WP. 52
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 (ix): Statistical disclosure limitation for table and analysis servers: how to make outputs of modern data access infrastructures safe

Protecting Confidentiality in a Remote Analysis Server for Tabulation and Analysis of Data

Prepared by James Chipperfield and Frank Yu, Australian Bureau of Statistics, Australia

Protecting Confidentiality in a Remote Analysis Server for Tabulation and Analysis of Data

James Chipperfield* and Frank Yu*

*Australian Bureau of Statistics, ABS House, Belconnen, ACT 2614, Australia, james.chipperfield@abs.gov.au and frank.yu@abs.gov.au

Abstract: Many national statistical offices are looking to improve their output dissemination strategy by enhancing access to microdata through the use of remote access analysis servers. In this approach, results of statistical analyses or tabulation of the data are released in a form that will not enable any microdata to be linked to individuals. Typically, the original data are altered before release for example, by perturbing, coarsening or other techniques of statistical disclosure limitations.
The Australian Bureau of Statistics (ABS) is developing a remote access service which will enable users to submit requests for tabulation of count data, and analysis outputs from statistical models fit by using the dataset, while ensuring confidentiality of individuals' information is strictly maintained. In this paper we give an overview of the methodology used for protecting confidentiality for the datasets. The methodology has to be robust enough to support a wide range of queries and analyses, against risks from different kinds of attacks, such as differencing, transformations, analysis of residuals and outliers, and other inferential disclosures.

The remote data access server has two components: a Survey Table Builder to support requests for tabulation output, and an Analysis Server for regression output. The Survey Table Builder methodology utilises a range of perturbation and suppression (for very sparse tables) techniques for disclosure control. Weighted counts are automatically confidentialised to prevent disclosure through differencing attacks and repeated requests for identical tables. The method extends that which was used by the ABS for the Census Table Builder of the 2006 Population Census.

The Analysis Server that will enable external users to submit queries remotely to analyse de-identified microdata. It will protect against the identification of respondents through a number of features that include query control and perturbation methods that are applied to model parameter estimates, variances and diagnostics. As with Table Builder, only confidentialised, aggregate outputs are permitted. The techniques for disclosure control involve perturbation of the score function, imposing some restrictions on the type of analysis performed and coarsening diagnostic information such as residual plots. The Analysis Server will allow users to undertake exploratory data analysis, perform data sub-setting and variable transformations, undertake a range of common statistical regression analyses and manage their user workspace for storing and printing analytical outputs and transformed datasets.

Keywords: confidentiality, remote access analysis server, perturbation.

Introduction

Vast amounts of micro-data are collected by agencies from Censuses, surveys and administrative sources. Such micro-data can be used in the development and evaluation of policy for the benefit, or utility, of society. For this reason, there is very strong demand from analysts, within government and universities, to access such micro-data. When allowing analysts access to its micro-data, the agency is often legally obliged to ensure that the risk of disclosing information about a particular person or organisation is acceptably low. Managing the risk of disclosure is commonly referred to as Statistical Disclosure Control (SDC). Even after removing personal identifying information, such as name and address, from the micro-data the risk of disclosure remains (see for example Willenborg and De Waal, 2001).

Methods of SDC for micro-data include reducing the level of detail, replacing real values with synthetic values (see for example Reiter, 2002), sub-sampling, micro-aggregation, swapping attributes between records, and perturbing categorical values. Mathews and Harel (2011), Duncan and Pearson (1991) and give good summaries of many of these, as well as a few more. However, these methods may reduce the utility of the micro-data in the following ways:

- Perturbing, or adding noise to, variables introduces measurement error which can significantly reduce the accuracy of estimates and the power of hypothesis tests.
- Statistical modelling may be more complicated (see Little, 1993), if it takes into account the fact that variables may not be equal to their true values
- Replacing true values with synthetic values, generated from a model, can be difficult and time-consuming since social relationships are often subtle and complex.
- Releasing a single set of micro-data, after applying SDC, is in some respects a one-size-fits-all solution. For example, the ABS releases micro-data from a 1% sub-sample of its national Census of Australian Households. A 1% sub-sample could amount to a significant loss of utility for analysis about small areas but not for Australia-level analysis.

