# Alternative Ways of Assessing Confidentiality of Economic Statistics at Statistics Canada

## Author: Daniela Ravindra, Statistics Canada

## 1. Introduction

Statistical agencies across the world are facing more and more the challenge of having to balance their mandate of publishing timely, relevant and good quality information with their duty of protecting respondent information against divulgation. As sophisticated users demand access to more detailed data and as statistical information is made available by providers others than statistical agencies, National Statistics Offices (NSO) are under pressure to publish data that currently may be considered confidential. This is exacerbated by the fast change in attitudes towards confidentiality in an era dominated by social media giants, on-line services and innovation driven by crowdsourcing.

There are a number of options for a statistical organization which wants to decrease the number of suppressed data cells or increase the number of published data cells:

- Adjust the confidentiality provisions such that they are less stringent
- Improve the statistical software used to identify confidential cells such that it takes into account more parameters (e.g. wavers, weights, quality indicators, etc.)
- Use different methods of assessing confidentiality for different data sets depending on the perceived sensitivity of the data (e.g. two year old data may be perceived as less sensitive than last month's information)

This study focuses on confidentiality at Statistics Canada with respect to economic statistics only. All economic statistics programs, with the exception of the International Merchandise Trade, use an "active" approach to confidentiality: that is to say that cells deemed to be confidential based on internally developed rules were suppressed. This included sensitive data cells as well as data cells supressed for secondary confidentiality which, for tabular data publications increase exponentially with the number of dimensions in the table. The International Merchandise Trade program uses a "passive" approach, by which no information was suppressed unless a respondent requested it. Most programs use a common set of suppression rules which are embedded in a software package developed at Statistics Canada[1].

Statistics Canada started to review its approach to confidentiality in 2016 by focusing first on the System of National Accounts (SNA) publications. The reason for selecting the SNA was twofold: first because Statistics Canada was suppressing far more SNA information than other statistical agencies and second because the SNA outputs are the result of a sophisticated integration of survey and administrative data, hence a few levels removed from the information provided by respondents.

The review led to the development of an **Economic Statistics Data Suppression Decision Tree** which would provide program managers with a clear and consistent set of rules that can be used to determine when a given data point can be published such that it is not possible ***"to relate the particulars obtained from any individual return to any identifiable individual person, business or organization.[2]"*** Appendix 1 provides the detailed description of the Data Suppression Decision Tree. Based on step #3 of the Data Suppression Decision Tree, Statistics Canada stopped applying a confidentiality mask to the supply-use tables.

However, when it comes to the vast majority of outputs which are based on surveys, the question of what is confidential becomes more complicated as most disseminated data are closely based on reported data. Regardless, the options listed above still merit an investigation.

This study is focused on testing whether a number of annual economic surveys which are processed through a generic approach, could be published in their entirety by virtue of step #5 in the Decision Tree which refers to the data having been transformed through statistical procedures to such an extent that they are different (+/-3% on average) from the information provided by the respondent.

[1] For more information see http://www5.statcan.gc.ca/olc-cel/olc.action?ObjId=10H0109&ObjType=22&lang=en&limit=0

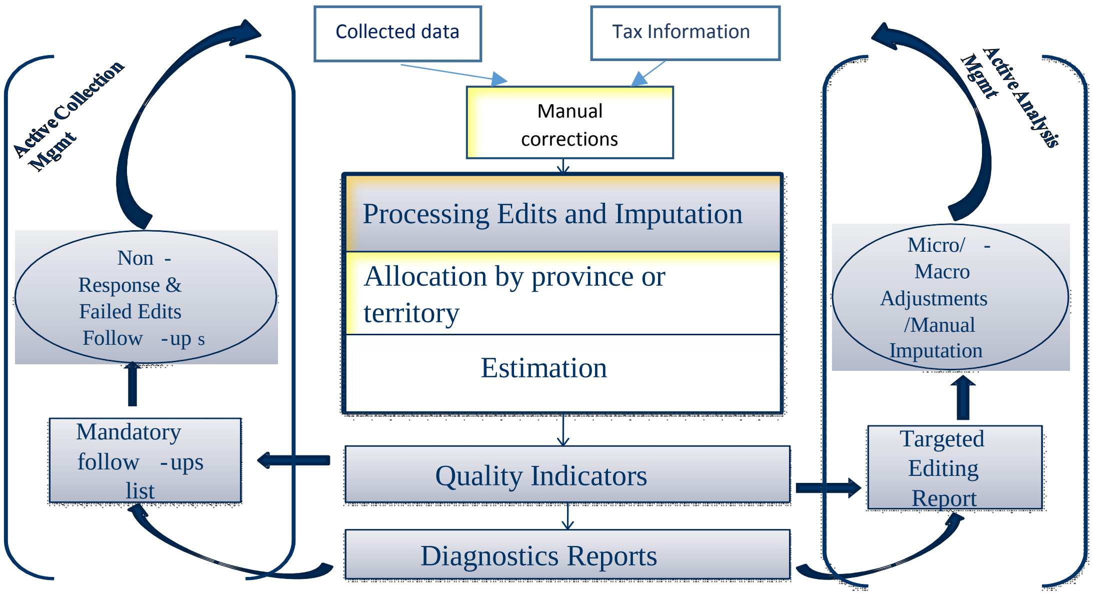
[2] Section 17(1) (b) of the Statistics Act which can be found at: http://laws-lois.justice.gc.ca/eng/acts/S-19/FullText.html.

