# Real Time Remote Access at Statistics Canada: development, challenges and issues.

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

## 1. Introduction

Access and analysis are key components of the mandate of Statistics Canada. By providing access to its statistical outputs to researchers, the Agency's data continue to make an immeasurable and vital contribution to public policy debates in Canada that benefit many of the nation's citizens and businesses. Statistics Canada currently provides access to its statistical data through many longstanding means: aggregate data are posted on the Agency's Web site, public use microdata files (PUMF) have been produced for many years, as have special and customized tabulations of aggregate data. More recently the creation of over 24 Research Data Centers (RDCs) has permitted access to confidential microdata files to researchers across the country. Although these RDCs are located in universities, they operate as satellite offices and provide researchers with controlled and secured access to Statistics Canada microdata for purposes of statistical research.

Like many other national statistical organizations, Statistics Canada has faced increasing demands from the research community for even greater access to detailed microdata. This greater interest reflects the importance of increasing access both in the development and analysis of government policies and programs. The interest extends not only to academia and policy makers within the country but also to the international research community for cross-national studies conducted by the World Bank, World Health Organization, etc. Furthermore, ongoing advances in informatics technology have significantly opened up avenues for statistical agencies to produce and disseminate detailed data and for researchers to conduct data-mining and analytical studies.

In recent years, Statistics Canada has invested much time and effort into examining ways to meet the demands of researchers while, at the same time, ensuring that the legislative requirement under the Statistics Act[1] to protect the confidentiality of respondent data is met. A key option that Statistics Canada has decided to explore 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.

While researchers appreciate the access to the microdata in the RDC, many find the administrative process of submitting research proposals, the inconvenience of travelling to the

[1] Canada's *Statistics Act* does not allow confidential data to be provided to anyone who is not a Statistics Canada employee (or "deemed" employee), except under the provisions of Sections 11, 12 or 17(2) of the *Act*. Sections 11 and 12 cover the sharing of information with provincial statistical agencies and other organizations, respectively. Section 17(2) allows certain types of release of identifiable individual data by the Chief Statistician.

centers and the restricted hours of operation cumbersome. The RTRA is a more transparent process for accessing StatCan information. In theory, when fully operational, the RTRA application would 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.

Each access mode has its benefits and associated risks. The challenge in the development of the tool is to be able to implement methods that will manage the risks while providing benefits to the users. Table 1 presents both risks and benefits for some access modes.

***Table 1. Benefits and risks of access modes***

| Mode of access | Benefits | Level of Disclosure Risk |
|---|---|---|
| PUMFs | Users can obtain and use microdata anywhere without restrictions (except licensing) | Very small risk of disclosure.<br><br>Very small risk from detailed cross-tabulations or linking with other files |
| Research Data Centres (RDC) | Users have access to detailed confidential microdata | Increase in risk of disclosure, though remains small.<br><br>Small increase in risk as users see confidential data plus residual disclosure from multiple submissions.<br><br>The risk is also subject to increase with the sensitivity of the data such as business microdata. |
| Real Time Remote Access | Users have access in real time to detailed microdata from their desktop which have been automatically vetted for non-disclosure | Larger risk of disclosure.<br><br>Increase in risk from inadvertent slips in automatic vetting and from residual disclosure with multiple submissions vetted automatically.<br><br>The risk would increase to unacceptable levels (for the moment) should users have desktop access without automatic vetting or adequate security. |

The next section describes the thinking behind the development of the RTRA and presents the different developmental phases. Section 3 discusses the challenges of the current version focusing on the disclosure control methods, section 4 discusses future challenges and issues and section 5 presents the conclusion.

## 2. Development

The objective of the RTRA project is to deliver, within three to five years, a full RTRA program with remote job submission for modeling and intricate programming as well as a tabulation tool for metadata driven descriptive statistics. RTRA will be seen as another mode in Statistics Canada's spectrum of means of releasing data, from the highly protected Public Used Micro-data Files (PUMFs) to the custom tables released on the agency website to the RDCs. There are multiple phases to the development. The first phase of development 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 as described in Simard (2009). This phase involved learning from the experiences of other national statistical agencies in order to establish an overall direction for the project. A working group examined current Real Time Remote Access facilities, among them; Statistics Netherlands, the Australia Bureau of Statistics, the Office of National Statistics (UK) and the Institut de la Statistique du Québec. This

phase involved determining the key requirements of a Statistics Canada model, i.e. the basic features, procedures and governance of the program. These were identified as follows:

- Identifying the scope: descriptive statistics and modelling for survey and administrative data;
- Defining an approach, i.e., a gradual implementation with the construction of a prototype using the software Statistical Analysis System(SAS) that will be tested in conjunction with limited partners (other federal government departments); expansion would follow after a careful evaluation of risks;
- Developing the process model, that is to say determining access to the system. This step would involve examination of informatics as well as contractual and legal aspects;
- Developing the governance model for the application;
- Ensuring proper organizational implementation: privacy impact assessment and risk and threat assessment, communication/marketing plan for the launch of the system. Establish partnerships for the development of the RTRA and determine infrastructure costs to accommodate the RTRA system.

