WP. 6
ENGLISH ONLY

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE (UNECE)

CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION

STATISTICAL OFFICE OF THE EUROPEAN UNION (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Tarragona, Spain, 26-28 October 2011)

Topic (i): Disclosure risk assessment

Progress with Real Time Remote Access

Prepared by Michelle Simard, Statistics Canada

Progress with Real Time Remote Access

Michelle Simard[1]

[1] Statistics Canada, Household Survey Methods Division, RHC Building 18-Q, Tunney's Pasture, Ottawa, Ontario, Canada, J8N 4A5, Michelle.simard@statcan.gc.ca

Abstract. In recent years, Statistics Canada has invested much time and effort into examining ways to meet the demands of researchers. A key option that Statistics Canada has explored is the development of a Real Time Remote Access (RTRA) application. This application is essentially an on-line remote access facility that would allow researchers to run—more or less in real time—data analyses on microdata or lightly masked microdata sets kept in a central and secure location under the control and care of Statistics Canada. Although with some limitations in terms of users and statistics offered, a prototype of the RTRA application was successfully launched in Spring 2010. The paper describes the development and challenges of the RTRA project and the improvements made in its methodological aspects.

Keywords. Confidentiality, Microdata access, Statistical Disclosure Control methods

1. Introduction

The objective of the RTRA project is to build an application, allowing users to have on their desktop anonymized descriptive and analytical statistics, in real time with remote job submission. RTRA will be defined as another mode in Statistics Canada's means of releasing data, from the highly perturbed and anonymized Public Used Micro-data Files (PUMFs) to the custom tables released on the agency website to the Research Data Centres (RDC). As shown in Figure 1, the RTRA should be positioned in term of utility, to be close to the RDC, but with less difficulty of access for researchers.

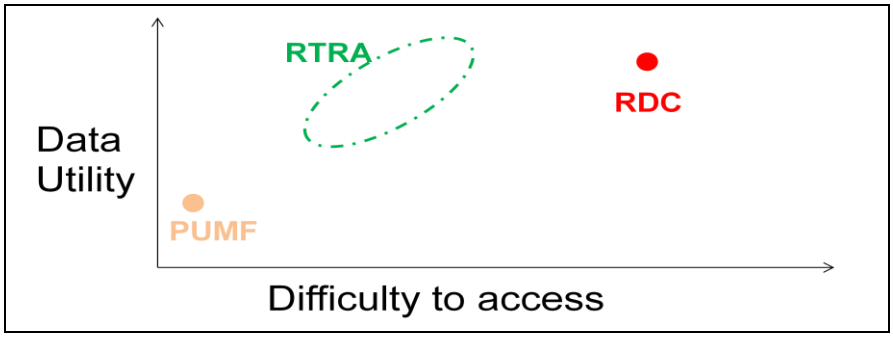

Figure 1 Position of RTRA in Statistics Canada's means of releasing data.

When fully operational, the RTRA application will provide faster access to more data and give more flexibility to researchers by providing a system that potentially gives them desktop access to data 24 hours a day, seven days a week.

There are multiple phases of development in the RTRA project. The first phase was to gather business requirements allowing the Agency to gain a deep understanding of the different components of remote access such as the security, legal and functionality requirements. This is described in Simard (2009). The second phase was to build the tool. This is described in the next section.

2. The Prototype

In the second phase, based on the business requirements, the specifications were determined for the actual development and building of a first prototype. There were two driving forces in the development of the tool: confidentiality and cost. Statistics Canada must protect the confidentiality of the data collected under the Statistics Acts. The team developed mitigation processes to manage the risk of disclosure in the RTRA. They are:

- Methodology processes through Statistical Disclosure Control (SDC) methods, e.g. controlled rounding, removal, suppression, perturbation, etc...
- System and Informatics processes through some pre-programmed or manual protection and secure infrastructure, e.g. maximums set, pre-request controls, post-requests controls.
- Management processes through proposal approval process and instructions provided to the users.
- Legal processes through terms and conditions signed by the users and the liability clause in the contract

Given that most of the funding for the project comes from external partners, and is limited, whenever possible, already-built systems were leveraged.

The first version of the Prototype has a limited number of users: researchers employed by some federal government departments. It also places certain restrictions on both the types of request that could be submitted and the level of detail of the statistical outputs. It only allows tabulations of counts for seven social surveys and with the use of SAS only.

The first time a researcher uses the system, it may take days before all the approval and legal processes are completed; however this is reduced significantly for subsequent accesses. Once the legal documents are signed and the approval process

is completed, researchers are issued a username and password that they use to link to a Statistics Canada external server through the Internet. All researchers are required to sign a contract that outlines rules and responsibilities as well as penalties and disciplinary actions in case of a breach.