## 2. Scope of Study

In 2014 Statistics Canada deployed a generic approach to collecting, processing and analyzing for most of its annual business surveys. The approach relies on common methods and tools to produce estimates and data suppression is subject to a common set of rules which use a minimum number of respondents per published cell, as well as dominance rules. The Annual Survey of Manufacturing (ASM)[3] is one of approximately eighty surveys using this common approach and it was selected for the study; results from the ASM study could be inferred to apply to the other surveys using the same survey processing methods.

**Figure 1** provides an overview of the processing: micro data collected at the national level is combined with tax data, it is manually corrected whenever necessary, edited and imputed for missing data, allocated by province or territory and rolled into estimates.

**Figure 1.**


Collected data
Tax Information
Manual corrections
Active Collection Mgmt
Active Analysis Mgmt
Processing Edits and Imputation
Allocation by province or territory
Estimation
Non-Response & Failed Edits Follow-ups
Micro/-Macro Adjustments /Manual Imputation
Mandatory follow-ups list
Quality Indicators
Targeted Editing Report
Diagnostics Reports

a

Estimates from the ASM are produced at the sub-national (provincial and territorial) level. Although some respondents are able and willing to provide information at this level of geography, many are not. Hence, allocating data at the provincial level is an important transformation step which, along with editing, imputation and manual corrections are the objective of this study to determine if these procedures add enough noise to provide a confidentiality mask to estimates which would otherwise be considered too sensitive to publish.

For ease of analysis, the study focused on revenue data cells where data quality was acceptable for publication but they were suppressed for primary confidentiality (the cell is sensitive) for all Canadian provinces and territories and all manufacturing industries The microdata in the suppressed cells were examined at each stage of data processing to determine the extent of the transformations applied to them. The ASM estimates are normally a combination of reported and administrative data: administrative data, really tax information, is used in lieu of reported information for the take-none component of the estimates, namely the small firms below a pre-defined threshold which are excluded from the collection processed and modelled. The take-none estimates were part of the study as well, but only in so much as to determine whether they were the reason for the suppression.

[3] Information on the ASM definitions, data sources and methods may be found at: http://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=2103

The impact of processing the data through the various steps was calculated by measuring the distance, in percentage points, between the reported values and the value at each stage of the processing chain. Three processing steps were examined: manual corrections, edit and imputation and allocation. Results were tabulated by geographical region (provinces and territories) as they represent the most common level of publication.

## 3. Results

In total 576 suppressed cells were examined. The idea was to determine if there was any evidence that the hypothesis that processing may introduce sufficient noise or distance from the reported data such that the cells should not be deemed confidential. First the results showed that the take-none estimates were almost never the reason behind the primary suppression. So the study then focussed just on the rest of the estimates (take all and take some). Table 1. below provides the average distance in percentage points between the data at each stage of processing and the reported data. It can be seen that each stage of processing ads to that distance and that generally, allocation contributes the most. This is not a surprising result as dominance tends to be the cause of a large number of primary suppressions and dominant firms often have establishments in more than one province and therefore requiring the allocation of the reported data by province and/or territory. The cumulative impact, which is measured after the allocation stage, varied from 8.59% in Province 2 to over 7,000% in Province 7. Manual corrections have a relatively low impact on the estimates, while edit and imputation contributes more than manual corrections but less than allocation. If we were to apply the 3% rules identified in the Data Suppression Decision Tree, it would seem that many cells currently deemed confidential would no longer be supressed.