Given the recent explosion in the amount of administrative data, facilitated through technological advances, the risk of disclosure for micro-data is arguably ever increasing. One way of potentially improving the trade-off between utility and disclosure risk is a remote analysis server. A simple model for a remote analysis server is:

A. An analyst submits a query, via the internet, to the agency's analysis server
B. The analysis server processes the analyst's query on the sensitive micro-data. The statistical output (e.g. regression coefficients) from the query is modified for the purpose of SDC. Some output may be restricted on the basis that it could allow an analyst to reconstruct the attributes of an arbitrary record.
C. The analysis server sends the modified output, via the internet, to the analyst.

Some advantages of a remote analysis server are:

- although the statistical output is modified, it is based on the real micro-data. This means complex relationships in the micro-data are essentially retained.
- the degree to which a particular output is modified can depend upon the output itself. For example, estimates at a broad level may require proportionally less modification than estimates at a fine, or small area, level. Since an analyst is restricted from viewing the attributes of any record, less modification is needed than would otherwise be the case.
- the impact of the modifications on the output can be broadly indicated to the analyst. If the impact is large the analyst may decide to ignore the results altogether.
- once the server is set up, it can process multiple analyses in real time.
- all submitted programs can be logged and audited. If an audit concludes an attempt at disclosure was made, the agency can revoke the analyst's access to the server and take legal action.

There are some disadvantages of a remote analysis server:

- Some statistical outputs may be aggregated (e.g. record-level residual plots may be replaced with box plots) or perturbed (e.g. regression coefficients), and others may be restricted altogether.
- The analyst may be restricted to use only analysis techniques supported by the server
- Analysis through a remote server may take longer than if the micro-data were available on the analyst's personal computer.

There has been some work on managing the disclosure risk of analysis and tabular output, (i.e. on point (B) above) . In respect to analysis output see Gomatam et. al (2008), Lucero and Zayatz (2010), Bleninger et. al (2010) and Sparks et. al (2008) and in respect to tabular output see Shlomo (2007). The goal of this literature is to protect against data attacks, which involves an

analyst using output from an analysis server to reconstruct attributes for one or more records which, if successful, could be used to attempt disclosure by linking to other micro-data.

The Australian Bureau of Statistics (ABS) is developing a remote access service which will enable users to submit requests for tabulation of count data, and analysis outputs from statistical models fit by using the dataset. Our main approach to disclosure control is by perturbation of the outputs, supplemented by some limitation of potentially identifying information. Sections 2 outline the ABS' method of managing disclosure risks for count tables, and Section 3 describes the approach for protecting analysis output. Method for protecting tables of continuous measurements is being implemented and will not be discussed in this paper.

2. Count Data

Cells in a table are either *internal* or *marginal*. The count for a marginal cell is a sum of two or more other counts appearing in the table. If a cell is not a marginal cell it must be an internal cell. We now describe the ABS' method for perturbing unweighted (section 2.1) and weighted counts (section 2.2) for internal and marginal cells of a table.

2.1 Census Table Builder

Here we describe the method of perturbing unweighted counts as implemented in Census Table Builder (CTB), an ABS remote server which allows analysts to remotely request contingency tables to be calculated from the Australian Census' micro-data. The perturbed tables are automatically returned to the analyst, with generally no intervention from ABS staff. The analyst can define the dimensions of the table and the attributes of the records contributing to the table with only limited restriction (only tables with a high percentage of cells with counts of 0 or 1 are not released).

