

Disclosure Control and Random Tabular Adjustment

Mark Stinner

Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, ON, CA, K1A 0T6
mark.stinner@canada.ca

Proceedings of the 2018 Federal Committee on Statistical Methodology (FCSM) Research Conference

Abstract

Statistical agencies are interested in publishing useful statistical data but doing so may lead to the disclosure of individuals' private data. This is a problem as it leads to a trade-off between the utility of the published data and the risk of disclosure of confidential data. Disclosure control can be seen as the use of methods to deal with this problem by assessing and controlling the risk of disclosing confidential data while also providing researchers with useful statistical data.

In this paper a disclosure control model based largely on Bayesian decision theory is described as well as a method of disclosure control called Random Tabular Adjustment (RTA). This method controls the risk of disclosure by randomly adjusting the data and is intended to be an alternative to the common practice of suppressing cells at many statistical agencies. It fits naturally into the disclosure control model described. Comparisons with cell suppression and applications to real survey data are described.

1 Introduction

Statistical agencies usually collect the private data of individuals under the requirement that this data will not be disclosed while at the same time publishing as much statistically useful data as possible. When the published data can be used to disclose the private data of an individual, these goals come into conflict. To resolve this conflict the data must be altered in some manner before it is published so that the risk of disclosure is controlled.

In this paper a model is proposed that attempts to formalize this situation. To model the effectiveness of the disclosure control, the users of the published data and their targets are considered. Some users are interested in inferring general features of the population but some users are interested in inferring confidential attributes of the individuals who provided the data. These users are called analysts and attackers respectively. For the statistical agency, the published data is useful, if analysts can make sufficiently good inferences about the population, and safe, if attackers cannot make good inferences about the providers of the data. To assess how useful and safe the published data is, models of the knowledge that the users have of their targets before and after publication and measures of the uncertainty that users have in making inferences are selected. This fits naturally into a Bayesian decision theoretic framework and this allows uncertainties and other quantities in the model to be expressed in terms of familiar statistical quantities such as expectations and variances.

The proposed model for disclosure control and the associated disclosure control problem are described and then used to formulate and solve a simplified disclosure control problem. The general model and problem are introduced in two parts. The first part introduces a high level model, called the *basic disclosure control model*, that is sufficiently abstract that it describes most disclosure control models. The associated problem, called the *basic disclosure control problem*, is then described. The second part introduces a more detailed model, called the *general disclosure control model*, that fills in the details of the basic model using concepts from Bayesian decision theory. The associated problem, called the *general disclosure control problem*, is also described. A simplified disclosure control model, called the *simple Random Tabular Adjustment (RTA) model*, is then described and its associated problem solved. This problem is solved analytically and so the solution provides a simple function of the input parameters which can be evaluated without the use of any complex numerical algorithm.

Many different models and methods of disclosure control have been proposed. A good overview is found in Willenborg and De Waal (2001). Bayesian decision theory has been used in disclosure control before. The disclosure control model presented here builds on similar approaches that can be found in Duncan and Lambert (1986) and Fienberg and Trottini (2002). The proposed model in this paper both simplifies and extends this previous work.

The RTA model was used to assess and control the risk of disclosure on data from Statistics Canada's Monthly Retail-Trade Survey (MRTS). MRTS currently uses suppression as its method of disclosure control. Parameters were chosen for RTA which provide a similar level of protection to that of the current suppression methodology and a comparison of the results was made.

2 Disclosure Control

Disclosure control can be seen as the use of methods of assessing and controlling the risk of disclosing confidential data while also providing researchers with useful information when publishing data. To do this an appropriate model and an appropriate formulation of the problem are needed.

2.1 Basic Disclosure Control Model

The basic disclosure control model formalizes the above description of disclosure control. A *disclosure control method* is selected to change the original data into altered data. A measure of the usefulness of the altered data, called the *utility*, is selected and a measure of the risk of disclosure of the altered data, simply called the *risk*, is also selected. The disclosure control method depends on a disclosure control parameter and this parameter is chosen so that the method provides sufficient control of the risk of disclosure while maximizing the utility of the altered data that results from using the method.