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. There are:

- Methodology processes through disclosure control methods, e.g. controlled rounding, removal, suppression, perturbation.
- 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

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

As for cost, this project is not funded by Statistics Canada; it is a cost-recoverable project, i.e., the federal partners are financing a large part of the development. Costs must be kept to a minimum. One of the results of this situation is the fact that there will be maximum recycling of current infrastructure and systems to keep the cost of development to a minimum. Trade-offs had to be made between building the ideal tool and having to reuse current systems.

The second phase was the building of a prototype; the RTRA Pilot, which was operational in the spring of 2010. The initial version of the Pilot had a limited audience: researchers employed by some federal government departments. It also placed certain restrictions on both the type of requests that could be submitted and the level of detail of the statistical outputs. It only allowed tabulations of totals for 7 social surveys and with the use of SAS.

How the current version of RTRA Pilot currently works

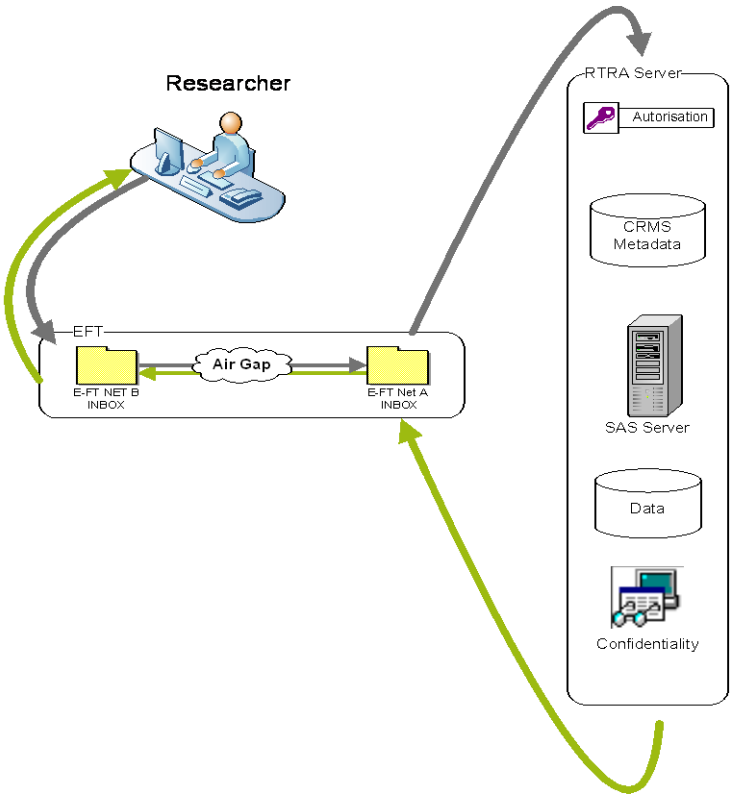
Once the legal documents are signed and the approval process is completed, researchers are issued a username and password that they use to link to a Statistics Canada external server through the Internet. All researchers are required to sign a contract that outlines rules and responsibilities as well as penalties and disciplinary actions.

Only a limited number of researchers who work as employees in policy departments within the federal government, namely Human Resources and Skills Development Canada and the Department of Indian and Northern Affairs have been accepted as researchers to work with the RTRA Pilot. Each researcher takes part in a brief RTRA training session to familiarise themselves with the system and understand the implications of accessing the data remotely. Currently, the researchers do not pay to have access as they are the financing partners. However new cost structure and penalties and disciplinary actions framework will be put in place for the next wave of users (probably the academic communities) who will be different than the first wave of users.

As shown in diagram 1, the RTRA system 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), 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.

Currently the timeliness of a request ranges from a minimum of 4 minutes plus processing time to a maximum of 3 hours plus processing time. The team is working on reducing the maximum time for both processes. For the prototype, users have to wait an average of 30 minutes.

***Diagram 1. The RTRA process.***



## 3. The Challenges

There are numerous challenges in developing a real time remote access infrastructure. Some of them relate to the informatics: programming rules and solutions, scanning algorithms and security requirements around the transit and housing of the data. Others relate to the legal considerations: approval processes and developing appropriate penalty and disciplinary actions. However the biggest challenge remains the protection of the respondents' data. The next section presents the current methods and a discussion on the future ones.