Denote the i th unweighted sample count for an *internal* cell in a contingency table by $n_i = \sum_{j=1}^{n} \delta_{ij}$, where i =1, …, C, δ_{ij} =1 if the j th record on the micro-data belongs to the i th cell and δ_{ij} =0 otherwise, j=1, 2, .., n and $n = \sum_{i=1}^{n} n_i$. CTB releases n_i^* to the analyst instead of n_i , where

$$n_i^* = n_i + e_i^* + a_i^* ,$$

$n_i^* \geq 0$, $\left|e_i^*\right| \leq L_e$, $\left|a_i^*\right| \leq L_a$, and L_e and L_a are positive integers specified by the agency. Clearly, the difference between n_i and n_i^* is restricted to be less than $L = L_a + L_e$. The e_i^* s represent the random integer perturbation of the i th cell count. The a_i^* s are derived so that the internal and marginal counts are consistent and so that the changes to the marginal counts are bound (for details see Appendix).

Define $Var_*(\ )$ and $E_*(\ )$ to be the variance and expectation with respect to the perturbation distribution of e_i^*, which meets the following criteria:

a) $E_*\left(e_i^*\right) = 0$

b) $Var_*\left(e_i^*\right) = \sigma^2$

c) $\mathrm{Cov}_*\left(e_i^*, e_j^*\right) = 0 \quad if \ i \neq j$

d) whenever the same set of records contribute to a cell count, the value for e_i will always be the same (see Fraser and Wooton, 2005).

e) e_i^* is an integer

Criterion a) ensures the count data are unbiased over the perturbation distribution. Criterion b) means that any cell count has a fixed perturbation variance. Criterion c) ensures that differencing two cells counts does not remove the effect of perturbation. Criterion d) ensures the effect of perturbation is not removed by repeatedly requesting the same cell count.

Table 1 gives an illustrative example of tabular counts before and after perturbation. Perturbed counts are asterisked while original counts are not. For example, a true count of 1 is perturbed to 3.

Table: Example of tabular counts before and after perturbation.

	Treatment A				Treatment B			
	Success	*Trials*	*Success**	*Trials**	*Success*	*Trials*	*Success**	*Trials**
Clinic 1	1	5	3	6	10	20	9	17
Clinic 2	9	10	9	11	5	20	4	18
Totals	10	15	12	17	15	40	13	35

* Perturbed counts

2.2 Survey Table Builder

In 2011 the ABS is implementing an extension to Table Builder, called Survey Table Builder (STB), which applies SDC to survey-weighted count data. Denote the i th weighted count in a contingency table by $N_i = \sum_j d_j \delta_{ij}$, where d_j is the survey weight for the j th record. The corresponding perturbed count is $N_i^* = \left[\tilde{d}_i n_i^*\right] + A_i^*$, where $\tilde{d}_i = n_i^{-1} N_i$ is the average weight for records belonging to the i th cell, n_i^* is the perturbed sample count described previously, $[x]$ rounds x to the nearest integer, and A_i^* performs an analogous function to a_i^* but for weighted counts (for details see Appendix). STB will not release any information about $\tilde{d}_i$, e_i^*, n_i^* or N_i to the analyst. If $\tilde{d}_i = 1$ for all i, then the CTB and STB methods of SDC are equivalent. Marley and Leaver (2011) studied the measures of risk and utility associated with STB.

3. Analysis Server

3.1 Without Statistical Disclosure Control (Standard Case)

First we consider the standard case for estimating regression coefficients in a regression model. Consider micro-data from which an analyst specifies an outcome variable y and K covariates $\mathbf{x}$, where the data are $\mathbf{d} = \{(y_j, x_j) : j = 1, \ldots n\}$. Consider fitting a regression model with parameter $\boldsymbol{\beta}$

using an unbiased estimating function $H(\boldsymbol{\beta})$ (see Chambers and Skinner, 2003). In particular we consider the estimating equation

$$H(\boldsymbol{\beta}) = \sum_{i=1}^{n} G_j(\boldsymbol{\beta})\{y_j - f_j(\boldsymbol{\beta})\},$$

where $f_j(\boldsymbol{\beta}) = E\left(y_j \middle| x_j\right)$ and $G_j(\boldsymbol{\beta})$ is a vector of order *K* with *k* th element $G_{jk}(\hat{\boldsymbol{\beta}})$ which is a function of $\boldsymbol{\beta}$ and x_j but not of y_j. The solution to $H(\boldsymbol{\beta}) = \mathbf{0}$ gives the standard estimate , $\hat{\boldsymbol{\beta}}$, of the regression coefficients.