2.1.1 Disclosure Control Method

2.1.1.1 Definition (Disclosure Control Method). *A disclosure control method is a function that takes the original data and yields the altered data. The method depends on a disclosure control parameter φ .*

There is a large variety of disclosure control methods. Disclosure control methods may be deterministic or probabilistic and may involve restriction or perturbation of the data. Suppression is a common method for tabular magnitude data and rounding is common for frequency data. Suppression restricts while rounding perturbs the data. Both of these are deterministic. Data swapping and the addition of random noise are common probabilistic methods.

The disclosure control parameter controls how the original data is altered and depends on the method used. For example if rounding is used, a rounding base needs to be selected, and so the disclosure control parameter is the rounding base in this case. If random noise is added, a distribution for the noise needs to be selected and so the disclosure control parameter is the distribution of the noise.

2.1.2 Utility and Risk

2.1.2.1 Definition (Utility and Risk). *The utility U and risk R of the altered data are functions of the altered data and depend on the disclosure control parameter φ . These functions measure the usefulness and disclosure risk of the altered data respectively.*

As with disclosure control methods there is a large variety of utility and risk measures. If suppression is used as the method on tabular magnitude data, the value of the published cells in a table is often used as a measure of utility and risk is measured using the lengths of the feasible intervals of the suppressed cells. If rounding is used, the distance of the rounded table from the original is often used as a measure of utility and the rounding base may be used to measure the risk.

2.2 Basic Disclosure Control Problem

Statistical agencies typically want to find the method that maximizes utility while constraining risk to an acceptable level. Using the above concepts we can formulate the basic disclosure control problem as an optimization problem.

2.2.0.2 Problem (Basic Disclosure Control). *Find the disclosure control parameter φ that maximizes $U(\varphi)$ under the constraint $R(\varphi) \leq r$ where r is the selected risk threshold.*

2.3 General Disclosure Control Model

The utility and risk of the altered data in the basic disclosure control model need to be described in more detail before any practical use can be made of this model. To do this we need models of who is using the data, what they are estimating and how well they are estimating it. The general disclosure control model is introduced and this model includes the concepts of *target* and *user*, *prior*, *posterior* and *base* distribution, *loss* and *uncertainty*. These concepts are formalized using concepts from Bayesian decision theory.

2.3.1 Targets and Users

2.3.1.1 Definition (Target). *A target is a function of the original data. Targets may be confidential or analytical.*

The values of analytical targets are to be made available and a publication that reveals these values contributes to utility. Conversely the values of confidential targets are to be made unavailable and a publication that reveals these values contributes to risk. A typical analytical target is a population parameter such as a mean or total and a typical confidential target is the contribution to the original data of a single individual.

2.3.1.2 Definition (User). *A user is an individual who estimates a target using the published data. A user who estimates an analytical target is an analyst. A user who estimates a confidential target is an attacker.*

Different users may know different things about their targets and the data. This may contribute to how well they can estimate their targets and so should be accounted for in the model. Users may even be data providers and so know their own contribution to the original data.

2.3.2 Prior, Posterior and Base Distributions

For each target and user we have a model representing the user's knowledge of the target and the altered data. From this the user's knowledge of the target after the data is published can be determined. The user's knowledge of the target and altered data is represented by a probability distribution in this model. In addition for each target a probability distribution is selected to be used as a benchmark against which the user's knowledge of the target is compared.

2.3.2.1 Definition (Prior Distribution). *The user's knowledge of the target and the altered data before the altered data is published is represented by the user's probability distribution of the target and the altered data. This distribution is the prior distribution.*

2.3.2.2 Definition (Posterior Distribution). *The user's knowledge of the target after the altered data is published is represented by the user's probability distribution of the target given the altered data. This distribution is the posterior distribution.*

2.3.2.3 Definition (Base Distribution). *The user's knowledge of the target after the altered data is published is compared to a selected distribution. This distribution is the base distribution.*

2.3.3 Loss and Uncertainty

A measure of how well the users can estimate their targets is needed. This is formalized by describing the user's estimation problem as a problem from decision theory. The concepts of *loss* and *uncertainty* provide a way to do this. Using the expected loss to measure uncertainty in estimation is done in other contexts (see DeGroot (1962)). The following descriptions make use of a target variable A , an observation variable B and a loss function f . The descriptions also make use of the estimator Θ and the uncertainty Ψ . Also note that in what follows E and V are the expectation and variance operators.

2.3.3.1 Definition (Loss Function). *The loss function f is the function that determines the measure of error or cost $f(a, \theta)$ of a user in using an estimate θ for a target value a .*

Before making an observation it is assumed that the user selects the estimator that minimizes the user's prior expected loss. The user solves the following problem.

