geographic areas (Section 2.1);
- an expanded definition of the population universes used for suppression routines (Section 2.1.2);
- an additional note explaining special statistic calculations as it refers to dollar value, number of weeks, number of hours and age (Section 2.3.2);
- a tighter control on confidentiality adjustments for population and dwelling counts at the census subdivision (CSD) level. 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 (Section 3.1);
- modification to the rule for 2D suppression based on the 2B and 2D population sizes (Section 5.1);
- increased background information for public use microdata files (Section 3.6);
- a clarification on area suppression for income characteristic data (Section 4.1); and
- removal of the 2001 Census 5,000 population threshold for same-sex common-law couples.

New for 2006 are the following

- Implementation of outlier statistics suppression (Section 2.3.3);
- inclusion of a subsection on confidentiality adjustment for daytime population counts (Section 3.3);
- inclusion of a subsection on calculation of order statistics (Section 4.4); and

- improved application of the maximum/minimum/mean statistic suppression (for 2006 tables only) (Section 2.3).

1.3. Census release criteria

The following questions were given consideration when reviewing and assessing the current release criteria and established policy for census data tabulations:

- Has the release criteria been relaxed over the last three or four census cycles? And, if so, why?
- What additional data outside the agency, if any, might give cause for concern with respect to current population release thresholds?
- What new data mining tools or software being used today might impact on census data release thresholds?
- Are our release criteria for tabular data more or less stringent than other statistical agencies?
- Does the census random rounding policy still work as an effective disclosure avoidance technique, and should we be looking at others?

The current release criteria and established policy for census weighted tables have proven to be successful at ensuring non-disclosure. Evolving client demands, data mining, the ever-increasing data combination possibilities and software developments give reason to perform regular assessments, research new methods, and make changes where necessary.

The census release criteria have been modified slightly over time since the 1981 Census, adapted to meet changing needs and conditions. In general, although some population thresholds for geographic area suppression were adjusted downwards from 1981 to 1986 and remained constant since then, there was no evidence of any negative impact resulting from this change. In other ways, the criteria have actually been strengthened, such as the additional protection added to the random rounding rules, the addition of the higher suppression threshold for postal code information, and suppression of certain statistics (e.g., standard deviation = 0). The population universes upon which the population thresholds are based were also modified in 1991 and 1996 to add protection against disclosure. In addition, a new suppression method has been implemented in 2001 that will prevent the dissemination of tables in which the number of units (individuals, families or households) is below a given threshold. The suppression rules and random rounding practices specific to the dissemination of 2006 data are provided in Section 2 of this document. A summary table of historical confidentiality rules, 1981 to 2001 can be found in Appendix A at the end of this document.

2. Confidentiality (non-disclosure) rules

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

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

The specified population size for all standard areas 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 for areas below a population size of 40.

The specified population size for six-character (FSA-LDU) postal codes, 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 for these area types below a population size of 100. Generally, blocks and individual urban block-faces (one side of the street between two intersections) will be too small to meet the above threshold specified population sizes. 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. Additional area suppression is applied for data quality reasons if the census tabulation contains any data showing income characteristics for individuals, families or households (see Section 4.1 for further information).

These threshold specified population sizes are applied to 2006 Census data as well as all previous census data.

2.1.1. Population universes used for suppression routines

The population under consideration for all 100% data tabulations is the total population.

For all other tabulations, except place of work data, the population under consideration is the lower of the 2A (100% data) or 2B (20% sample data) non-institutional population.

For place of work data, the population under consideration is the employed labour force having a usual place of work or worked at home.

Population universes used for suppression routines

2A	2B	Place of work geographic areas
Total population	Lower of the 2A or 2B non-institutional population	Employed labour force having a usual place of work or worked at home

For census tabulations that are based on place of work geographies or areas, all criteria are to be based on the employed labour force having a usual place of work or worked at home counts. That is, the 40 population, 100 population and 250 population thresholds are employed labour force having a usual place of work or worked at home counts, rather than the population of the areas. Tabulations containing both places of residence and places of work as geographic areas have the 40, 100 and 250 size limits applied to both place of residence (population) and place of work (employed labour force having a usual place of work or worked at home).