Data attacks involve obtaining $\hat{\theta}$ from one or more queries in order to reconstruct attributes for an individual record. These attacks can involve differencing , leveraging a single record, isolating a record with a covariate, and by making inferences from a highly accurate model. These are well discussed for example by Gomatam (2008). Data attacks can of course use other outputs, such as plots, diagnostic statistics, p-values in a data attack.

When designing a set of perturbations and restrictions to apply to a set of analysis output, it quickly becomes clear that a series of regressions designed to find the optimal model could be indistinguishable from a sophisticated data attack. Therein lies the challenge: not restricting the former while thwarting the latter.

3.2 With Statistical Disclosure Control

Below we discuss the approach ABS is considering to implement in its remote analysis server.

3.2.1 Estimation of Parameters

Instead of solving $H(\boldsymbol{\beta}) = \mathbf{0}$ and releasing $\hat{\boldsymbol{\beta}}$, the server solves

$$H(\boldsymbol{\beta}) = \mathbf{E}^* \qquad (1)$$

and releases the resulting estimator $\hat{\boldsymbol{\beta}}^*$, where $\mathbf{E}^* = (E_1^*, E_k^* \ldots, E_K^*)'$ are perturbations introduced for the purpose of SDC, $E_k^* = u_k^* e_k$, u_k^* is the uniform distribution on the range (-1,1), and $e_k = \max_j \left\{G_{jk}(\hat{\boldsymbol{\beta}})\left(y_j - f_j(\hat{\boldsymbol{\beta}})\right)\right\}$ is the maximum influence a record may have on the *k* th estimating equation. For example, for the case of binary variables and the logistic model $e_k = 1$. The distribution of the perturbations, E_k^*, are independent and if the same model is fitted the same value of $\mathbf{E}^*$ is used- this stops an analyst estimating $\hat{\boldsymbol{\beta}}$ by fitting the same model a number of times and averaging over the regression parameters obtained from solving (1).

The size of the perturbation is designed to be of sufficient size to mask the contribution of any record to the estimating equation. Applying the perturbation to the score function is important, since this is where $\hat{\boldsymbol{\beta}}$ imposes a constraint on the data values.

3.2.2 Inference

To make valid inference with $\hat{\boldsymbol{\beta}}^*$ an analyst will need to account for the variance from both the model and the perturbation of the estimating equation. The variance of $\hat{\boldsymbol{\beta}}^*$ is

$$\mathbf{V}_{m*}\left(\hat{\boldsymbol{\beta}}^*\right) = V_m\left(\hat{\boldsymbol{\beta}}\ \right) + V_*\left(\hat{\boldsymbol{\beta}}^*\right)$$

where $V_m\left(\hat{\boldsymbol{\beta}}\ \right)$ is the variance of $\hat{\boldsymbol{\beta}}$ due to the model (i.e. the absence of any perturbation) and $V_*\left(\hat{\boldsymbol{\beta}}^*\right)$ is the variance of $\hat{\boldsymbol{\beta}}^*$ due to the perturbation. We propose estimating $V_m\left(\hat{\boldsymbol{\beta}}\ \right)$ using the delete-a-group Jackknife (Rao and Wu, 1988). A benefit of the Jackknife is that it is simple to calculate and is unbiased when the micro-data have been collected from a sample with a complex design (e.g. clustered sampling), as is the case for many ABS surveys. The Jackknife method involves allocating all selection units to one and only one replicate group in the same way that the sample was selected from the population. Using a similar approach to the sandwich variance estimator (see Chambers and Skinner, 2003 pp.105), we derive $V_*\left(\hat{\boldsymbol{\beta}}^*\right) = \left(\mathbf{X}'\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{D}\left(\mathbf{X}'\mathbf{W}\mathbf{X}\right)^{-1}$, where $\mathbf{D} = Var_*(\mathbf{E}^*)$.