2.3.3.2 Problem (User's Prior Target Estimation). *Find a value θ that minimizes $E(f(A, \theta))$. A solution to this problem is the user's prior estimator and is denoted by $\Theta(A)$.*

After making an observation it is assumed that the user selects the estimator which minimizes the user's posterior expected loss given an observation. The user then solves the following problem.

2.3.3.3 Problem (User's Posterior Target Estimation). *Find a value θ that minimizes $E(f(A, \theta)|B)$. A solution to this problem is the user's posterior estimator and is denoted by $\Theta(A|B)$.*

A large value for the minimal expected loss indicates that the user's estimate may be poor and so the user is uncertain about their knowledge of the target. Note that if the user selects a different estimator, then the user's expected loss can only be larger. Assuming that the user selects the estimator that minimizes the expected loss is the same as assuming the best case from the user's perspective.

2.3.3.4 Definition (Prior Uncertainty). *A user's prior uncertainty in estimating a target is the user's expected loss. The prior uncertainty $\Psi(A)$ is given by $\Psi(A) = E(f(A, \Theta(A)))$. A user's prior certainty is the reciprocal of the user's uncertainty.*

2.3.3.5 Definition (Posterior Uncertainty). *A user's posterior uncertainty in estimating a target is the user's posterior expected loss. The posterior uncertainty $\Psi(A|B)$ is given by $\Psi(A|B) = E(f(A, \Theta(A|B))|B)$. A user's posterior certainty is the reciprocal of the user's uncertainty.*

A useful loss function in this context is the quadratic loss function. The quadratic loss function is a good choice for continuous data although others are possible. It leads to uncertainties that may be expressed in terms of familiar statistical quantities as the following lemma states.

2.3.3.6 Definition (Quadratic Loss). *The quadratic loss function is given by $f(a, \theta) = (a - \theta)^2$.*

2.3.3.7 Lemma (Quadratic Loss). *Given a quadratic loss function the prior estimator $\Theta(A)$ is given by $\Theta(A) = E(A)$ and the prior uncertainty $\Psi(A)$ is given by $\Psi(A) = V(A)$. The posterior estimator $\Theta(A|B)$ is given by $\Theta(A|B) = E(A|B)$ and the posterior uncertainty is given by $\Psi(A|B) = V(A|B)$.*

2.4 General Disclosure Control Problem

The concept of uncertainty leads to natural definitions of risk and utility. To this end let τ be a confidential target, v be an analytical target, α be an attacker and β be an analyst. Also let $X_{\tau\alpha}$ be a user target variable for target τ and user α , Y_τ be a base target variable for target τ and Z_α be a user observation variable for user α . Finally let z be an observed value.

The distribution of the user target variable $X_{\tau\alpha}$ represents the knowledge of the user α of the target τ before publication and the distribution of the user observation variable Z_α represents the knowledge of the user α of the altered data before publication. The joint distribution of the user target variable $X_{\tau\alpha}$ and the observation variable Z_α is the prior distribution. The distribution of the base target variable Y_τ is the base distribution and is used to make comparisons with uncertainties in risk and utility calculations.

The utility and risk are defined in terms of uncertainties of the above variables. Using these definitions the basic disclosure control problem becomes the *general disclosure control problem*.

2.4.0.8 Definition (Utility). *The utility of the altered data is the smallest relative certainty of a user in estimating an analytical target. The utility U is a function of the disclosure control parameter ϕ given by*

$$U(\phi) = \inf_{v\beta z} \frac{\Psi(Y_v)}{\Psi(X_{v\beta}|Z_\beta(\phi) = z)}.$$

The cost of the altered data is the reciprocal of the utility of the altered data.

2.4.0.9 Definition (Risk). *The risk of the altered data is the largest relative certainty of a user in estimating a confidential target. The risk R is a function of the disclosure control parameter ϕ given by*

$$R(\phi) = \sup_{\tau\alpha z} \frac{\Psi(Y_\tau)}{\Psi(X_{\tau\alpha}|Z_\alpha(\phi) = z)}.$$

The safety of the altered data is the reciprocal of the risk of the altered data.

Here the convention that if the prior uncertainty of the base variable and the posterior uncertainty of the user variable are zero then the ratio is not considered when determining the infimum or supremum.