### 3.1 Developing Disclosure Control Methods

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 threshold between negligible and non-negligible risk. Therefore, in terms of disclosure control, Statistics Canada must apply numerous risk management practices to safeguard the confidentiality of respondents. It is the same approach that was taken in the development of the RTRA application. It is necessary to develop numerous rules and guidelines for disclosure control.

Based on the literature and the practices of other statistical organizations, there are usually three key aspects: pre-request or control rules for the inputs (manual and automatic); post-request or control rules for the outputs (manual and automatic); and perturbation methods applied on the microdata files.

**Elements of the pre-request controls**

- 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 some 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 post-requests controls**

- Disclosure methods appropriate for the statistics being produced. Note: currently, only tabular outputs of totals are offered.
- Two formats for output (SAS and HTML)
- 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 additive controlled rounding method**

The disclosure control method chosen for the tabulated data is additive controlled rounding (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 among other properties. It also has only a small impact on precision. As demonstrated in Diagram 2, the first step is to create a rounded additive table close to the original table with a controlled grand total called a semi-controlled table. The second step is to use an iterative process to improve the semi-controlled result to get the controlled results. In this example, the 35 is semi-controlled based on the original count and it is set to 40 and subsequently the row is reset.

***Diagram 2. An example of the method.***

Original table

| | | | |
|---|---|---|---|
| 14 | 14 | 3 | 31 |
| 15 | 3 | 22 | 40 |
| 29 | 17 | 25 | 71 |

Semi-controlled

| | | | |
|---|---|---|---|
| 15 | 15 | 5 | 35 |
| 15 | 0 | 20 | (35) |
| 30 | 15 | 25 | 70 |

Controlled

| | | | |
|---|---|---|---|
| 15 | 15 | 5 | 35 |
| 15 | 5 | 20 | 40 |
| 30 | 20 | 25 | 75 |

The challenge here is to keep this process in an acceptable time frame so that the term —real-time" true. There is a maximum of 1,000,000 iterations allowed in the controlled rounding part of the algorithm. The larger the number of initial constraints (number of cells and margins) the longer the program runs. In some of our testing, it took about 3 hours to process one million iterations. The number of initial constraints grows exponentially with the number of dimensions. For a table $a_1 \times a_2 \times ... a_n$ with $n$ dimensions, where $1 \leq n \leq 5$, the number of constraints is:

$$(a_1+1)\times(a_2+1)\times...(a_n+1) \times \left[\frac{1}{(a_1+1)}+\frac{1}{(a_2+1)}+...\frac{1}{(a_n+1)}\right]$$

For example, for a table of 3 dimensions with 3 categories each: 3x3x3, we have 48 initial constraints and for a 6 x 8 x 4 x 2 x 3, 5 dimensions, we have 3,921 initial constraints.

Reaching 1,000,000 iterations does not always result in a better table. In tests, the smallest number of uncontrolled constraints at the end of the iteration step occurred at the 100,000 iteration limit. Another result was the fact that the choice of rounding-base affects the number of iterations required to find a solution in an unpredictable way. The rounding-base is survey-driven.

To reduce the processing time, while still protecting the confidentiality, it was decided to pre-determine the number of iterations to apply to a table as a function of its initial number of constraints. Consequently tables having few constraints can be processed more quickly and receive more iterations, resulting in more additive final tables. Tables with many constraints should be processed with a smaller number of iterations to be run more quickly. Keeping in mind that a maximum of 10 tables should be executed within about 90 minutes in RTRA, rules were suggested and are presented in Table 2. There is also a maximum of 5 dimensions for a given table that was put in place for the prototype.

***Table 2. Rules for execution time***

| Range of # of constraints | Recommended # of iterations |
|---|---|
| < 50,000 | 10,000,000 |
| 50,000 – 99,999 | 2,000,000 |
| 100,000 – 199,999 | 200,000 |
| ≥ 200,000 | 100,000 |

Another challenge for the development of the disclosure methods for the pilot was to come up with some common rules that could be applied to the surveys involved. At this point, no standardised or uniform method is applied for all surveys. Each survey program has a set of disclosure rules on their own. However, the managers agreed to use the RTRA rules for the pilots which are, along with the additive controlled rounding:

- The smallest level of geography available is the province.
- Only rounded weighted counts are available.
- All logs are scanned and modified for protection. It is impossible to determine the number of observations through them.
- No more than 500 different categories can be produced for a given variable.
- In the rare case where a user would use a PROC FREQ on a (dollar) amount, an adaptive rounding technique is applied on the amount shown in the final table.