We argue that the uncertainty in the Jacknife variance estimator(see pp.196 Shao and Tu, 1996), due to the allocation of selection units to replicate groups, is such that the total variance, $\mathbf{V}_{m*}\left(\hat{\boldsymbol{\beta}}^*\right)$, cannot be used in a data attack.

3.2.3 General Restrictions

Several authors have noted that fixed-distribution perturbation (as used above) alone is not sufficient to protect analysis outputs in the context of multiple queries. Approaches to managing the additional risks have included imposing restrictions into the analysis server (see Gomatam et al., 2005; Sparks et al., 2008) On the other hand, when designing a set of restrictions to manage disclosure risk, it quickly becomes clear that a series of regressions designed to find the optimal model could be indistinguishable from a sophisticated data attack. Therein lies the challenge: not restricting the former while thwarting the latter. In this subsection, we mention a set of restrictions that do not defend against a particular data attack, but are designed to significantly hinder a data attacker while only making a minor reduction in utility. These general restrictions include:

- $n > 50$.
- $n/K > 10$
- $K > 5$
- models can be fitted to a subset of records, where the subset is defined by at most 4 (always less than K)
- binary variables originally on the micro-data
- new binary variables can only be created from two other binary variables that are originally on the micro-data.
- new continuous variables can be only be created by using certain transformations
- variables must be non-zero for at least 15 records.

- for models with only binary covariates, the number of covariate patterns in **x** must be greater than 50
- $\mathbf{X}'\mathbf{X}$ must be full rank, where $\mathbf{X} = (\mathbf{x}_1, ...\mathbf{x}_j, ...\mathbf{x}_n)'$ and $\mathbf{x}_j$ is the *K* column vector of covariates for the *j* th record.

The values (e.g. 50) used in the above restrictions are used for illustration and can of course be changed.

3.2.4 Additional Attack-Specific Restrictions

As mentioned above, there are some well documented data-attacks (see for example, Gomatam 2008). It makes sense to impose restrictions, in addition to those mentioned in section 3.2.3, to explicitly defend against them. These restrictions are not discussed here, for space, but can be found in Chipperfield and O'Keefe (2011). We do, however, briefly describe three attacks for which explicit defences are constructed.

One such attack is called a *differencing attack.* A differencing data attack involves fitting the same model to two sets of records that are identical except that one record is dropped from one of the sets. Differences in the regression coefficients from the two models could be used in an attempt to reconstruct attributes of the dropped record. For example, if the covariates of the dropped record are known to the attacker, the change in the regression coefficients would allow a binary outcome variable for the dropped record to be derived.

Another such attack involves *fitting different models to the same set of records* and their attributes (i.e. the same data set) by:

1. Swapping the choice for the outcome variable
2. Using a different link function (e.g. linear, logistic and probit)
3. Using variables that are different transformations of the same attributes

Each model imposes *K* constraints on a set of records' attributes, which are unknown to the analyst. The aim of this attack is to impose enough constraints so that it is possible to solve for the values in the underlying data set.

3.2.5 Diagnostics

A range of test statistics (see Hosmer and Lemeshow, 2000) are available to assess the model assumptions (e.g. normality of residuals) and model fit (e.g. AIC, R-squared). Again, when releasing such statistics the agency needs to balance the disclosure risk against the utility. Ideally, an analyst's model selection should not be influenced by statistical disclosure control.

The approach to SDC for the estimate of the dispersion parameter or diagnostic statistics closely follows that for regression coefficients. Denote such a parameter or statistic by $t^* = t(\hat{\boldsymbol{\beta}}^*, \mathbf{d})$. Instead of releasing $t^* = t(\hat{\boldsymbol{\beta}}^*, \mathbf{d})$ we release,

$$t^{**} = t^* + u^* s(\hat{\boldsymbol{\beta}}^*, \mathbf{d})$$

where u^* is a random variable on the range (-1,1) and $s(\hat{\boldsymbol{\beta}}^*, \mathbf{d})$ bounds the maximum influence that a single record in **d** can have on the statistic t^* given $\hat{\boldsymbol{\beta}}^*$.