The RTRA Prototype allows a researcher to submit a SAS program to a controlled SAS server that has been modified to prevent the use of particular commands and to comply with rules regarding the nature and size of the statistical outputs. The request passes through Statistics Canada IT security firewalls through the E-file transfer infrastructure (EFT), and then is screened by the system to ensure that a valid user submitted the request before running on a secure internal Statistics Canada SAS server. The data sets being used contain confidential microdata that have been lightly masked to remove sensitive variables and detailed geography. All tabular outputs are weighted and vetted for confidentiality. Following the vetting for disclosure of confidential information the tables are sent back to the researcher in the specified format.

The time it takes for a request is highly dependent of the EFT system and of course the size and number of tables requested. Currently the timeliness of a request ranges from a minimum of 4 minutes (EFT time) plus processing time to a maximum of 3 hours (EFT time) plus processing time.

3. Developing Statistical Disclosure Control Methods

There are aspects in developing a real time remote access infrastructure that are associated with a certain level of risk. As mentioned, some of the aspects relate to the systems and informatics, others relate to the legal considerations, and finally some are strictly methodological. From a methodological point of view; there is no absolute criterion for defining confidential data. However, the boundary between confidential and non-confidential data can be interpreted as the boundary between negligible and non-negligible risk. Therefore, in terms of disclosure control, Statistics Canada must apply strict rules to protect the privacy of respondents. The SDC methods are essential to any release tool; they have to be developed carefully as these are often seen as the last barriers of control. Nonetheless, one must not forget that they are part of a combination of protective controls and they must not be considered as the only method when evaluating risk.

3.1 The disclosure control methods

Based on the literature and practices, there are usually three key aspects of controls: before-the-request rules or control rules for the inputs (manual and automatic); after-

the-request rules or control rules for the outputs (manual and automatic); and perturbation methods applied on the microdata files.

Elements of the before-the-request controls

The elements listed below are the rules implemented in the first version of the prototype.

- Limited number of requests. A maximum of 10 requests per 24 hours with a maximum of 10 tables per request are set per user.
- Ensures that the programming guidelines have been followed. Scanning for appropriate codes, use of variables and programming. User support is offered.
- Use of limited available processes to be able to control output. *Note*: currently, only a modified version of SAS PROC FREQ is offered with DATA step and PROC SORT.

Requests that do not comply with the guidelines and parameters of submission will not be run by the RTRA application. To monitor such incidents, a log will be generated by the system indicating how the program did not comply with the guidelines and it will be sent back to the researcher.

Elements of the after-the-request controls

The elements listed below are the rules implemented in the first version of the prototype.

- SDC methods appropriate for the statistics being produced.
- Log and outputs checked and modified, if need be.
- All output generated during the prototype phase will be kept indefinitely for auditing purposes.

Note that the requests are run on the original micro-data files where only some geography variables have been removed. The first version of the prototype only offered tabular frequency counts for some household surveys. Thus there was one method developed for that statistic.

4. How to choose the best Statistical Disclosure Control Methods

For tabular outputs, the development and choice of disclosure methods was relatively simple. There are two basic options: rounding and suppression. The disclosure control method chosen for the tabulated frequency data is additive controlled rounding (ACROUND), a method developed for the Canadian Census described in Boudreau, Filep and Liu, (2004). The method was chosen for its simplicity of application and programming and its ability to protect against potential links to

PUMFs and multiple query submissions. It also has only a small impact on precision. More details on the ACROUND as used in the RTRA can be found in Simard (2011).

4.1 The next round of statistics

Sometime in 2011, the second version of the Prototype will be released. Additional statistics and surveys and some quality indicators are the new features of this second version. When thinking about the SDC methods for the second version, the challenge facing the team was to decide what the next best strategy was. It was agreed with the partners that the next statistics offered will be means, percentiles and proportions. With these in mind, the question then became: should the team develop, for each combination of statistic and software, a disclosure control method appropriate for each of them individually? For example, develop a method for the median in SAS, a method for a proportion in SPSS, another method for proportion in SAS, etc. This is referred to as the output control approach.

Alternatively, the team was wondering if there is a series of masking procedures that could be applied to the microdata files directly at the beginning of the process. This lightly masked file could then be used by any or some large sub-set of the analytical procedures without worrying about disclosure risk. By identifying the risky variables and the unique records and masking them in the microdata, no other vetting processes would be needed. Masking techniques include the removal of variables, top-coding, regrouping sensitive variables, data swapping, etc. This is referred to as the input control approach. The next sub-sections discuss the two approaches.

4.2 The Output Control Approach

As described in Brandt et al (2010), outputs can be classified basically by the types of statistics. In their paper, they present types of statistical outputs with an evaluation for each in terms of “safe” or “unsafe” in an RDC-type environment.

