UNECE

CONFERENCE OF EUROPEAN STATISTICIANS

Expert Meeting on Statistical Data Confidentiality

1-3 December 2021, Poland

Increasing utility of economic statistical information.

Steven Thomas (Statistics Canada)
steven.thomas@canada.ca

Abstract

Statistics Canada is implementing transformational change to its approach to protect business data. For some business surveys, the use of complementary cell suppression (CCS) for tabular results will be phased out entirely in what is termed the 'Move to Zero Suppression'. To this end, Statistics Canada has starting releasing business data protected by perturbation methods that balance the precision of the estimate with the uncertainty associated with a contributor's value. For those business surveys that continue to use CCS, Statistics Canada is improving both the assessment of sensitivity and the implementation of the suppression pattern. To meet the needs of a broader group of researchers, Statistics Canada is examining solutions to release microdata. Synthetic business data is being examined, while other options within the paradigm of the five safes are also being tested and implemented.

Increasing utility of economic statistical information

Steve Thomas[1]

Statistics Canada, steven.thomas@statcan.gc.ca

Abstract: Statistics Canada is implementing transformational change to its approach to how it balances the necessity to produce data for Canadians while continuing to protect Canadian business data. For some business surveys, the use of complementary cell suppression (CCS) for tabular results will be phased out entirely in what is termed the 'Move to Zero Suppression'. To this end, Statistics Canada has started releasing business data protected by perturbation methods that balance the precision of the estimate with the uncertainty associated with a contributor's value. For those business surveys that continue to use CCS, Statistics Canada is improving both the assessment of sensitivity and the implementation of the suppression pattern. To meet the needs of a broader group of researchers, Statistics Canada is examining solutions to release microdata. Synthetic business data is being examined, while other options within the paradigm of the five safes[2] are also being tested and implemented.

1 Background

The Government of Canada and more specifically Statistics Canada aims to ensure that that Canada has the data it needs to make evidence-based decisions. This includes ensuring that as much information as possible is made available to its user community. Information is currently made available to the research community through a variety of access options, one of which is the tabular data that is available on the Statistics Canada website. As many programs attempt to increase the granularity of data being published with these tables, confidentiality risks begin to arise as analytical information approaches individual information. These confidentiality concerns are typically assessed through a standard risk assessment and the risks are mitigated through a standard complementary cell suppression (CCS) strategy. As more granularity is added, the risk of disclosure increases and our promise to protect that information becomes more challenging.

The move away from suppressing data is obviously complicated. The current process works at identifying and treating risks. To move away from that involves a complete evaluation of (1) the understanding of what is confidential, (2) how to identify what is considered a disclosure (Risk scenarios), and (3) how to treat those disclosure risks when they are identified. The entire strategy of risk identification and treatment must be balanced with the other obligation of the Statistics Act[3] which is to publish information relating to the economic activities of Canadian Businesses.

This paper aims to illustrate some of the progress that Statistics Canada has made towards what has been termed 'Move towards zero suppression'. This includes understanding what is

[1] The views expressed in this presentation are those of the author alone and do not necessarily represent the position of Statistics Canada or the Government of Canada in these matters.

[2] www.fivesafes.org

[3] Statistics Act (R.S.C., 1985, c. S-19)

confidential, developing more appropriate risk scenarios, using the appropriate methods and tools to identify those agreed upon risks to finally mitigating those risks with solid defendable strategies. With each step, strategies will be identified to help reduce the amount of suppression, while still controlling for the risk of disclosure, in the final output tables.

2 Understanding the risks

The Statistics Act points out many terms and ideas that make it difficult to implement in a data driven mathematical framework. In the case of business surveys, a mathematical framework based on risk and standard tools to identify those risks have been developed and used for decades. A move away from this risk model must first include an understanding of what risks are being identified and treated with the current risk strategy.

The Statistics Act states 'No person who has been sworn under section 6 shall disclose or knowingly cause to be disclosed, by any means, any information obtained under this Act in a manner that it is possible from the disclosure to relate the information obtained to any identifiable individual person, business or organization.'