Diagnostics that involve plotting individual record values (e.g. residual plots) will be aggregated in some way, following Sparks et. al (2008). For example, Q-Q plots will be replaced by a smoothed non-parametric regression line and residual plots with be replaced by parallel box plots.

Appendix

Denote the internal and marginal cells of a table by *t*=1, 2, ..., C, C+1,..., T, where *t*=1, 2, ..., C denotes the internal cells of the table. Denote the *t* th cell count by n_t . Instead of releasing n_t , TB releases $n_t^* = n_t + e_t^* + a_t^*$ which is obtained in two steps. The first step involves calculating the preliminary counts $m_t^* = n_t + e_t^*$, where e_t^* has properties a)-e) from section 2.2. The table's preliminary counts are not consistent: sums of preliminary counts for internal cells are not guaranteed to equal corresponding preliminary marginal counts. The second step involves finding the value for a_t^* that so that

The table with counts n_t^* is consistent and $\left|a_t^*\right| \le L_a$ for all *l*=1, ..., *L*. This means no preliminary count, for a marginal or internal cell, is changed by more than L_a .

References

Bleninger, P., Drechsler, J. and Ronning, G. (2010) Remote Data Access and the Risk of Disclosure from Linear Regression: An Empirical Study, Privacy in Statistical Databases, Springer.

Chambers, R.L. and Skinner, C.J. (2003) Analysis of Survey Data, John Wiley & Sons.

Chipperfield, J. O. and O'Keefe, M. C. (2011), Disclosure-Protected Inference using Generalised Linear Models, Submitted for publication.

Fraser, B. and Wooton, J. (2005) A proposed method for confidentialising tabular output to protect against differencing, UNECE work session on Statistical Data Confidentiality.

Gomatam, S. , Karr, A. F., Reiter, J. P., and Sanil, A. P. (2008) Data Dissemination and Disclosure Limitation in a World Without Microdata: A Risk –Utility Framework for Remote Access Analysis Servers, Statistical Science, 20, pp.163-177.

Hosmer, D. W. and Lemeshow, S. (2000) Applied Regression Analysis, John Wiley and Sons.

Little, R., J., A. (1993) Statistical Analysis of Masked Data, Journal of Official Statistics, 2, 407-426.

Marley, J. K and Leaver, V. L. (2011) A Method for Confidentialising User-Defined Tables: Statistical Properties and a Risk-Utility Analysis, Proceedings of the International Statistics Institute.

Mathews, G, J. And Harel, O. (2011), Data Confidentiality: A Review of methods for statistical disclosure limitation and methods for assessing privacy, Statistical Surveys, 5, pp. 1-29.

Rao, J.N.K. and Wu, C.F.J. (1988) Resampling Inference with Complex Survey Data, Journal of the American Statistical Association, 83(401), pp. 231–241.

Reiter, J. P. (2002) Satisfying Disclosure Restrictions with Synthetic Data Sets, Journal of Official Statistics, 18, 531-543.

Shao, J. and Tu, D. (1996) The Jackknife and Bootstrap, Springer.

Shlomo, N. and Skinner, C. (2010) Assessing the Protections provided by Missclassification-based Disclosure Limitation, The Annals of Applied Statistics, 1291-1310.

Sparks, R., Carter, C. Donnelly, J., O'Keefe, C.M., Duncan, J., Keighley, T. and McAullay, D. (2008) Remote Access Methods for Exploratory Data Analysis and Statistical Modelling: Privacy-Preserving AnalyticsTM, Computer Methods and Programs in Biomedicine 91, pp. 208-222.

Willenburg. L. and de Waal, T. (2000) Elements of Disclosure Control, Springer.