3 Simple RTA

In simple Random Tabular Adjustment (RTA) the use of the general disclosure control model is demonstrated using a simple situation. Here the original data consists of a set of individuals and their contributions to a single cell. The contributions are real-valued, unbounded and continuous data. The total of this cell is to be published. The contribution of each individual to the cell total is confidential. The users of the published data are individuals that may include the contributors. To control the risk of disclosure, a random value is added to the cell total.

3.1 Simple RTA Model

The parts of the general disclosure control model for simple RTA are specified and the resulting expressions are derived so that the simple RTA problem can be solved.

3.1.1 Disclosure Control Method

Disclosure is controlled by adding a random variable Δ to the cell total where $\Delta \sim N(0, \sigma^2)$. The parameter σ^2 controls the variance of the random variable added to the total and is the disclosure control parameter in this model.

3.1.2 Targets and Users

There is one analytical target, the cell total. The user target variable for the cell total is denoted by X_{g^*} for user individual g . There are many confidential targets, namely each target individual's contribution, and many users estimating these targets. The user target variable for the individual's contribution is denoted by X_{gh} for user individual g and target individual h .

3.1.3 Prior, Posterior and Base Distributions

It is assumed that the users have some prior knowledge of each individual's contribution to the cell and that the set of contributors to the cell is known. The distribution of the user data variable D_{gi} represents the knowledge of the user individual g of the contribution to the cell of the individual i . It is assumed that $D_{gi} \sim N(m_{gi}, v_{gi}^2)$ and that these variables are independent. A base data variable E_i for individual i is selected such that $E_i \sim N(n_i, w_i^2)$ and again it is assumed that these variables are independent.

The target variables and observation variables can be expressed in terms of these data variables. The confidential user target variable for user individual g and target individual h is given by $X_{gh} = D_{gh}$, the analytical user target variable for user individual g is the unaltered cell total $X_{g^*} = \sum_i D_{gi}$ and the observation variable for user individual g is the altered cell total $Z_g = \sum_i D_{gi} + \Delta$. Similarly the confidential base target variable is given by $Y_h = E_h$ and the analytical base target variable $Y_* = \sum_i E_i$.

The posterior distribution for the confidential target variables can be determined from the prior distributions then given by

$$\begin{bmatrix} X_{gh} \\ Z_g \end{bmatrix} \sim N \left(\begin{bmatrix} m_{gh} \\ m_{g^*} \end{bmatrix}, \begin{bmatrix} v_{gh}^2 & v_{gh}^2 \\ v_{gh}^2 & v_{g^*}^2 + \sigma^2 \end{bmatrix} \right).$$

From this it follows that the posterior distribution is given by

$$X_{gh}|Z \sim N \left(m_{gh} + \frac{v_{gh}^2}{v_{g^*}^2 + \sigma^2} (Z_g - m_{g^*}), v_{gh}^2 - \frac{v_{gh}^4}{v_{g^*}^2 + \sigma^2} \right).$$

The posterior distribution for the analytical target variables can also be determined from the prior distributions. The joint distribution of X_{g^*} and Z_g is then given by

$$\begin{bmatrix} X_{g^*} \\ Z_g \end{bmatrix} \sim N \left(\begin{bmatrix} m_{g^*} \\ m_{g^*} \end{bmatrix}, \begin{bmatrix} v_{g^*}^2 & v_{g^*}^2 \\ v_{g^*}^2 & v_{g^*}^2 + \sigma^2 \end{bmatrix} \right).$$

From this it follows that the posterior distribution is given by

$$X_{g^*}|Z_g \sim N \left(m_{g^*} + \frac{v_{g^*}^2}{v_{g^*}^2 + \sigma^2} (Z_g - m_{g^*}), v_{g^*}^2 - \frac{v_{g^*}^4}{v_{g^*}^2 + \sigma^2} \right).$$

The distributions of the base variables can also be determined. Let n_* and w_* be given by $n_* = \sum_i n_i$ and $w_*^2 = \sum_i w_i^2$ so that the distributions of the base variables are given by

$$Y_h \sim N(n_h, w_h^2), \quad Y_* \sim N(n_*, w_*^2).$$

3.1.4 Loss and Uncertainty

Recall that the uncertainty is the minimal expected loss associated with a user estimation problem and so measures how poor the user's best estimate is in estimating the target.