2.2. Random rounding

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

For 2A (100%) data, all counts are rounded to a base of 5. This means that all 2A counts 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 pre-determined frequency. The table below shows those frequencies. Note that counts ending in 0 or 5 are not changed and remain as 0 or 5.

Random rounding frequency (100% data)

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

2B (20%) data require a slightly different random rounding algorithm. All counts greater than 10 are rounded to base 5, as is done for 2A data. Counts less than 10 are rounded to base 10. This means that any 2B counts less than 10 will always be changed to 0 or 10. The table below shows the effect of rounding on 2B counts with a value less than 10.

Random rounding frequency (20% sample data)

Count of	Will round to 0	Will round to 10
1	9 times out of 10	1 time out of 10
2	8 times out of 10	2 times out of 10
3	7 times out of 10	3 times out of 10
4	6 times out of 10	4 times out of 10
5	5 times out of 10	5 times out of 10
6	4 times out of 10	6 times out of 10
7	3 times out of 10	7 times out of 10
8	2 times out of 10	8 times out of 10
9	1 time out of 10	9 times out of 10
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.

2.3. 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.

2.3.1. Statistic suppression

The following three situations will result in the suppression of statistics:

(1) It is possible (mainly for cells with small counts) that quantitative values were imputed from a single donor record. For example, an income cell with three individual records may in fact be only one actual response to income and the other two income amounts were imputed from the first record. When this occurs, the income characteristics of a single individual could be

disclosed if the mean and standard error statistics are produced. To prevent this, and more generally to prevent disseminating statistics based on a narrow range of values, all statistics of a cell are suppressed if the relative difference between the minimum and the maximum is less than a specific percentage.

(2) For all quantitative variables, a statistic is suppressed if the number of actual records used in the calculation (not rounded or weighted) is less than a specific number.

Note: The number of records used in the calculation is not necessarily the number of records in the cell but, rather, the number of records that are applicable or available to the calculation of the statistic in the cell.

(3) For all quantitative variables, all statistics are suppressed if the sum of the weights is less than 10.

2.3.2. 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, that is, never rounded.

(3) All dispersion statistics (standard error) are calculated the usual way, that is, never rounded.

(4) When a sum is specified, if the program sums a dollar value, a number of weeks, a number of hours, or an age, then the program multiplies the unrounded average of the group in question by the rounded, weighted frequency. Otherwise, the program rounds the actual weighted sum.

When a division is specified (averages, percentages, ratios, etc.), the program must apply the point (4) 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 why no additional confidentiality measures are applied to them.

Note: The average of dollar value, a number of weeks, a number of hours or 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.

2.3.3. Outlier statistic suppression

It is possible, though highly unlikely, that an outlier can be estimated accurately on the basis of an average. To reduce the risk of such a disclosure, all statistics for a cell will be suppressed if the ratio of the absolute value to the sum of the absolute values is greater than a specific percentage.

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) level. 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. Finally, all census division adjusted population counts and actual values agree, which means that the population counts for all census divisions (CD) remain unchanged.

3.1.1. Confidentiality adjustment for Forward Sortation Area (FSA) population and dwelling counts

The population counts for FSAs less than 15 will be rounded to a count ending in '0' or '5' to reinforce the confidential nature of the data. This adjustment will be controlled at the province level.

3.2. Confidentiality adjustment for place of work counts

The place of work counts for census blocks are available on a custom basis. These counts will be adjusted to reinforce the confidential nature of the data. In fact, all census block counts for employed labour force having a usual place of work or worked at home will be rounded to a base of 5. This adjustment, however, will be controlled. That is, aggregates of the adjusted population counts for dissemination areas will always be within 5 of the actual values.

3.3. Confidentiality adjustment for daytime population counts

Daytime population counts will be determined by taking the population living in a specific area, adding in the workers who live elsewhere and commute into the area, and subtracting the workers who live in the area and commute out of the area. The number of workers will be based on persons in the employed labour force having a usual place of work or worked at home. Daytime population counts will be adjusted to reinforce the confidential nature of the data by controlled rounding of the counts to a base of 5.