The interpretation of this phrase has severe implications on how to implement tools to identify and treat disclosure risks. Readers should note that in this phrase, the term 'disclose' is not defined in a way that allows implementation in a mathematical context. Neither are the specifics of what is meant by 'relate the information obtained' or 'identifiable person business or organization'.

A literature review will not immediately help with defining these terms in a way that easily allows implementation in a statistical program. The basic Oxford definition of disclosure suggests 'the action of making new or secret information known'. The term Statistical Disclosure may be more applicable. One of the early definitions (Dalenius, 1977) suggests 'If the release of the statistic S makes it possible to determine the value more accurately than it is possible without access to S, a disclosure has taken place'. Another definition (Cox, 1980) suggests 'Disclosure in both categorical and magnitude data arises from unacceptable narrow estimation of the values of statistical cells'. Both of these early definitions define the disclosure control problem in terms of uncertainty although accuracy and 'narrow estimation' (precision) are two different concepts that adds to the confusion of disclosure. Can a precise value for a respondent be released if it is not an accurate portrayal of what was collected? Can an accurate value be released if it is imprecise?

This leads to the UNECE definition of statistical terms in the OECD Glossary of Statistical Terms[4]. This list was developed and supported by the UNECE and the following are the standard definitions that will be used throughout this article.

> **Disclosure:** Disclosure relates to the inappropriate attribution of information to a data subject, whether an individual or an organization. Disclosure has two components: identification and attribution.
>
> **Disclosure Risk:** A disclosure risk occurs if an unacceptably narrow estimation of a respondent's confidential information is possible or if exact disclosure is possible with a high level of confidence.

[4] Source OECD Glossary of Statistical Terms https://stats.oecd.org/glossary

> **Identification:** The association of a particular record within a set of data with a particular population unit.
>
> **Attribution:** The association or disassociation of a particular attribute with a particular population unit.

The idea of disclosure risk will be used through the rest of the document yet still requires some interpretation. ‘An unacceptable narrow estimation’ has traditionally been related to ‘precision’ – the closeness of values together. This is seen by some as a conservative approach where values that are precise but not accurately reflecting what was collected could be considered for release under alternative risk models. The challenge is setting up such a model that allows for this nuance to be considered and these ideas are not fully developed in the theory and tools used with disclosure control. As an example suppose that ‘total revenue’ was derived from ‘total sales’ using a transformation that only the data custodian was aware of. Traditionally, any risk of precise values coming out of total revenue would be considered a disclosure while more recently there has been appetite to consider it non-disclosive if the transformed value is far enough or biased enough from the value collected under the Statistics Act (total sales).

The second interpretation is around ‘unacceptably narrow’. Statistics Canada has developed thresholds in the application of rules for precision measures. These thresholds have been set as the standard for any financial information being disseminated. The default thresholds are not necessarily applied in the case of information that is less sensitive where its disclosure will do less harm to a business. An example here might be total acreage of land in an agricultural context where the information is practically public or easily determined through outside sources of information. It may not require the same protection as more sensitive information such as total revenue from sales.

The final piece of information that is missing is an understanding of the risk scenario that is being mitigated through our disclosure control strategies. This is often termed as the attack scenario. Even though not included in the definition of disclosure risk, there is an idea of ‘learning something new’ that underlies the idea of disclosure. The attack scenario helps us to model what is already known and determine if something new can be learned through a release. It is generally accepted that in a business survey context the contributors are known and that the values of each contributor are already known to a certain level of precision. This is also being reconsidered when information that is less visible in nature is being collected. There are some situations where even participating in the activity would not be known or easily determined. An example is research spending in specific areas.

