CONFERENCE OF EUROPEAN STATISTICIANS
Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality Session 5
29-31 October 2019, the Hague, the Netherlands 23 September 2019

Successes and Challenges in Increasing Accessibility at Statistics Canada

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

Abstract and Paper

Statistical Disclosure Control and protection of confidentiality have always been of the utmost importance to Statistics Canada. The legal obligations to protect personal information are reinforced with our ethical and pragmatic reasons to protect the information entrusted with us from our survey and administrative respondents. These obligations are what give confidence to the public that their data are safe, but at the same time are what limit our ability to meet the needs of the ever-growing group of researchers and analysts who want access to our information. The challenge for Statistics Canada is to ensure that we continue to respect and even strengthen our promise to keep personal information safe while at the same time ensuring that researchers, analysts and policy makers are given access to the information that they need to move ideas forward in a data driven society. This contribution will give an overview of some of the initiatives Statistics Canada has undertaken to expand access to data and some of the areas that need further work as it moves forward.

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE
Work Session on Statistical Data Confidentiality
(The Hague, Netherlands, 29-31 October 2019)

Successes and Challenges in Increasing Accessibility at Statistics Canada

Prepared by Steven Thomas, Statistics Canada[1]

1. Abstract

1. Statistical Disclosure Control and protection of confidentiality have always been of the utmost importance to Statistics Canada. The legal obligations to protect personal information are reinforced with our ethical and pragmatic reasons to protect the information entrusted with us from our survey and administrative respondents. These obligations are what give confidence to the public that their data are safe, but at the same time are what challenge our ability to meet the needs of the ever-growing group of researchers and analysts who want access to our statistical information. The challenge for Statistics Canada is to ensure that we continue to respect and even strengthen our promise to keep personal information safe while expanding access in a data driven society. This contribution will give an overview of some of the initiatives Statistics Canada has undertaken to expand access to data and some of the areas that need further work as it moves forward.

2. Background

2. The Government of Canada has developed a data strategy roadmap[2] that recognizes the need for a modern look at how the government accesses and shares data. Under the vision of modernization at Statistics Canada, one of the main goals is to 'increase access to data and microdata to drive innovation and inclusion'[3]. The department has prioritized research and development of methodological solutions to promote access options that respect privacy and offer appropriate confidentiality protection. This includes research for both economic and social data which includes strategies for the next iteration of the Canadian Census of the population.

[1] The views expressed here are those of the author and do not necessarily represent those of Statistics Canada

[2] Government of Canada, *Report to the Clerk of the Privy Council: A Data Strategy Roadmap for the Federal Public Service*, < https://www.canada.ca/en/privy-council/corporate/clerk/publications/data-strategy.html>.

[3] Statistics Canada, *2017/2018 Year in Review: Modernization in motion*, accessed September 2019, <https://www.statcan.gc.ca/eng/about/yir2018>

3. As we adopt and develop new access solutions, it seems that the confidentiality strategies must be more varied and specific to the details of the survey and how the information will be accessed. There is no one-size-fits-all access solution and therefore there is no one-size-fits-all confidentiality strategy. From a Five Safes Framework[4] perspective, it is clear that the confidentiality or statistical disclosure control strategy for 'safe outputs' or 'safe data' will heavily depend on the type of data and the other mechanisms put in place to make the access solution safe. The challenge is to develop new access solutions that take existing constraints into consideration. These challenges will be described below.

4. From our experience, the challenges in accessing economic versus social data are very different because of the risk scenarios associated with the two types of information. As well, the Census provides special confidentiality considerations because of its visibility and its importance. The initiatives listed below are broken down into social, economic and census. Confidentiality strategies from both a data perspective and an output perspective will be discussed. Achievements will be highlighted, research areas discussed and challenges moving forward will be presented.

3. Economic Data Access Solutions

3.1. Perturbation alternative for tabular data

5. One of the major achievements at Statistics Canada in the field of economic statistics was the first publication of survey estimates using a cell perturbation technique as an alternative to the standard cell suppression techniques. In 2019, estimates for the 2017 Survey of Innovation and Business Strategy[5] (SIBS) were released using a Random Tabular Adjustment (RTA) methodology developed by Mark Stinner[6]. Releasing information this way was not easy. Similar methods like Controlled Tabular Adjustment were considered in the past but because of both technical and practical reasons these methods were not implemented. An review of activities of other statistical offices suggests that there are limited examples of perturbation techniques being used for a variety of reasons including:

1. Perceived disclosure – These methods will allow the publication of cells even in the extreme case of having a single contributor. Misinterpreting the estimate could be perceived as a disclosure of a reported value.
2. Noise by design – Many business surveys depend on producing very precise estimates with little error. Cross-validation and calibration ensure that estimates are consistent

[4] The Five Safes Framework <www.fivesafes.org>

[5] Statistics Canada, *Survey of Innovation and Business Strategy*, <https://www150.statcan.gc.ca/n1/daily-quotidien/190326/dq190326b-eng.htm>

[6] Stinner, M. (2017). *Disclosure Control and Random Tabular Adjustment*. Proceeedings of the Survey Methods Section, SSC Annual Meeting. Winnipeg: Statistical Society of Canada.

with other data sources. Adding noise will undo some of the work put in place to improve the quality of the estimates.