3.4. 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 rooms in the dwelling and detailed groupings of dwelling value 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.

3.5. Census of Agriculture tabulations

Census of Agriculture and Census of Population 2B (long form) data are matched using geographic information and the age and sex of farm operators. Match rates are about 95% and weighting 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 socioeconomic data, including education, income and occupation of families and household members. Pre-planned standard products are produced at the province level only. Some multidimensional tables are available in pre-planned products with the usual being two or three dimensions.

Custom products are available at subprovincial levels based on aggregations of weighted areas, where 5,000 persons is the usual population threshold, although it can go as low as 2,500. The data are random-rounded and low-bounded to ensure confidentiality. Suppressions are done manually if cells are below a specified size.

All verification of tabulations is done internally by Census of Agriculture staff. There is a group responsible for reviewing all agriculture-population data before release.

3.6. Public use microdata files (PUMF)

The 2001 PUMF product consisted of a series of files based on different focuses: individuals, families and households. Each file contains approximately 3% of the Canadian population. The recommendation for the 2006 PUMF product is an individual file based on the individuals, families and households focuses.

Microdata files are unique among census products in that they give users access to non-aggregated data. This makes PUMF a powerful research tool. The files contain a large number of variables. Users can group and manipulate these variables to suit their own data and research requirements. Tabulations not included in other census products can be created, or relationships between variables can be analyzed using various analytical tools.

The census public use microdata files (PUMF) provide quick access to a comprehensive social and economic database about Canada and its people. They consist of samples of anonymous responses to the census long questionnaire. The PUMF files contain statistical information about Canadians, the families and households to which they belong and the dwellings in which they live.

Statistics Canada has to protect the confidential information that it collects. Owing to the very nature of a microdata file, various measures are taken to fulfil this commitment. The Microdata Release Committee reviews all requests for release of microdata.

Data for small geographic areas are not available in these files. The user will find information only for selected census metropolitan areas, the provinces and the territories. The breakdown of some sensitive variables was reduced for the Atlantic region. Some of the values of sensitive variables were suppressed because their combination could have been used to identify a person, a family or a household. Also, in 2001, the income variables were subjected to reduced low, and high, income limits. This will require review for 2006.

4. Data quality practices

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

4.1. Area suppression for income characteristic data

Area suppression, when applied for data quality purposes, is used to replace all income characteristic data with zeroes for geographic areas with populations and/or number of households below a specific threshold.

If a census tabulation contains any data showing income characteristics for individuals, families or households, then the following rule applies. Income characteristic data are zeroed out for areas where the population is less than 250 or where the number of private households is less than 40. These thresholds are applied to 2006 Census data as well as all previous census data. The threshold of 40 private households is based upon the fact that weighted data are being used. With the weighting factor for each household being 5, setting a threshold of 40 ensures that there will be at least 8 households used in the calculation. The private household threshold does not apply for tabulations based on place of work geographies.

4.2. Data quality measures

4.2.1. Data quality indicators for tabulations based on place of residence geographies

Data quality indicators (commonly referred to as data quality flags) are attached to each standard geographic area disseminated. In the census database environments, the data quality indicators consist of a five-digit numeric field. On the database and in electronic products browsed via Beyond 20/20, these flags are displayed as a five-digit numeric code (example: 0 2 1 3 1). On the census website and in print publications, flagging to end users partially enumerated areas is done through the use of symbols. Specific symbols in use for the 2006 Census are documented as part of the print publication standards.

4.2.1.1. Incompletely enumerated areas

In the 2006 and previous censuses, some dissemination areas for Indian reserves were not enumerated due to non-participation/non-cooperation. Data quality rules require these non-enumerated areas to be identified and removed from products. As well, higher-level geographic areas containing non-enumerated areas must be identified in the products.

Although there is no census data collected for non-response areas, the areas themselves are included as part of the standard geographic hierarchies on the census databases. Retrieval and tabulation software will retrieve these areas but with no data.