Their evaluation used for RDCs should be used as guidelines for developing the RTRA rules as these are protective processes similar in nature. However, there are two major distinctions between the two modes of release: i) in the application of the checking rules and ii) in the automation and programming of the corrective measure. In the RDC context, almost everything is done manually. In a RTRA environment, as if it was not challenging enough, rules have to be applied through automated programs. Consequently, the chosen approach for RTRA has to take into consideration the fact that the rules and the corrective measures (if any) need to be programmed relatively easily. Simple rules, such as a minimum of 10 unweighted units or a minimum number of degrees of freedom seem simple to program, but others might be difficult. Furthermore, the corrective actions can be rather complex.

First, there is a need to develop what is the appropriate uniform corrective measure and then another need to program it along all other subsequent modules of the system. One advantage of the RTRA however, is the possibility of ruling out some complex outputs right from the outset, if deemed too risky or too complex, as opposed to the RDC analysts who can basically run any program.

The approach is less risky, produces more precise statistics, but requires more resource as it is statistics and software-dependant.

4.3 The input control approach

The main issue in this approach is to determine if there is a series of perturbation techniques applied to the original micro-data set that could be sufficient to protect the confidentiality for all outputs without having to develop output control methods for each procedure.

Noise introduction and perturbation techniques are often used in business surveys. The U.S Census Bureau used the EZS noise method (Massell, Zayatz and Funk, 2006) to perturb their microdata before producing tables. This technique works well with magnitude data. In social surveys, most of the variables are not quantitative data but categorical. One strategy could be to use some noise introduction techniques such as the EZS method on the magnitude data such as the revenue, weight and height and some other technique on categorical data.

For the categorical data of an individual record, it is the combination of variables that makes it unique, thus highly risky. Some other perturbation techniques for categorical data (swapping, recoding variables, etc) could be applied once the record has been identified as unique or risky. The key here is to use a measure of disclosure risk such as the Skinner-Elliot measure for example (Skinner and Elliot, 2002) and use it intelligently in the strategy.

The rule of a minimal number of units or degrees of freedom could be difficult to automate and program. In the U.S. Microdata Analysis System, (Lucero, Singh and Zayatz, 2009) they have pre-determined universes with a pre-specified number of observations and also implemented a universe sub-sampling routine called the Drop-Q rule. It will be worth investigating the potential to include and program this rule in the RTRA system.

The approach is a little more risky, produces less precise statistics, but requires fewer resources once the strategy is developed and implemented. It is neither statistics nor software-dependant. However, it could potentially take very long before coming up with a defendable to all approach.

4.4 The chosen SDC

It became apparent that it was impossible to implement a proper input control approach without jeopardizing our delivery dates with major delays. The fact that some analysts did not like the data being modified had also some weight in the decision. Based on the guidelines of Brand et al. (2010) and some discussion with experts, the following rules were developed for the given statistics.

Maxima, minima and percentiles.

What Brand et al. suggest:

- Usually minima and maxima are not released.
- Percentiles are treated like magnitude data. The rules for these cases include a minimum of unweighted units, a group disclosure rule (i.e. no cell or group can contain more than X% of the total row or column) and a dominance rule (i.e. in a given cell, the largest contributor cannot exceed Y%). Usually X = 90 and Y = 50.

What the RTRA rules are:

- No minima will be released.
- No maxima will be released.

For percentile, the following four rules were selected.

- Release the percentile only if there is a minimal number of observations above and a minimal number of observations below the percentile value.
- Release the percentile if it is ≠ minimum or maximum value.
- Release the percentile if the total number of unweighted observations is larger than a certain number of units.
- Release the percentile if the rounded frequency associated (from ACROUND) with the percentile is ≠ 0.

Modes, indices, means, ratios and indicators.

What Brand et al. suggest:

- For modes, usually a group disclosure rule is applied.

- For means, indices and ratios, they should be derived from at least 10 units and a dominance rule is applied (see above).

These would be challenging to automate. Another approach often considered is to evaluate the complexity of the formula of the index itself. Usually an index is a summary of variables $I = f(X, n)$. The formula and the population size (n) should be

factored in. Some are so complex that it is basically impossible to figure out individual values even for only one unit. There is also some consideration if it is appropriate to publicly divulge the formula or not. Again the challenge would be in the automation of this in the system. Means and ratios are simpler formulas with only two components and are more problematic. One approach often used is to release if all the components can be released. However smart or synchronized perturbation techniques should be applied on both components to control the change and minimize the impact on the precision of these two statistics.

What the RTRA rules are:
For means, the following 2 rules were selected:

- Release the mean only if there are a minimal number of observations present in the domain.
- Release the mean if the rounded frequency associated with the mean (from ACROUND) is $\neq 0$.

Note that the real parameters for the RTRA rules will not be made public.

Modes, indices and indicators have not been identified so far by the partners as key statistics. SDC rules for proportions and ratios are currently being developed.