3 Traditional Risk Assessments

In application, Statistics Canada has several disclosure control strategies that were originally built on minimum count and the n,k dominance style rules where:

> A cell is regarded as confidential, if the *n* largest units contribute more than *k %* to the cell total, e.g. *n*=2 and *k*=85 means that a cell is defined as risky if the two largest units contribute more than 85% to the cell total. The *n* and *k* are determined by the statistical authority. In some National Statistical Institutions the values of *n* and *k* are confidential. (OECD Glossary of Terms)

Although related to the disclosure risk described above, the reader will note that there is no direct relation between dominance rules and disclosure risk. The weaknesses of these approaches have been described in the past, for example in the following scenario by Robertson and Ethier (2002):

In the dominance rule, let $n = 1$ and $k = 0.6$ (60%). Then a cell with value 100 and contributions 59, 40, 1 is declared not sensitive, while a cell with value 100 and contributions 61, 20, 19 would be declared sensitive. Assume now that the second largest respondent of both cells knows the total 100 and is interested in estimating the contribution of the largest respondent. Then, for the (59, 40, 1) cell, she removes her contribution and gets an upper bound $100 - 40 = 60$ for the largest contribution. For (61, 20, 19) the upper bound she gets is $100 - 20 = 80$, much farther from the real largest contribution. So the cell declared non-sensitive by the rule allows better inferences than the cell declared sensitive.

As a more direct measure, Statistics Canada and other statistical organizations have been using the prior-posterior rule (Willenborg and de Waal, 2001) also termed the *p-percent rule* or the *p,q rule* as the standard which is defined as

> It is assumed that, prior to the publication of the table, every respondent can estimate the contribution of each other respondent to within *q* percent. A cell is considered sensitive if someone, e.g. one of the respondents, can estimate the contribution of an individual respondent to that cell within *p* percent after (i.e. posterior to) publication of the table.

This better formalizes disclosure and the idea that competitors should not have their values disclosed to each other through statistical publications. Traditionally, the formalization of this prior-posterior rule used a geometric approach where the idea of precision is described with intervals and lengths. Cells where the contributed value can be determined within the relative value *p* are considered sensitive.

4 Traditional Mitigation Strategies

Mitigating the risks that are identified with the standard *p*% rule involve not releasing the statistic through the process of suppression. Users of this method realize that suppressing the cell without recognizing the additive structure of the table does not adequately mitigate the risk where a precise value can be recalculated through simple subtraction from the aggregated totals. With this approach, the uncertainty is introduced through complementary cell suppression (CCS). By suppressing complementary cells that correspond to *p%* of the sensitive value, an attacker will be left with a feasible range for the cell which can be used to derive a feasible range for the individual contribution. One nuance with this is that for the naïve user, it may appear that they do not have any information since the cell value nor its feasible bounds are released but for the more advanced user, they will realize that bounds are available for the cell and any value within the bounds could be calculated to represent the cell value.

5 Challenging the Norm

5.1 The introduction of perturbation strategies

Alternative to the traditional approach, a similar formalization of the risk assessment can be made using a probabilistic approach with distributions and variances instead of intervals and lengths. In this formulation, Stinner (2017) presents the Random Tabular Adjustment (RTA) strategy where the precision of prior and posterior knowledge can be described with related measures of standard errors or variances. The coefficients of variation (CV) defined as the standard error of the attackers estimate divided by the contribution to the cell offers a relative measure that can be considered the value *p* in the Willemborg definition. Under this interpretation, cells where the posterior relative precision is less than *p* (a prescribed CV) are considered sensitive.