4.2.1.2. Partially enumerated areas

Any geographic area that contains an incompletely enumerated area is considered a partially enumerated area. Partially enumerated areas are flagged to end users as containing incompletely enumerated areas.

4.2.1.3. Global response rates (100% data quality flag and 20% sample data quality flag)

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.

4.2.1.4. 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 2006 or the 2001 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 2006 or the 2001 Census portion of the Statistics Canada website at www.statcan.ca.

4.2.1.5. 2001 adjusted population flag

Users wishing to compare 2006 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 2001 counts are adjusted as needed to take into account boundary changes between the 2001 and 2006 censuses. The flag is also used to refer to corrections to the 2001 counts and to identify areas that have been created since 2001, such as newly incorporated municipalities (census subdivisions) and new designated places. However, most of these flags are the result of boundary changes.

The following table 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.

Data quality indicators for place of residence – 2006 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)	100% 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 2006 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 2001 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 2001 Census data presented in these tables, the 2001 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 2006 and 2001 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.

Data quality indicators for place of residence – 2006 Census (continued)

Digit	Description	Flag	Flag description
4th (XXX0X)	20% sample data quality flag	0	Default.
		1	Data quality indexes showing, for the long census questionnaire (20% sample data), a global non-response rate higher than or equal to 5% but lower than 10%.
		2	Data quality index showing, for the long census questionnaire (20% sample data), a global non-response rate higher than or equal to 10% but lower than 25%.
		3	Data quality index showing, for the long census questionnaire (20% sample data), a global non-response rate higher than or equal to 25% (suppressed).
5th (XXXX0)	2001 adjusted population flag	0	Default.
		1	2001 adjusted count; most of these are the result of boundary changes.

Note: The 100% and 20% sample data quality flags do not apply to the population and dwelling counts. The flag legend for historical census years can be found in Appendix B at the end of this document.

4.2.2. Data quality indicators for tabulations based on place of work geographies

Place of work areas are suppressed for data quality reasons when the following three conditions are met:

1. the area is not already suppressed due to confidentiality reasons
2. the equivalent residence area is suppressed for data quality reasons
3. most of the workers in the area are also residents in the area.

The data quality indicator for place of work uses only the 4th digit of the five-digit numeric code. A value of 3 on this indicator for a place of work geography indicates the area is to be suppressed.

Data quality indicator for place of work – 2006 Census

Digit	Description	Flag	Flag description
4th (XXX0X)	**20% sample data quality flag**	0	Default.
		3	- Data quality index showing, for the long census questionnaire (20% sample data); - a global non-response rate higher than or equal to 25% (suppressed); and - most of the employed labour force working in the area, also reside in the same area.

4.3. 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.

4.3.1. Distribution suppression

The most common example of other methods of data quality suppression is distribution suppression. This occurs in selected standard income products where income distributions are suppressed when the total number of units (persons, families, households) within the income distribution is less than 250. A variation of this procedure is applied to standard income products that feature number and average employment or total income only.

Further, when there are indications that there is high degree of variability among responses and thus the possibility of extreme income values, the earnings and/or income statistics may also be suppressed for data quality purposes. Therefore, more specific rules are in place that account not only for population size but also for the likelihood of uncertainty in the estimates due to extreme values and sample variability.

4.4. Calculation of order statistics

Medians and more generally quantiles are calculated using linear interpolations. The quantile interval (that is the interval where the value of the quantile is located) is determined using two methods based on the kind of values of the statistical variables:

(1) Variables that take values with decimals and any variables with dollar values.

The quantile interval is constructed to ensure that relative errors made by using the linear Interpolation are less than 0.78%. For example, if the true quantile is $30,000, the error made of using the built-in algorithm is less than $234.

(2) Variables that take integer values that are not dollars.

For these variables, the quantile interval is always of size 1. For example, if the true quantile is 23.46, the interpolation is applied to the interval [23, 24].