When developing the rules for both statistics, it became evident that the rules had to be developed with the ACROUND output in mind, i.e. the anonymized final count for the same domain. If one user requests a table of counts for a given domain in one submission and in another submission, the user requests for the same domain the mean, the two outputs have to be consistent. For example, for a given cell, if the ACROUND algorithm sets the count to zero, the "MEANS" rules should not provide a mean for that cell. Consequently for every tabulated statistic, when applicable, the ACROUND program will be run and the counts will be produced in parallel and actually used in the output of the requested statistics.

4.5 Balancing precision, confidentiality and data quality

To make the exercise a little more complex, developing the proper SDC also has to be balanced with not compromising too much the precision of the estimates produced, i.e., the difference between the true estimates and the released one. Moreover, the data quality indicators associated with the statistics, whether it is the variance and/or the Standard Error (SE) and/or the Coefficient of Variation (CV) and/or Confidence Interval (CI) must not provide additional information that could be useful to identify respondents, but still provide an indication of the quality of the estimates.

For the second version of the Prototype, a CI along with the SE will accompany the estimates. Furthermore an indication of quality derived by the CV calculated with the real data will also accompany the statistic. A possible quality indicator range based on the CV is provided in Table 1 and a potential output table is provided in Table 2. Note that these are only for illustrative purposes only. The final parameters are not yet finalised and will not be made public.

Value of CV	Quality Indicator	Guideline
$0 \leq$ C.V. ≤ 0.20	(a)	very good
$0.20 <$ C.V. ≤ 0.40	(b)	acceptable
$0.40 <$ C.V. ≤ 0.50	(c)	marginal
C.V. > 0.50	(d)	very poor
Suppressed/ Unreleasable	(z)	

Table 1. Example of a derived Quality Indicators.

Gender \ Vision	Bad	Good	Excellent
Male	300 (b), S.E = 48 **INCOME** P50 = 17,600 (c) S.E. = 4,550	**0** (.), S.E. = n/a **INCOME** **P50 = X** **S.E. = X**	100 (d), S.E. = 3.4 **INCOME** **P50 = X** **S.E. = X**
Female	0 (.), S.E. = n/a **INCOME** **P50 = X** **S.E. = X**	400 (c), S.E. = 131 **INCOME** P50 = 0 (z) S.E. = 55.8	400 (c), S.E. = 118 **INCOME** P50 = 0 (z) S.E. = 0

Table 2. Example of an output table for median.

5. Conclusion

The RTRA project continues its development on different fronts: more statistics, more SDC, more surveys and more users. The next phase will allow access to the academic community. Censuses and administrative data are also on the radar for future versions. On the SDC front, until now, the output control methods approach

was the preferred one, but input control methods may be re-evaluated when more complex analytical statistics such as linear regression will be made available.

Acknowledgement

The author would like to thank David Price, Martin Lessard and Patrick Gallifa for their contributions and their tireless work toward the project. She also would like to thank Jack Gambino, Jean-Louis Tambay and Pierre Caron for their advices and editorial comments.

References

Boudreau, J.-R., Filep, K. and Liu, L.(2004). *Iterative rounding for large frequency tables.* Proceedings of the American Statistical Association, Joint Statistical Meeting. Toronto, 2004.

Brandt, M., Franconi, L., Guerke, C., Hundepool, A., Lucarelli, M. Mol, j., Ritchie, F., Seri, G. and Welpton, R. (2010) *Guidelines for the checking of output based on microdata research 2010*, published; with NSIs of UK Germany, Netherlands, Italy

Lucero, J., Singh, L., and Zayatz, L. (2009), *The Current State of the Microdata Analysis System at the Census Bureau*, Proceedings of the American Statistical Association, Joint Statistical Meeting. Washington, 2009.

Massell, P., Zayatz, L., and Funk, J. (2006), *Protecting the Confidentiality of Survey Tabular Data by Adding Noise to the Underlying Microdata: Application to the Commodity Flow Survey, Privacy in Statistical Databases, CENEX-SDS Project International Conference*, PSD 2006, Proceedings, Lecture Notes in Computer Science (LNCS) 4302, Springer 2006, ISBN 3-540-49330-1.

Simard, M. (2009). *Development of a Real Time Remote Access Infrastructure at Statistics Canada.* Paper presented at the UNECE/EuroStat workshop on data confidentiality in Bilbao, Spain, in December 2009.

Simard, M. (2011). *Real Time Remote Access at Statistics Canada: development, challenges and issues.* Proceedings of the International Statistical Institute Conference, Dublin, August 2011.

Skinner, C. J. and Elliot, M.J., (2002), *A measure of disclosure risk for microdata*
J. R. Statist. Soc. B, 64, Part 4, pp. 855–867.