3. Trend analysis – Similar to point 2, noise is generally added in an independent fashion between survey occurrences which can greatly affect trend estimates.
4. Reliance on quality measures – The addition of noise on the survey estimate is the assurance that individual contributions are protected. Reporting of quality indicators is integral to demonstrating that disclosure control methods are in place and ensure there is no perceived disclosure. It is not clear how to calculate these measures for some perturbation techniques.

6. The SIBS survey was a great opportunity to apply a perturbation technique. One of the main advantages of this survey was the lack of historical ties to a dissemination approach. Users were accepting of the added noise in favour of having estimates for cells that would have been typically suppressed. They were also not interested in a picture of trends as much as levels. To help in the interpretation of the estimates several communication pieces were written including a note to users with the release and a report on the StatCan Blog[7]. The blog helped with the statistical capacity building of the data users to ensure a proper interpretation of the estimates. It also helped to ensure that the survey respondents understood that their information was protected.

7. With the success of this initiative, the next step is the ambitious goal of integrating RTA in the development of a zero suppression approach for economic surveys. The RTA method may not be applicable to all surveys and other solutions will be proposed when the RTA method is not appropriate. Adopting a new dissemination strategy which relies on perturbation rather than suppression is somewhat simple from a statistical perspective but is a challenge for the subject matter areas that have to analyse and defend the estimates. The reasons mainly surround the ideas presented above. In some cases, the move to this technique involves a complete rethink of the dissemination approach which requires consultation with the users of the information to ensure that we continue to meet their requirements.

8. In order to broaden its application, one of the main research areas is to consider correlated noise which would allow the application on trend estimates and other situations where there are variables where the relationships should be maintained. Another challenge is the development of a simplified, more user-friendly version of the software that will allow researchers to easily apply the method on their own.

3.2. ‘Safe data’ for Economic Surveys

9. Access to economic microdata is an attractive option for researchers. However, economic data is very sensitive and presents a high risk of identification by its very nature. Some of the access options for microdata have been dismissed in the past because of this detail. Policies, communication strategies and trust frameworks will have to be developed to encourage this type of access.

---

[7] Statistics Canada, March 26, 2019, *Random Tabular Adjustment is here!*, <www.statcan.gc.ca/eng/blog/cs/rta>

10. Synthetic data for social surveys has been produced by some programs for several years. These data access solutions have traditionally been limited to 'Dummy' files that allow access to simulated data that is non-disclosive but offers limited utility. More recently, higher utility data has been produced for social data and the success is described below. Ideally, some of the same methods could be applied in an economic data context. Statistics Canada is investigating the application of the same methods used in the social context on business data. The application will facilitate the virtual data labs and the current business data access options such as the Canadian Center for Data Development and Economic Research (CDER)[8].

11. Another alternative are public-use microdata files. Similar to synthetic data, the files should be created in a way to prevent disclosure but instead of being randomly generated, they are often created through disclosure control methods applied directly to the original data (collapsing, suppressions, rounding, and perturbation). These have been investigated and dismissed in the past because of the risk of identification but will be reviewed as an access solution in the context of broader ideas like the virtual data lab where additional levers of safe access can be used (User accreditation for example).

12. One possible area where synthetic data or public use files could be applied when the population of interest is homogeneous, or when the size of business is not an important analytical factor, synthetic data or public use files seem especially well suited. One example could be surveys of small and medium sized businesses.

3.3. Disclosure Control Software Solutions for economic data

13. The methodology branch supports the generalized software disclosure control tool G-Confid[9]. It is a software solution that supports the evaluation of disclosure risks for continuous data such as income or revenue. It applies a suppression method to address the identified risks. It is simple to use when the tabular output is simple in nature as is the case with most outside research. Some value may be found in access solutions that put the researcher at arm's length from the microdata. A wrapper for G-Confid will be investigated along with developing a strategy that integrates the analysis from the agencies estimation system G-Est[10] with the confidentiality strategy from G-Confid. This would be useful in the access scenario where the researcher is not permitted to work with microdata or not able to run SAS routines directly.