5. Data suppression – Indian reserves

5.1. Indian reserves – 2D suppression

Suppression of data also occurs when certain census questions are not asked of all respondents. Persons living on Indian reserves and Indian settlements who were enumerated with the 2006 Census 2D questionnaire were not asked the questions on citizenship (Question 10), landed immigrant status (Question 11) and year of immigration (Question 12). However, it was possible that a census subdivision (CSD) or lower geographic area was enumerated using both the 2D questionnaire (for the on-reserve population) and the 2B questionnaire (for the off-reserve population). In this case, the following rules were used to determine if suppression had to be applied to all citizenship and immigration data for that CSD (or lower geographic area):

1. If the population count from 2B questionnaires is higher than the population count from 2D questionnaires (based on weighted results), then include citizenship and immigration counts.

2. If the population count from 2D questionnaires is higher than the population count from 2B questionnaires (based on weighted results), then exclude citizenship and immigration counts.

3. In cases where the weighted results are the same, apply rules 1 and 2 to the unweighted counts to make the determination.

4. In cases where the unweighted results are the same, exclude the citizenship and immigration counts for the CSD (or lower geographic area).

Consequently, citizenship, landed immigrant status and period of immigration data are suppressed for Indian reserves and Indian settlements at census subdivision and lower levels of geography where the majority of the population was enumerated with the 2D form. These data are, however, included in the totals for larger geographic areas, such as census divisions and provinces.

In 2006, zeros were used to show where data were suppressed. In 2001, data were suppressed either by the removal of the geographic area from the table (e.g., in the 2001 topic-based tabulations) or by showing the data as 'null' (e.g., profile products displayed the 'null' data value for the area either as a '–', 'N' or as a 'D,' with the 'D' referring users to a special note).

For a complete list of Indian reserves and Indian settlements for which citizenship, landed immigrant status and period of immigration data are suppressed in 2006 and 2001, please refer to:

http://www12.statcan.ca/english/census06/reference/notes/supplist2D.cfm

.

6. Data suppression – Other

As indicated in Section 5, suppression of data occurs when certain census questions are not asked of all respondents. Additionally, 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.

6.1. 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.

6.2. 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.

6.3. Place of work flow suppression

Flow suppression is a process used to remove records with very low counts from place of work flow tables. It can be used as a method to select only those commuting flows with a count greater than a threshold value for inclusion in a product. The default value is 20. These are typically product-specific requirements and therefore are not part of the automated suppression systems.

Appendix A Summary table of confidentiality rules (historical), 1981 to 2001

Census years ▸ Confidentiality rules ▾	1981	1986	1991	1996 to 2001
Random rounding	Rounding up or down to base 5.	Rounding up or down to base 5.	For 100%, rounding to base 5. For 20%, rounding to base 10 for weighted frequencies less than 10. Weighted frequencies over 10, rounding to base 5, (custom tabulations only).	For 100%, rounding to base 5. For 20%, rounding to base 10 for weighted frequencies less than 10. Weighted frequencies over 10, rounding to base 5, (all tabulations).
Area suppression	Geographic areas < 50 for self-enumeration areas were suppressed. Geographic areas < 25 for canvasser areas were suppressed. For products containing income characteristics, distributions or quantitative statistics, the threshold was 250.	Standard geographic areas with population count < 40 were suppressed. Non-standard geographic areas with population count < 100 were suppressed. For products containing income characteristics, distributions or quantitative statistics, the threshold was 250.	All standard and non-standard geographic areas with population count < 40 were suppressed. For 100% data, the decision was based on population count. For 20% data, the decision was based on non-institutional population estimates. For products containing income characteristics, distributions or quantitative statistics, the threshold was 250.	All standard and non-standard geographic areas with population count < 40 were suppressed. For 100% data, the decision was based on population count. For 20% data, the number used was the lower of the 2A non-institutional population count or the 2B non-institutional population estimates. For products containing POW geographic areas, population counts < 40 were suppressed. For POW flow tables, both residence and work suppression rules applied. For products containing income characteristics, distributions or quantitative statistics, the threshold was 250.