A natural choice of loss function in this context is the quadratic loss function. Under this loss function the uncertainties are easy to determine as they are the variances of the above distributions. The posterior uncertainties for the confidential user target variables are given by

$$\Psi(X_{gh}|Z_g = z) = V(X_{gh}|Z_g = z) = v_{gh}^2 - \frac{v_{gh}^4}{v_{g^*}^2 + \sigma^2}.$$

The posterior uncertainties of the analytical user target variables are given by

$$\Psi(X_{g^*}|Z_g = z) = V(X_{g^*}|Z_g = z) = v_{g^*}^2 - \frac{v_{g^*}^4}{v_{g^*}^2 + \sigma^2}.$$

For the base target variables the corresponding uncertainties are given by

$$\Psi(Y_h) = V(Y_h) = w_h^2, \quad \Psi(Y_*) = V(Y_*) = w_*^2.$$

3.2 Simple RTA Problem

Using the above expressions for the various uncertainties in the model we find simple expressions for the utility and risk functions given by

$$U(\sigma^2) = \inf_{gz} \frac{\Psi(Y_*)}{\Psi(X_{g*}|Z_g(\sigma^2) = z)} = \inf_g \frac{w_*^2}{v_{g*}^2 - \frac{v_{g*}^4}{v_{g*}^2 + \sigma^2}},$$

$$R(\sigma^2) = \sup_{ghz} \frac{\Psi(Y_h)}{\Psi(X_{gh}|Z_g(\sigma^2) = z)} = \sup_{gh} \frac{w_h^2}{v_{gh}^2 - \frac{v_{gh}^4}{v_{g*}^2 + \sigma^2}}.$$

3.2.0.1 Problem (Simple RTA). *Find a disclosure control parameter σ^2 that maximizes*

$$\inf_g \frac{w_*^2}{v_{g*}^2 - \frac{v_{g*}^4}{v_{g*}^2 + \sigma^2}}$$

under the constraint

$$\sup_{gh} \frac{w_h^2}{v_{gh}^2 - \frac{v_{gh}^4}{v_{g*}^2 + \sigma^2}} \leq 1.$$

This problem can be solved analytically. The objective function is a decreasing function of σ^2 so if there is a solution, the value that maximizes the objective function is the smallest that satisfies the constraint. If $v_{gh}^2 > w_h^2$ for all target individuals h and user individuals g , then the solution is given by

$$\sigma^2 = \sup_{gh} \left(\frac{v_{gh}^4}{v_{gh}^2 - w_h^2} - v_{g*}^2 \right)$$

provided the right hand side is non-negative and $\sigma^2 = 0$ otherwise. If $v_{gh}^2 \leq w_h^2$ for some target individual h and user individual g , then there is no solution.

Once the variance σ^2 is determined, a realized value of Δ is randomly selected and added to the total x to get z . The value z together with the variance σ^2 are published.

3.3 Simple RTA Parameters

One way to interpret the prior and base distribution parameters in simple RTA is in terms of user knowledge and protection. The prior distribution parameters determine the most knowledgeable user protected against and the base distribution parameters determine the amount of protection given to the target. Smaller prior variances determine more knowledgeable users and larger base variances determine more protected targets. There are many ways to select these parameters but they all involve making assumptions about the knowledge of the users estimating the targets and how much protection a target requires.

Here is one way to select the prior and base distribution parameters that leads to some simplifications under some reasonable assumptions. It is assumed that the contribution of each individual is a target and that each contributing individual is an attacker. Furthermore it is assumed that each individual knows their own contribution and that every other individual knows this contribution equally well. This situation corresponds to setting the prior variances using

$$v_{gh}^2 = \begin{cases} r_h^2 & \text{if } g \neq h \\ 0 & \text{if } g = h \end{cases}$$

Note that the variance is zero when $g = h$ indicating that there is no variability associated with an individual's own contribution and when $g \neq h$ the variance depends only on the target individual h .

Using the above solution to simple RTA, if $r_h^2 > w_h^2$ for all target individuals h then

$$\sigma^2 = \sup_{gh} \left(\frac{r_h^4}{r_h^2 - w_h^2} + r_g^2 - r_*^2 \right)$$

provided the right hand side is non-negative and $\sigma^2 = 0$ otherwise. If $r_h^2 \leq w_h^2$ for some target individual h , then there is no solution. Here $r_*^2 = \sum_i r_i^2$. Note that determining the supremum in this case does not require a search through all possible pairs of individuals g and h . Only the pairs where user individual g maximizes r_g^2 or target individual h maximizes $r_h^4 / (r_h^2 - w_h^2)$ need to be considered. This reduces the amount of work needed to calculate the disclosure control parameter σ^2 .