**Table 1. Average Impact of Processing on Cells Deemed Confidential by Geography**

| Geography | Impact after Allocation | Impact after Edit and Imputation | Impact after Manual Correction |
|---|---|---|---|
| **Province 1** | 16.07% | 14.65% | 7.14% |
| **Province 2** | 8.59% | 3.63% | 0.20% |
| **Province 3** | 15.83% | 6.15% | 5.87% |
| **Province 4** | 27.52% | 15.65% | 8.83% |
| **Province 5** | 15.87% | 12.45% | 7.60% |
| **Province 6** | 152.77% | 187.40% | 186.95% |
| **Province 7** | 72231.4% | 72222.7% | 3.77% |
| **Province 8** | 17.15% | 8.14% | 5.31% |
| **Province 9** | 49.92% | 8.68% | 5.23% |
| **Province 10** | 23.92% | 7.10% | 7.00% |
| **Territory 1** | 10.36% | 3.27% | 0.92% |
| **Territory 2** | 51833.2% | 51817.0% | 51814.8% |
| **Territory 3** | 41.86% | 24.85% | 8.92% |

However, as average results provide only one element of the picture, Table 2 illustrates the frequency of impacted data cells by geography and impact range, Less than 3% was selected as it is in line with the provisions of the Data Suppression Decision Tree, while the other ranges were chosen to reflect more or less the broad categories of the results. The results reinforce the picture provided by Table 1, namely that each stage of processing takes the reported cells further from the reported data. After allocation, 31% of cells are impacted by more than 3% of their original value.

**Table 2. Number of Confidential Cells Impacted by Processing by Geography and Range of Impact**

| Geography | Impact after Allocation | | | Impact after Edit and Imputation | | | Impact after Manual Correction | | |
|---|---|---|---|---|---|---|---|---|---|
| | [0%-3%) | [3%- 25%) | [25% +) | [0%- 3%) | [3%- 25%) | [25% +) | [0%-3%) | [3%-25%) | [25% +) |
| Province 1 | 65 | 5 | 13 | 68 | 3 | 12 | 76 | 3 | 2 |
| Province 2 | 51 | 6 | 7 | 57 | 5 | 2 | 64 | . | . |
| Province 3 | 44 | 9 | 14 | 52 | 10 | 5 | 56 | 6 | 5 |
| Province 4 | 40 | 8 | 22 | 48 | 9 | 13 | 58 | 4 | 8 |
| Province 5 | 10 | 3 | 4 | 11 | 3 | 3 | 14 | 1 | 2 |
| Province 6 | 10 | 1 | 1 | 10 | 1 | 1 | 11 | . | 1 |
| Province 7 | 53 | 5 | 13 | 55 | 10 | 6 | 59 | 9 | 3 |
| Province 8 | 42 | 4 | 12 | 47 | 5 | 6 | 50 | 5 | 3 |
| Province 9 | 18 | 3 | 12 | 25 | 5 | 3 | 29 | 2 | 2 |
| Province 10 | 14 | 1 | 7 | 19 | 1 | 2 | 19 | 1 | 2 |
| Territory 1 | 28 | 1 | 5 | 28 | 4 | 2 | 30 | 4 | . |
| Territory 2 | 20 | 4 | 10 | 23 | 7 | 4 | 28 | 3 | 3 |
| Territory 3 | 6 | . | 9 | 6 | 3 | 6 | 9 | 3 | 3 |
| Total | 401 | 50 | 129 | 449 | 66 | 65 | 445 | 41 | 34 |

This focused investigation has shown that, in the case of the ASM, data processing puts an important amount of distance between the reported and the final microdata point. However, given that a great number of cells are supressed to protect dominant firms, the question that also needs to be answered is whether these adjustments made to the reported data represent sufficient protection for data providers. They answer may be "yes" as pre the requirements of the Statistics Act however, it may still not address the moral duty of Statistics Canada for protecting respondents' right to keep information about their companies confidential. Even though, technically, Statistics Canada would not be releasing the reported data, would it still be giving the impression that it is publishing data which companies would consider sensitive, data which could provide competitors with a glimpse of their operations and possible a competitive advantage.

These questions are very pertinent not just to this study, but to another confidentiality protection method which Statistics Canada is currently testing: Random Tabular Adjustment (RTA). The RTA includes both a disclosure control framework based on Bayesian decision theory for the identification step, and random tabular adjustment which adjusts the domain totals for the protection step. It works by selecting quality thresholds and a prior model, it determines quality of posterior estimates and then finds the smallest variance for the cell adjustments so that the quality of the posterior estimates are below the threshold of the prior model. Currently, Statistics Canada is doing tests of the RTA to determine whether the model provides adequate protection without distorting the quality of the estimates beyond the acceptable level.

## 4. Conclusion

This small and focused study is but one of the many avenues that Statistics Canada is currently pursuing to assess the potential for reducing the number of suppressed cells and improve the availability of data for users. Although the study is not conclusive, it does provide some avenues for further investigation. It has also served to provide confirmation that data obtained from respondents for the ASM is of good quality given the low impact of edit and imputation and manual corrections on the cells examined. This is welcome news as the generic data processing approach introduced in 2014 were meant to reduce manual intervention and to improve data quality.