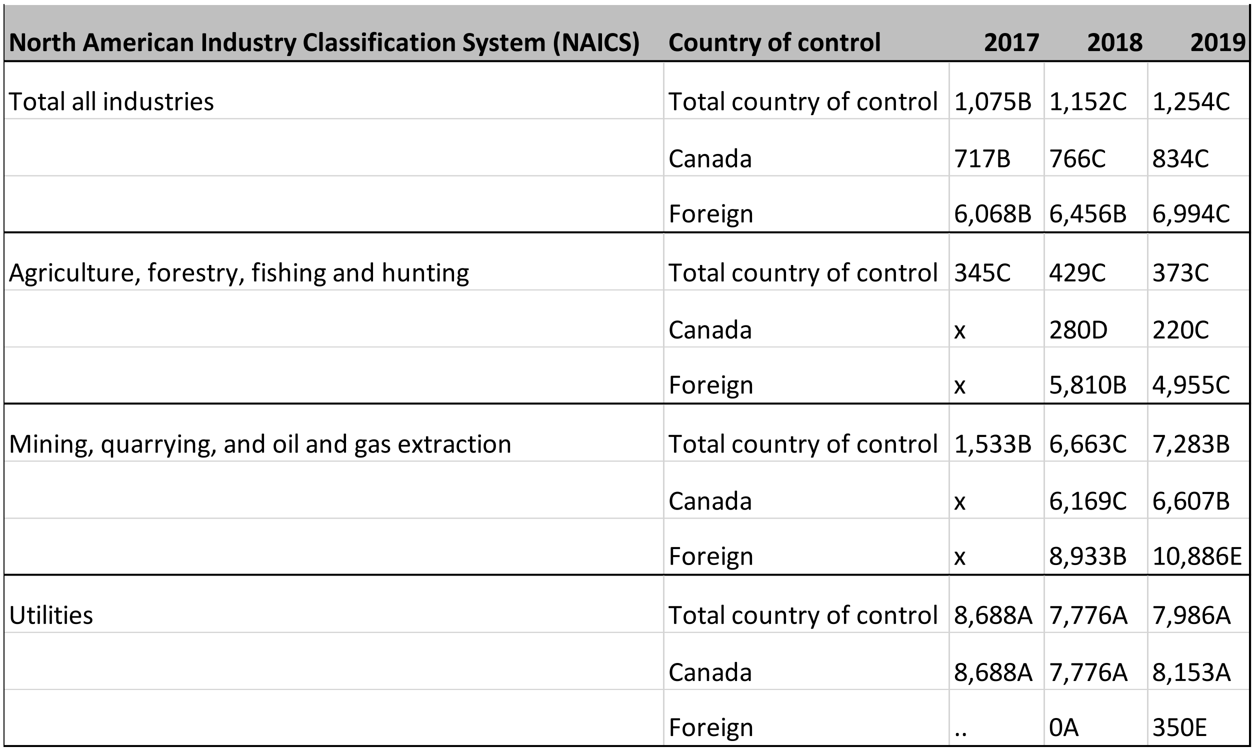
For the risk mitigation, instead of suppressing secondary cells, the uncertainty is introduced by adding a random adjustment based on a normal distribution to the cell. In this situation, the user is left with an estimate along with some quality measure (CV, variance, standard error, etc) that reflects the noise of the estimate. In a table structure, the noise will propagate into aggregated values. One interpretation of this fact suggests that choosing the RTA method means to choose releasing high quality internal cells over releasing precise aggregates.

In the move to releasing more information, this strategy helps release an explicit estimate for each cell. Sensitive cells will be affected directly. Non-sensitive internal cells will not be affected while aggregate cells will be affected to a lesser relative degree. The explicit nature of RTA is perhaps its drawback from a user acceptance point of view. By providing an estimate, the naïve user may feel that we have released more information than we have with the geometric approach when in fact, with proper parameter specification, the protection can be quite similar. This challenge may be result of the strategy being less intuitive than the suppression strategy. The RTA strategy relies on a strong understanding of disclosure to be able to accept that we have not disclosed and; an understanding of estimation and precision measures to be able to interpret the risk and utility associated with this strategy. The challenges to the approach seems to be more related to the pragmatic and ethical considerations of disclosure control rather than the legal aspects.

5.2 RTA Example

As an example, the 2017 version of the Annual Survey of Research and Development in Canadian Industry use a suppression methodology while the 2018 and 2019 version used RTA (See Table below). The example illustrates that estimates for cells that would be suppressed under a geometric strategy have been published with the RTA strategy along with the appropriate quality indicators. There are a few things to note. First, the quality of sensitive cells are generally poor while non-sensitive cells are pretty good. This reflects the desired feature that sensitive estimates should not be precise. Second, note the slight decrease in quality for published statistics where noise at the detailed level has a slight effect on the larger aggregated statistics. For some programs these features have been seen as limiting factors preventing adoption of the approach. Finally, note the footnote where Statistics Canada is explicit and transparent to the user that the data have been perturbed to protect the confidentiality of the respondents.