If a size measure s_i is available for each individual i , coefficients of variation can be selected for the prior and base variances. This provides a simple and understandable way of determining all the prior and base parameters. When a prior coefficient of variation ε and a base coefficient of variation η are selected, the prior and base variances are given by $r_i^2 = \varepsilon^2 s_i^2$ and $w_i^2 = \eta^2 s_i^2$.

Using the above solution, if $\eta < \varepsilon$ then

$$\sigma^2 = \sup_{gh} (\lambda^2 s_h^2 + \varepsilon^2 s_g^2 - \varepsilon^2 s_*^2)$$

provided the right hand side is non-negative and $\sigma^2 = 0$ otherwise. If $\eta \geq \varepsilon$ then there is no solution. Here $\lambda^2 = \varepsilon^4 / (\varepsilon^2 - \eta^2)$ and $s_*^2 = \sum_i s_i^2$. Since $\lambda > \varepsilon$, this supremum is attained when the individual h has the largest size and the individual g has the second largest. It follows that

$$\sigma^2 = \lambda^2 s_{(1)}^2 + \varepsilon^2 s_{(2)}^2 - \varepsilon^2 s_*^2$$

where $s_{(1)}$ is the size of the largest individual and $s_{(2)}$ is the size of the second largest individual. This expression is similar to the expression used in traditional sensitivity rules such as the pq -rule. A discussion of sensitivity rules can be found in Willenborg and De Waal (2001).

4 Application and Comparison

RTA has been applied to data from the Canadian Monthly Retail-Trade Survey (MRTS) as a test of the methodology. In this section an overview of this survey is given as well as a description of the current disclosure control methodology. A description of the application of RTA to MRTS is then given and a comparison of the two disclosure control methodologies is made.

4.1 A Brief Overview of MRTS

MRTS is a survey of businesses that publishes the estimates of total retail sales by geography and industry each month. The geographic levels of publication consist of provincial and territorial levels and national level of Canada as well as some selected Census Metropolitan Area (CMA) levels. The industry levels consist of selected NAICS (North American Industry Classification System) levels from 441100 to 454110.

MRTS is a sample survey with take-all, take-some, and take-none strata. There are many large businesses in the population. These are always sampled and so are included in the take-all part of the sample. The contributions of these businesses have the highest risk of disclosure as they tend to make up a large proportion of the estimates which they are a part of. Most estimates have contributions from businesses in the take-some part of the sample and so most estimates have sampling errors. Sampling variances are calculated for each estimate.

In addition confidentiality waivers are collected from some businesses. When a business has given a confidentiality waiver, the business waives their right to the confidentiality protection provided by the *Statistics Act*. A confidentiality waiver allows us to discount the waived contribution when determining disclosure risk. Waivers should be taken into consideration in any disclosure control methodology for MRTS.

4.2 Current Disclosure Control Methodology for MRTS

Currently MRTS uses suppression as its method of disclosure control. This involves conducting a risk assessment using a cell sensitivity measure and then finding a suppression pattern by solving a linear programming problem using Statistics Canada's G-Confid system. Similar approaches are discussed in Willenborg and De Waal (2001). In

the suppression pattern the sensitive estimates along with other estimates (secondary suppressions) are suppressed in order to prevent the recalculation of the sensitive estimates using the other published estimates. The confidentiality waivers are taken into account when determining the sensitivities of the estimates. Sampling variances and weights are not taken into account in the current MRTS disclosure control methodology. Cell sensitivities are calculated as if the sample were a census. This is a conservative assumption.

Around 20% to 25% of the estimates in the publication table are suppressed each month although the total dollar value of the suppressed estimates is very small (about 2%). However for some provinces and industries the total dollar value of the suppressed estimates is large, particularly in small provinces and territories, whose estimates are often selected for secondary suppression.

4.3 Application of RTA to MRTS

The RTA methodology was applied to MRTS. This was done by selecting a set of low level estimates, selecting appropriate parameters, solving the RTA problem for each estimate, randomly altering the estimates, and aggregating to the higher level estimates.