Census years ▸ **Confidentiality rules ▾**	**1981**	**1986**	**1991**	**1996 to 2001**
Area suppression	Native's package, self-enumerated areas with native population count < 500 were suppressed; canvasser areas with native population count < 100 were suppressed. All suppressed data were included in higher geographic levels.	All suppressed data were included in higher geographic levels.	All suppressed data were included in higher geographic levels.	For products containing 6-digit postal code the threshold was 100. Suppression of data also occurs if and when certain questions are not asked of all respondents, such as on Indian reserves. All suppressed data were included in higher geographic levels.
Cell suppression	Custom tables involving industry, occupation or income, all cells < 25 were suppressed (set to 0).	Custom tables involving industry, occupation or income, all cells < 10 were suppressed (set to 0).	Income statistics based on cells < 10 were suppressed.	The standard deviation rule was established. If the standard deviation was equal to 0 indicating that all quantitative values were equal, then the statistic was suppressed. For all quantitative variables, statistics are suppressed if the associated unrounded count used in calculating the statistic is less than the random rounding threshold which is 10 for all 2B products.

Appendix B Table of data quality flags for historical census years

2001 Census

The following table describes the data quality indicator fields and their contents as they apply to the 2001 Census. Note that a zero in any of the five bytes is the default for the respective indicator and means that no data quality action is required.

Data quality flags – 2001 Census

Byte	Byte description	Byte value	Value description
1	**Incomplete enumeration**	0	<default> – Fully enumerated area.
		1	Incompletely enumerated Indian reserve or Indian settlement (suppressed).
		2	Excludes census data for one or more incompletely enumerated Indian reserves or Indian settlements.
2	**2A data quality**	0	<default>
		1	Data quality indexes showing a global 2A non-response rate higher than or equal to 5% and lower than 10%.
		2	Data quality indexes showing a global 2A non-response rate higher than or equal to 10% and lower than 25%.
		3	Data quality index showing a global 2A non-response rate higher than or equal to 25% (suppressed).
3	**Population and dwelling counts error flag**	0	<default>
		1	An error exists in the 2001 population and dwelling counts for this area.
4	**2B data quality**	0	<default>
		1	Data quality indexes showing a global 2B non-response rate higher than or equal to 5% and lower than 10%.
		2	Data quality indexes showing global 2B non-response rate higher than or equal to 10% and lower than 25%.

Data quality flags – 2001 Census (continued)

Byte	Byte description	Byte value	Value description
		3	Data quality indexes showing a global 2B or 2A non-response rate higher than or equal to 25% (suppressed).
5	**1996 adjusted population flag**	0	<default>
		1	1996 adjusted count; most of these are the result of boundary changes.

Prior to 2001 Census

The following table describes the data quality indicator fields and their contents as they apply to census years prior to 2001. Note that a zero in any of the five bytes is the default for the respective indicator and means that no data quality action is required.

Data quality flags – prior to 2001 Census

Byte	Byte description	Byte value	Value description
1	**Incomplete enumeration**	0	<default> – Fully enumerated area.
		1	Incompletely enumerated Indian reserve or settlement (suppressed).
		2	Excludes census data for one or more incompletely enumerated Indian reserves or Indian settlements.
2	**2A data quality**	0	<default>
		1	Data quality indexes showing a global non-response rate higher than or equal to 5% and lower than 25%.
		2	Data quality indexes showing a global non-response rate higher than or equal to 25% (suppressed).
3	**Boundary error**	0	<default>
		1	Areas where boundaries were found to be in error subsequent to collection.
4	**2B data quality**	0	<default>
		1	Data quality indexes showing a global non-response rate higher than or equal to 5% and lower than 25%.
		2	Data quality indexes showing global non-response rate higher than or equal to 25% (suppressed).

Data quality flags – prior to 2001 Census (continued)

Byte	Byte description	Byte value	Value description
		3	Employee resident error rate is higher than or equal to 5% and lower than 25% and global non-response rate is less than 5%.
		4	Both the employee resident error rate and global non-response rate are higher than or equal to 5% and lower than 25%.
		5	Employee resident error rate is greater than or equal to 25% and global non-response rate is less than 25% (suppressed).
5	**1996 adjusted land area**	0	No <default>
		1	Yes