The RTA approach also has an added feature that existing noise (sampling and non-sampling errors) can be considered when specified to the system. When there is already enough noise to protect the contributions, the RTA will not add additional noise.



North American Industry Classification System (NAICS)	Country of control	2017	2018	2019
Total all industries	Total country of control	1,075B	1,152C	1,254C
	Canada	717B	766C	834C
	Foreign	6,068B	6,456B	6,994C
Agriculture, forestry, fishing and hunting	Total country of control	345C	429C	373C
	Canada	x	280D	220C
	Foreign	x	5,810B	4,955C
Mining, quarrying, and oil and gas extraction	Total country of control	1,533B	6,663C	7,283B
	Canada	x	6,169C	6,607B
	Foreign	x	8,933B	10,886E
Utilities	Total country of control	8,688A	7,776A	7,986A
	Canada	8,688A	7,776A	8,153A
	Foreign	..	0A	350E

*Values may have been randomly adjusted to meet the confidentiality requirements of the Statistics Act. The quality indicators reflect the impact on the precision of the estimate of sampling, non-response and the random adjustment when applied.

Symbol legend:

..	not available for a specific reference period
x	suppressed to meet the confidentiality requirements of the Statistics Act
E	use with caution
A	data quality: excellent
B	data quality: very good
C	data quality: good
D	data quality: acceptable

Table: Annual Survey of Research and Development in Canadian Industry - Average business enterprise in-house research and development expenditures. Source: https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=2710000101

6 Other strategies to release more data

6.1 'Decision Tree' logic

Recent moves to releasing more data have put into question many of the standards that have been described here. Passive approaches to confidentiality have been used in specific situations. The branch has also developed a 'decision tree' logic which offers solutions to the question of whether a data cell should be considered disclosive or not for the system of national accounts (Ravindra, 2017). This includes first ensuring that the information must be protected and treated as confidential. Waivers or certain pieces of public information do not require the protection measures to be applied and programs are being encouraged to investigate those options. Another main consideration that is outlined is the question of 'Has the data cell proposed for release been subjected to a statistical and/or conceptual transformation(s) such that a third party cannot with certainty relate the data cell proposed to be released to the particulars obtained by Statistics Canada from any individual return'. This goes back to the questions that are outlined in section 2 and introduces new challenges for application including what transformations are acceptable and how to measure their effectiveness in protecting respondent values. These ideas must be developed further to be applied to survey responses. If the data point is precise but an inaccurate reflection of what was collected, does it need protection?

The 'decision tree' also offers a risk assessment that questions if 'three or more individuals, businesses or organizations in the population that could engage in the activity represented by the data cell'. If yes, the cell is considered safe for release. This is challenging since it is not always clear what 'could' implies. It also has an underlying assumption that the data cell cannot be related back to an individual. These ideas are inconsistent with the attack scenarios developed with the CCS approach and RTA. They require further development if automated solutions are to be implemented where mitigation strategies will be required for those cells with fewer than three in the population. Gray (2016) offers an alternative framework that may offer the flexibility required to incorporate some of these ideas.

6.2 Taking more disclosure risk

The challenge with both the geometric and the probabilistic approaches is assigning proper thresholds for the value *p*. Both approaches are quite flexible. Traditionally, the geometric rules are defined with fixed parameters in what Statistics Canada has labelled the C2 rule and are implemented in the automated disclosure control system G-Confid (Frolova, 2009). This strategy is generally accepted but with objectives to release more data some of the assumptions are being challenged in order to increase the utility. The challenges include questioning the standard attack scenario where the contributors are known, questioning the prior knowledge parameter *q* and questioning the amount of protection required *p*. Similar challenges exist with the implementation of the RTA method where a prior knowledge distribution must be determined along with the appropriate risk tolerance thresholds.