A set of low level estimates was selected that could have its estimates randomly adjusted independently and could determine values for all the other estimates by aggregation. The set selected included the estimates of the internal cells of the table and some additional estimates to accommodate the CMA level cells. This was done in order to preserve the additive relations among the estimates, reduce disclosure risk and make the implementation easier. Independent random adjustments provide some protection to individuals who contribute to multiple estimates. Correlated random adjustments are possible but make the risk assessment difficult. Correlations may allow an attacker to use one part of the table to make inferences about other parts and this complication is not accounted for in the simple RTA model. Allowing for correlations in the RTA model is a topic for future work.

The parameters for RTA were selected in a way that provides a similar level of protection to that of the current suppression methodology. This involved selecting values for ϵ and η (as in section 3.3) so that the values for σ^2 are comparable to that of the cell sensitivity rule used in the current suppression methodology. In particular it is desirable to have parameter values that produce a positive value for σ^2 whenever the cell sensitivity is positive and conversely. While this is desirable, it is not possible to have complete agreement on this as the RTA methodology is based on variance calculations and cell sensitivity calculations are not.

The simple RTA problem is solved (as in section 3.3) for each estimate using the selected parameters and the business' sales value itself as the size. Adjustments were made to the formulas in section 3.3 to account for survey weights and confidentiality waivers. Weights are interpreted as a count of the number of businesses in the population that the sampled business is representing and the formula adjusted accordingly. Waivers are accounted for by changing how the worst case attacker/target pair is found. For example if the largest contributor has a waiver but the second largest does not, then the worst case occurs when the second largest contributor is the target and the largest is the attacker. Sampling variance is also accounted for by subtracting the sampling variance from the variance of the random adjustment required by RTA since sampling variability may be considered as part of the disclosure control method in the RTA methodology.

4.4 Comparison of Methodologies

Comparing the RTA and the suppression methodology is difficult as they are described using different sorts of concepts. In RTA the concepts are probabilistic and in suppression the concepts are geometric. It is desirable to compare the two methodologies using similar levels of risk and similar measures of utility.

As mentioned above, in order to make the disclosure risks comparable, the RTA parameters were selected so that the level of protection against disclosure under RTA should be roughly equivalent to the level under the current suppression methodology.

Utility in the two methodologies is measured differently. To compare the two, it was decided to simply describe the differences instead of trying to find a common utility measure. The tables that follow illustrate and summarize these differences.

Table 1 shows the results of the current methodology next to the results of the RTA methodology. It shows the results for one month for one province at the lowest published industry levels using simulated MRTS data. Here the sensitive cells are highlighted.

This table is a good illustration of the differences between the RTA methodology and the current suppression methodology. Note that there are a number of estimates that are suppressed under the current methodology but are published with no adjustment (CV 0%) under the RTA methodology (like NAICS 44111). These are estimates that are secondary suppressions and used in the current suppression methodology to protect the sensitive estimates. This protection is not needed in RTA and so all the secondary suppressions may be published. Under the RTA methodology some sensitive cells are substantially adjusted (like NAICS 4521 with CVs about 10%) while some need only a small adjustment (like NAICS 44511 with CV about 3%). Note that regardless of how sensitive an estimate is, it is suppressed under the current methodology while under the RTA methodology an estimate that has a high sensitivity has a random adjustment with a large CV and an estimate that has a low sensitivity has a random adjustment with a small CV. There is also an estimate (NAICS 44112) that was suppressed under the suppression methodology that is published unadjusted under the RTA methodology. This is because the sampling variance is larger than the variance required by RTA.

North American Industry Classification System (NAICS)	Suppression	Adjustment	
	Sales	Sales	CV
New car dealers [44111]	X	255	0.0%
Used car dealers [44112]	X	15	11.9%
Other motor vehicle dealers [4412]	22	22	0.0%
Automotive parts, accessories and tire stores [4413]	27	27	6.8%
Furniture stores [4421]	21	21	0.0%
Home furnishings stores [4422]	11	11	0.0%
Electronics and appliance stores [443]	19	19	0.0%
Building material and garden equipment and supplies dealers [444]	10	10	0.0%
Supermarkets and other grocery (except convenience) stores [44511]	284	287	3.1%
Convenience stores [44512]	13	13	17.0%
Specialty food stores [4452]	X	15	0.0%
Beer, wine and liquor stores [4453]	X	55	8.8%
Health and personal care stores [446]	131	131	3.5%
Gasoline stations [447]	166	166	8.8%
Clothing stores [4481]	X	3	0.0%
Shoe stores [4482]	X	15	0.0%
Jewellery, luggage and leather goods stores [4483]	65	65	0.0%
Sporting goods, hobby, book and music stores [451]	31	31	7.2%
Department stores [4521]	X	65	10.5%
Other general merchandise stores [4529]	X	55	1.9%
Miscellaneous store retailers [453]	X	45	15.3%