## 5. References:

Saint-Pierre E. and Bricault M (2011), The Common Editing Strategy and the Data Processing of Business Statistics Surveys, Statistics Canada, Conference of European Statisticians, Slovenia 2011

Tebrake, J. (2017), Economic Statistics Data Supression Decision Tree, Statistics Canada Internal Document, Ottawa 2016

Wright, P. (2016), Current Developments in Disclosure Control for Business Surveys at Statistics Canada, Advisory Committee on Statistical Methods, Ottawa 2016

# Appendix 1.

**Economic Statistics Data Suppression Decision Tree**

**Definition**: The term "Data cell" represents the intersection of dimensions in a data table (such as the intersection of industry, geography and time, or the intersection of an industry, commodity) pertaining to a variable that is proposed for public dissemination. The secrecy provisions of the Statistics Act are applied at the data cell level.

***Revenue, 2016***

| **Industry/Geography** | **Geography 1** | **Geography 2** | **Geography 3** | **Geography 4** |
|---|---|---|---|---|
| Industry 1 | (Data Cell) | | | |
| Industry 2 | | | | |
| Industry 3 | | | | |
| | | | | |

Program managers are required to answer and provide supporting documentation for each of the following questions which will identify the data cells that can be released and those that need to be suppressed prior to the dissemination of the data table.

1. Has the data cell proposed for release been obtained by means other than the Statistics Act?

   No – The data cell proposed for release was obtained by means of the Statistics Act. Further assessment is required.

   Yes – The data cell may be published as per the terms and conditions under which Statistics Canada obtained the Information.

2. Has the data cell proposed for release been authorized to be disclosed by order by the Chief Statistician as per 17(2) of the Statistics Act?

   (Selected domain cells) Yes – the data cell can be released as per the order issued by the Chief Statistician.

   (Mostly) No – Further assessment is required.

3. Has the data cell proposed for release been mapped to a classification system (such as a product, industry or geographic classification system) or conceptual framework and aggregated according to the classification or conceptual framework such that it no longer represents any person, business or organization but rather an aggregation of persons, businesses or organizations?

   Yes – The data cell was derived after having been mapped to a classification system (NAICS and geographic classification system). The data cell has undergone a transformation such that it no longer represents the person, business or organization from which it was collected but rather an aggregation of persons, businesses or organizations related to an industry, product, geography or other grouping?

   No. Further assessment is required.

4. Are there 3 or more individuals, businesses or organizations in the population (e.g. business register, individual register) that could engage in the activity represented by the data cell such that it would not be possible by a third party (i.e. a party other than Statistics Canada or the responding individual, business or organization) to relate the information to be released to any identifiable individual person, business or organization?

   Yes – there are 3 or more individuals, businesses or organizations that could engage in the activity represented by the data cell such that, through the release of the information, a third party (a party other than Statistics Canada or the responding individual, business or organization) would not be able to relate the data cell released by Statistics Canada to any identifiable individual person, business or organization. The data cell can be released subject to residual and sensitivity disclosure rules.

   1388 of 1512 cells have 3 or more businesses

No –There are 2 or fewer individuals, businesses or organizations in the population that engage in the activity represented by the data cell and therefore users may be able to relate the data cell proposed for release by Statistics Canada to any identifiable individual person, business or organization. Further assessment is required.

5. Has the data cell proposed for release been subjected to a statistical and/or conceptual transformation(s) such that a third party (a party other than Statistics Canada or the responding individual, business or organization) cannot with certainty relate the data cell proposed to be released to the particulars obtained by Statistics Canada from any individual return?

   Yes – the data cell has been integrated with other information into an internationally or nationally accepted economic statistics conceptual framework; or has been subjected to statistical procedures (indexing, modelling, editing, imputation, allocation, estimation, valuation) such that studies have demonstrated that the data cell is on average +/- 3% different from the particulars obtained by Statistics Canada from any individual return.

   The data cell can be released subject to residual and sensitivity disclosure rules.

   No – The data cell should be suppressed as per the secrecy provision 17(1)(b) of the Statistics Act.

6. Does any data cell reveal by residual disclosure (that is by indirect calculation) a data cell that has been suppressed from release as per the secrecy provision of the Statistics Act or does the program manager feel that the release of the data cell is of a sensitive nature or compromises the future receipt of information?

   Yes – the information falls under the secrecy provisions of the Statistics Act and cannot be released.

   No – The data cell can be released because it cannot be used by indirect calculation to disclose a suppressed data cell and/or it is not of a sensitive nature (e.g. the main contributors to the data cell are publically traded firms that are required by legislation to publish their financial statements).