6.3 Risk Parameterization

The simplest and perhaps the most common strategy is to take more risk by redefining the required protection (*p*). In some cases, this decision is based off of the sensitivity of the information where it is recognized that more visible public information requires less protection than sensitive information. Public information may include things like acreage of a farm while sensitive information may include financial attributes. A sensitivity impact assessment is used at Statistics Canada when assessing the sensitive nature of information.

Sensitivity assessment based on impact	Description
Negligible	The information breach would not cause any harm as the information is considered to be publicly available.

Low	The information breach would cause minimal harm to an individual or business (e.g., because the information is easily accessible or released without concern but not publicly available).
Medium	The information breach would cause reputational damage or embarrassment to an individual or to a business.
High	The information breach would cause considerable harm to an individual or business (e.g., to the financial situation of an individual or business).
Severe	The information breach would cause severe harm to an individual or business (e.g., to the safety or health of an individual, to an enterprise's ability to do business, including expensive litigation or, in extreme cases, closure).

Table: Sensitivity impact assessment

One of the ways forward with determining the appropriate risk parameter is considering the balance with utility. The standard utility measures for CCS approaches have been the proportion of cells in a table being released or the proportion of table value. The RTA can offer a more direct measure where the user can see the direct effect on the quality of the cell. The approach allows the custodian to fine-tune the adjustment whereas the CCS approach gets to a point where the utility cannot be pushed further. In the figure below, it is seen that there is no way to increase the proportion of cells being released over 90% because at the very least cells with fewer than 3 units must be suppressed under our attack scenario and some complementary cells must also be suppressed. The RTA as an alternative method can be pushed to where its effects on CVs are negligible. The utility measures do not offer direct solutions but can be used to determine if the increase in utility are worth taking extra risks.

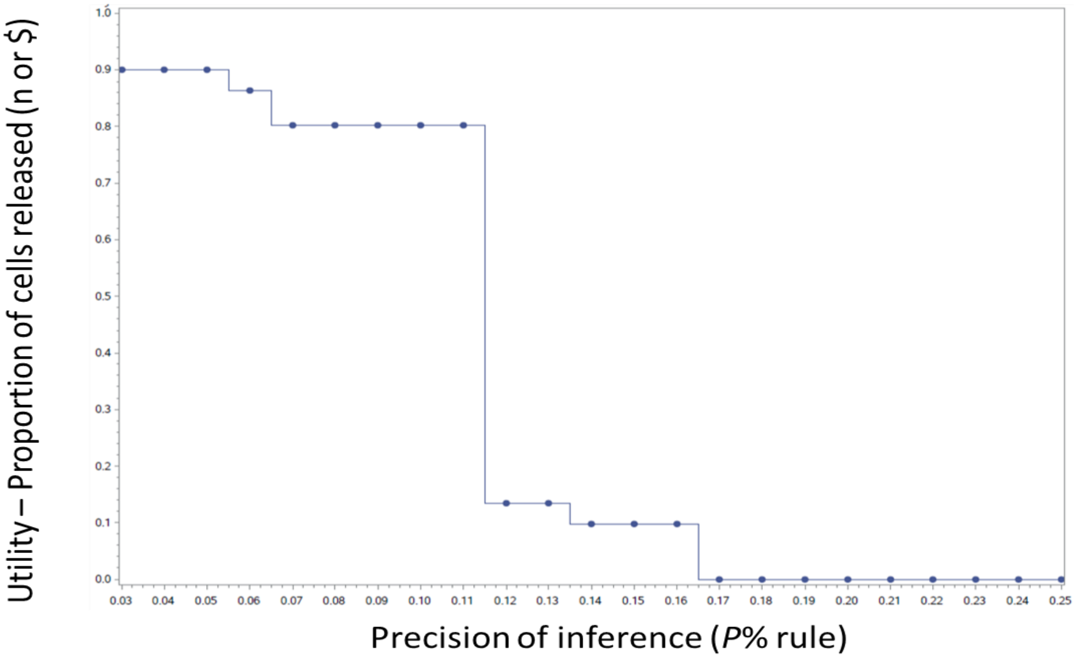


Figure: Risk versus Utility of Complementary Cell Suppression Strategy (Example)