Table 1

Tables 2 and 3 show the difference in quality for published and suppressed estimates under the different methodologies. The quality grades represent standard CV ranges and the values in the tables indicate the number of estimates. Here the grades are as follows: A (0% to 5%) B (5% to 10%) C (10% to 16.5%) D (16.5% to 25%) E (25% to 33%) F (over 33%).

Table 2 shows the change in quality among the published estimates. It can be seen that almost all the estimates have their original quality grade after any adjustment using the RTA methodology. There are 85 published estimates whose quality grade move from A to B out of 9869 estimates whose original quality grade is A. These estimates are usually marginal totals that are published under the suppression methodology but are required to absorb the random adjustment made to some internal cells under the RTA methodology.

Table 3 shows the change in quality among the suppressed estimates. The estimates included in this table are all suppressed under the current methodology but are published under the RTA methodology. Note that there are 803 suppressed estimates with original quality grade A that have grade A after the application of RTA. These estimates are mostly estimates that are chosen as secondary suppressions under the current methodology. There are also 110 suppressed estimates with original quality grade A that have grade C after the application of RTA. These estimates are mostly estimates that were originally sensitive under the current methodology and so require a random adjustment with a large CV.

	Quality of Published Estimates After Application of RTA						
Original Quality	A	B	C	D	E	F	Total
A	9784	85	0	0	0	0	9869
B	0	4975	0	0	0	0	4975
C	0	0	1996	0	0	0	1996
D	0	0	0	708	0	0	708
E	0	0	0	0	42	0	42
F	0	0	0	0	0	12	12
Total	9784	5060	1996	708	42	12	17602

Table 2

	Quality of Suppressed Estimates After Application of RTA						
Original Quality	A	B	C	D	E	F	Total
A	803	643	110	0	0	0	1556
B	0	146	0	0	0	0	146
C	0	0	108	0	0	0	108
D	0	0	0	35	0	0	35
E	0	0	0	0	27	0	42
F	0	0	0	0	0	14	14
Total	803	789	218	35	27	14	1886

Table 3

In summary it is seen that, for comparable levels of disclosure risk, using the RTA methodology, many more estimates are published at the cost of some loss in precision. In particular a large number of non-sensitive low-level estimates that were suppressed to protect sensitive estimates are published with no change in precision under RTA. In addition under RTA some estimates that are sensitive require only a random adjustment with a small CV, and so may be published with a small loss in quality. Some estimates may even have a large enough sampling variance to control the risk of disclosure without any further random adjustment and so may be published as is.

5 Conclusion

A general disclosure control model has been proposed and used to solve the simple RTA problem. This model uses concepts from Bayesian decision theory to formulate disclosure control problems. It does this by defining the utility and risk of the published data in terms of users who make inferences about targets using the published data. Solving the general disclosure control problem involves finding the value of the disclosure control parameter that maximizes the utility while constraining the risk. The simple RTA problem is formulated and solved analytically. This problem involves a single cell total that is to be published while protecting the contributions of the individuals who contributed to the cell. When the prior and base parameters are selected in certain ways, the disclosure control parameter that solves the problem is simple to calculate and is similar to traditional sensitivity rules used in the disclosure control of tabular magnitude data.

The RTA disclosure control methodology was applied to the monthly retail trade survey (MRTS) and the results compared to the current suppression methodology used by MRTS. For comparable levels of protection against disclosure, many suppressed estimates are published at the cost of some loss in quality.

References

DeGroot, M. H. (1962). Uncertainty, information, and sequential experiments. *Annals of Mathematical Statistics*, 3:404–419.

Duncan, G. T. and Lambert, D. (1986). Disclosure-limited data dissemination. *Journal of the American Statistical Association*, 81:10–18.

Fienberg, S. E. and Trottini, M. (2002). Modelling user uncertainty for disclosure risk and data utility. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10:511–527.

Willenborg, L. and De Waal, T. (2001). *Elements of Statistical Disclosure Control*. Springer, New York, NY.