14. The G-Confid methodology is based on the evaluation of disclosure risk associated with published totals. Most of the research conducted by economic analysts accessing CDER is

---

[8] Canadian Center for Data Development and Economic Research, <www.statcan.gc.ca/eng/CDER>

[9] G-Confid (Disclosure Avoidance – Generalized System) <https://www150.statcan.gc.ca/n1/en/catalogue/10H0109>

[10] G-Est (Estimation – Generalized System), <https://www150.statcan.gc.ca/n1/en/catalogue/10H0035>

complex in nature with analysis of ratios and model parameters. Both the methods and their application in G-Confid for multivariate analysis is under investigation.

15. GTAB is a generalized tabulation tool used internally at Statistics Canada. It offers a consistent approach for creating tabular data and applying confidentiality by regulating the process of tabulating data and creating precision measures (i.e. quality indicators). The Real Time Remote Access (RTRA)[11] system is an extension of GTAB that allows real-time analysis of microdata located in a central and secure location. There has been some research into developing a layered perturbation strategy (Tambay, 2017)[12] for GTAB that could have some application to economic data. A similar approach is under consideration by the Australian Bureau of Statistics and their TableBuilder[13] product.

16. One of the main challenges with any automated solution is the limited functionality of a system in terms of the analytical capacity available. Automation often puts a user in a situation where utility is limited to the functions included in the system. The tools available offer good solutions for descriptive statistics but fail when looking at applying statistical disclosure control rules for more complex statistics such as model parameters. User consultation will be required to understand their needs and abilities in performing disclosure assessments.

4. Social data access solutions

4.1. Synthetic Data

17. Synthetic data of high analytical utility is a challenge since as more utility is offered more risks of disclosure become apparent. Statistics Canada has successfully applied methods[14] for a Mortality data file as well as a Cancer file linked to demographic information from the Canadian Census. The data synthesis process was carried out using the R software package S*ynthpop*[15]. In

[11] The Real Time Remote Access (RTRA) system, <https://www.statcan.gc.ca/eng/rtra/rtra>

[12] Tambay, J. 2017 "A layered perturbation method for the protection of tabular ouputs" *Survey Methodology* 43: 31-40. Available at < https://www150.statcan.gc.ca/n1/pub/12-001-x/2017001/article/14818-eng.htm>

[13] Chipperfield, J. et al. 2019, "Prospects for Protecting Business Microdata when Releasing Population Totals via a Remote Server", *Journal of Official Statistics*, Vol. 35, No. 2, 2019, pp. 319–336, <http://dx.doi.org/10.2478/JOS-2019-0015>

[14] Sallier, K., Girard, C. 2018. Toward a Successful Implementation of Synthesis in a National Statistical Agency: A model for Cooperation. Proceedings of the 2018 Privacy in Statistical Databases.

[15] Nowok, B., G.M. Raab and C. Dibben. 2016. synthpop: Bespoke creation of synthetic data in R. Journal of Statistical Software 74(11), 1-26

both situations these files were created for hackathons[16][17] where participants needed access to a safe data file that allowed reasonably valid statistical inference. The data was suited for those researchers seeking to create statistical models that describe the many complex relationships existing in the original dataset. In both situations the data was fit for its use in a training / outreach situation but in both situations there were expressed interest of using the data for more than just a training exercise.

18. Statistics Canada is investigating better ways to evaluate the risk and utility with such files. The goal is to help users understand the value of such files while also understanding their limitations. The other challenge is to find quality indicators that clearly demonstrate the utility of the information derived from these files. Research will continue to help develop better synthetic data files that take issues such as survey weights and variance estimation into consideration.

19. One of the main advantages of a synthetic data file is that if properly created, it can be disseminated freely with no access concerns similar to any public-use file. This has great appeal especially when considering allowing more researchers to access the information under less secure access scenarios. One of the issues with any data source that is not the original is that it is always considered as lower quality / utility compared to the real data. Communication strategies will have to be developed to help convince users of the utility of such files and to consider synthetic data as a valuable access option.

4.2. Real Time Remote Access (RTRA)

20. RTRA offers a remote safe access option that automatically ensures safe output with no direct access to the microdata. It allows analysis of an anonymized data file with restrictions on the types of analysis that are possible for a trusted researcher. More research and development is required to expand the data available in the RTRA (42 datasets currently available). The tool is also limited in terms of the descriptive statistics available and work will be required to expand the analytical options of the system as well as expansion to administrative data solutions. This tool could also be considered as an access option with fewer constraints if used in a virtual context.