**The adaptive rounding method**

Again, this method was chosen for its simplicity of application and programming and for its ability to protect. The rule rounds the amount to a particular digit based on the value and also on the number of decimals required. Table 3 presents an example of the adaptive rounding method.

These rules and methods were developed for the second version of the Pilot system and for tabular output only. The third phase will be the development and the testing of the complete functionality of the application in terms of more statistics, more quality indicators, more outputs and more surveys.

***Table 3. Example of the adaptive rounding method***

| Original value | Decimals (max=5) | Output value (always released as a num type) |
|---|---|---|
| 123 | 0 | 120 |
| 1,234 | 0 | 1,230 |
| 12,345 | 0 | 12,300 |
| 123,456 | 0 | 123,500 |
| 1,234,567 | 0 | 1,235,000 |
| 12,345,678 | 0 | 12,346,000 |
| 123,456,789 | 0 | 123,457,000 |

## 4. Future Methodological Challenges and Issues

For tabular outputs, the development and choice of disclosure methods was relatively simple. There were two basics options: rounding or suppression. As the RTRA is entering its third phase, the challenge facing the team is to decide what the next best strategy is. Should the team develop, for each combination of statistic (simple and complex) and program language, 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 graphs in SAS, another

one for odds ratio, etc. This is referred to as the output control approach. This strategy seems burdensome and inefficient, but less risky and with the minimal number of changes to the original data files. Or, the team is wondering if there is a series of masking procedures that could be applied to the microdata files directly in the beginning of the process. This 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, swapping, etc.

### 4.1 The Output Control Approach

As described in Brandt et al (2010), outputs can be classified basically by the types of statistics. Table 4 presents their classification and their evaluation in terms of ~~"s~~afe" or ~~"u~~nsafe" in an RDC-type environment. Their evaluation of ~~"s~~afe vs. unsafe" is to provide a guideline on the likelihood of an output being released.

The evaluation used, whether for RDCs or for remote access processes, is similar in nature. However, there are two major distinctions: 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, if it was not challenging enough, it has to be done 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 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 to program it along all other subsequent modules of the system. One advantage of the RTRA however, is the ruling out of some complex outputs from the outset, if deemed too risky or too complex. RDC analysts can basically run any program.

***Table 4. Classification of output***

| Type of Statistics | Type of Output | Evaluation |
|---|---|---|
| Descriptive Statistics | Frequency tables | Unsafe |
| | Magnitude tables | Unsafe |
| | Maxima, minima and percentiles (incl. median) | Unsafe |
| | Mode | Safe |
| | Means, indices, ratios, indicators | Unsafe |
| | Concentration ratios | Safe |
| | Higher moments of distributions (incl. variance, co-variance, kurtosis, skewness) | Safe |
| | Graphs: pictoral representation of actual data | Unsafe |
| | | |
| Correlation and Regression Analysis | Linear regression | Safe |
| | Non-linear regression | Safe |
| | Estimation residuals | Unsafe |
| | Summary and test statistics from estimates ($R^2$, $\chi 2$) | Safe |
| | Correlation coefficient | Safe |

### 4.2 Some Early Issues

**Maxima, minima and percentiles.**
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.

**Mode, indices, means, ratios and indicators.**
For modes, usually a group disclosure rule is applied. For means, indices, ratios and indicators, 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.

**Linear regression coefficients.**
Usually, coefficients are safe provided at least one estimated coefficient is withheld (e.g. intercept). The other rules often used are: must be based on a minimum of 10 degrees of freedom; not solely based on categorical variables; and the coefficients are not based on only one unit. One example of this would be a time series on one company. This could be relatively simple to program and automate and a note to users should be prepared to be used as guidelines before using the RTRA.

This provides just a glimpse of the complexity of disclosure control methods in the output control approach in a remote access environment. More investigation and tests have to be done if this approach is chosen. Simard (2011) discusses this issue more.

### 4.3 The input control approach

The main issue in this approach is to determine if there is a series of perturbation techniques applied on 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 numerical magnitude 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 (MAS), 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. Details can be found in Lucero, Singh and Zayatz (2009).

## 5. Conclusion

The RTRA team has developed a system in a very short period of time and have faced challenges throughout the development. System-wise, informatics, legal and methodological challenges were overcome during the first phases of development. Phase 3 may be the biggest challenge so far in terms of methodology and appropriate use of disclosure control techniques. However, the benefits for the users and partners of being able to have access to more data outweigh any challenges. The team is confident in meeting all of them.

## 6. References

Boudreau, J.-R., Filep, K. and Liu, L.(2004). *Iterative rounding for large frequency tables*. Proceedings of the ASA 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, Government Statistics Section.

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). Development of the Real Time Remote Access. Internal document.

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.