7 A note on synthetic data

This paper has noted the main strategies being used to address user needs in accessing tabular data. Another strategy being investigated is the creation of synthetic business data. To date, no synthetic or public version of a microdata economic dataset has been released by Statistics Canada. The challenge has been investigated several times and it continues to get stalled at the issues with defining and understanding disclosure as described earlier. First and foremost, there are challenges with addressing disclosure as defined in section 2. In that definition, there is the idea that disclosure has two components: identification and attribution. Our standard strategies in anonymizing datasets has relied on preventing identification. In the case of economic activity in Canada, there are three issues with this - there is an ease in obtaining business information, the populations are small and they are highly skewed in terms of magnitude values. It has been continually realized that useful microdata sets will have challenges with re-identification.

A proposed way forward is to begin to accept the idea that identification in the absence of attribution is not disclosure. This approach is consistent with the tabular data strategies where it is assumed that intruders can identify which businesses contributed to a cell. Also supporting this strategy is the idea that under section 17 (2) (f) of the Statistics Act, the basic identifying information related to a business can be released. This suggests that standard methods of swapping, suppression, trimming, and model-based synthesis are appropriate. Once a safe dataset is made available, it is understood that the tabular outputs based off of those datasets will not be of concern.

It is unlikely that all economic datasets will be publicly available in the near future. We are still asking 'How much noise will make these files safe enough?', 'How can we illustrate the noise addition to the user and the data provider?' and 'What will the public's reception be?'. Instead, it is likely that our way forward will be taking advantage of an alternative access scenario where the files are seen as somewhat risky and therefore only available to trusted researchers, in trusted environments with a demonstrated need to access the data.

8 Conclusion

In order to implement transformational change to Statistics Canada's approach to protect business data, a novel interpretation and mathematical translation of risks may be needed. These risks appear whether releasing information based off of alternative risk scenarios or implementing alternative risk identification and mitigation under traditional scenarios. The 'move to zero suppression' had been a success for some programs while others may not be able to diverge from existing strategies because of operational constraints in their given context. Synthetic data is a solution for some but still requires work before it is seen in the mainstream of Statistics Canada releases.

References

Cox, L. (1980). Suppression Methodology and Statistical Disclosure Control. Journal of the American Statistical Association, Volume 75, Issue 370.

Dalenius, T. (1977). Towards a methodology for statistical disclosure control. Statistisk tidskrift. Statistical Review, Volume 15.

Elliot, M., A. Hundepool, E. Schulte Nordholt, J.L. Tambay and T. Wende, 2006. *Glossary on Statistical Disclosure Control*. In: Monograph of official statistics, Work session on statistical data confidentiality, Geneva, 9-11 November 2005, Office for Official Publications of the European Communities, 2006, pp. 381-392.

Frolova, O., Fillion, J., and Tambay, J. (2009) Confid2: Statistics Canada's new tabular data confidentiality software. Proceedings of the Survey Methods Section, SSC 2009.

Gilchrist, P. (2020) Public Use Microdata Files and Business Data: A Summary of Challenges for Confidentiality. Statistics Canada Internal Document.

Gray, D. (2016). *Precision Threshold and Noise: An Alternative Framework of Sensitivity Measures*. International Conference on Privacy in Statistical Databases.

Ravindra, D. (2017). *Alternative Ways of Assessing Confidentiality of Economic Statistics at Statistics Canada*. UNECE European Establishment Statistics Workshop 2017.

Robertson, D. and Ethier, R. (2002). Cell suppression: Theory and experience, in Inference Control in Statistical Databases, LNCS 2316, ed. J. Domingo-Ferrer. Berlin: SpringerVerlag, pp. 9-21, 2002

Stinner, M. (2017). Disclosure Control and Random Tabular Adjustment. SSC Annual Meeting, June 2017.

Willenborg, L. and De Waal, T. (2001) Elements of statistical disclosure control, Springer, New York, NY.