4.3. GTAB Development

21. GTAB is the engine that supports the analyses available in the RTRA solution. The methodology branch is finalizing the development of the Layered Perturbation Method (LPM) to completion which will allow a new safe output solution for users in the presence of data with some dominance. The analyses available are limited in some ways and the branch is working on

---

[16] The LIDIC hackathon: Linked Data Innovation Challenge, <https://ipdln.org/2018-conference-info/preconference/student-hackathon>

[17] Hack4Cancer Hackathon: Unleashing the power of linked cancer data, <https://www.partnershipagainstcancer.ca/news-events/event/hack4cancer/?utm_source=partner-toolkit&utm_medium=spring-summer&utm_campaign=hack4cancer>

the development of ratio functionalities and confidence intervals for weighted frequencies in the next year.

5. Automation of Simple Analyses

22. A small proportion of output vetting in the Research Data Center is dedicated to a standard approach for complex model parameters coming from regression models. Statistics Canada will investigate these approaches and ensure that they are appropriate with the goal to develop simple automated tools that can aid the RDC analyst or researcher with the disclosure risk evaluation. The value of such a tool will have to be evaluated if it is limited to only certain basic types of analyses.

6. Census Access Solutions

6.1. GTAB for the Census – G-Spec

23. For the 2021 Census, Statistics Canada is looking to improve protection measures with the anticipation of increased statistics and functionalities that are available along with an anticipated increased level of access to the data for users. Statistics Canada is developing a new tabulation tool, called G-Spec, which is used for producing estimates and applying confidentiality measures. The intent is to replace the previous system Computer Assisted Product Specification System (CAPSS) that has been used for well over 20 years. This new tool will give greater functionalities and options to produce more statistics and precision measures than before.

24. Furthermore, we anticipate increased access to Census data to the public through less controlled environments such as an online tabulation tool. As a result of this increased level of data available for users we have implemented new protection methods for new variables, statistics and precision measures. We have implemented perturbative methods such as a new rounding tool for statistics called OptRound which is available in G-Confid. The layered perturbation method is also being considered to allow for more statistics to be released rather than applying suppression measures. Other perturbative methods such as consistent random rounding for counts are also being used to strengthen areas of concern for past cycles.

6.2. Expanded RTRA

25. The changes and adaptions to GTAB will have to be applied in RTRA if we are to promote this access option to Census researchers. The GUI will have to be enhanced and tested with a risk evaluation of allowing researcher to access information through this tool. RTRA limits access to certain levels of detail. A challenge will be increasing the risk appetite to allow more detailed research through the RTRA.

6.3. Differential Privacy

26. The methodology branch is researching this statistical concept and assessing the impact it could have if it were to be used on Statistics Canada surveys. Differential privacy assesses the risks slightly differently than Statistics Canada's standard approaches to dealing with confidentiality. To address these risks, the utility of the data is affected differently. In a data-rich society it does have benefits in properly considering the risks and utility of our publications. Before adopting this strategy, the risk-utility of such a strategy must be well understood by all. Laplace noise has been successfully applied at Statistics Canada in a specific small project that illustrates some of the advantages and challenges with the approach. One of the main challenges is understanding the risk-utility model to help in choosing the appropriate parameters for the amount of noise to be added to the estimates. From a methodological perspective, any parameter value is appropriate and is considered as a user decision on the risks they are willing to take and the desired utility of the product. However, because of the technical nature of the approach, it is difficult and challenging for anyone to pinpoint exactly what these limits should be.

7. Overall Initiatives

27. Access solutions such as the Virtual Data Lab (VDL) will allow safe access to Statistical information based on an access trust framework with researchers. The idea is to have remote access facilities where there is a shared responsibility for ensuring confidentiality of personal and business information. The methodology branch is developing and enhancing user support documentation including documentation used in the RDCs. The methods and rules put in place will be evaluated for consistency between survey programs to make them easier for the researcher to implement. Where appropriate, simple and accessible confidentiality methods will be promoted

28. With a shared responsibility framework for access, the first step is to ensure that there is capacity building for understanding the importance of confidentiality, being familiar with the methods for applying rules and being able to apply the methods with disclosure control tools. The methodology branch will be heavily involved in developing the capacity of researchers to share in the responsibility of confidentiality and is actively examining and considering outside tools as a way to achieve this goal.

8. Conclusion

Over time, Statistics Canada has developed a large array of access solutions that give safe access to analysts to help them develop the policies that help shape the country. As we move forward, this paper has shown that we cannot rest on our past achievements. We must continue to develop new ideas to expand access to new areas while at the same time addressing emerging risks to the safety of the data entrusted to us by Canadians.