WP. 22
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Bilbao, Spain, 2-4 December 2009)

Topic (iv): Tools and software improvements

IMPLEMENTING A METHOD FOR AUTOMATICALLY PROTECTING USER-DEFINED CENSUS TABLES

Invited Paper

Prepared by Victoria Leaver, Australian Bureau of Statistics, Australia

Implementing a method for automatically protecting user-defined Census tables

Victoria Leaver*

* Data Access and Confidentiality Methodology Unit, Australian Bureau of Statistics, Locked Bag 10, Belconnen ACT 2616 Australia, victoria.leaver@abs.gov.au

Abstract: The Australian Bureau of Statistics has developed a method for automatically protecting tables of Census counts. This method protects against requests for similar tables and repeated requests for identical tables. These features allow it to be used in a web-based system for creating user-defined tables.
The method assigns a permanent numeric key to each unit record. These keys are used to generate consistent values for the perturbation that is applied to the cells in the table. Whenever the same contributors appear in a cell, the same perturbation will be applied. Perturbation is applied to all cells, to protect against differencing. An algorithm then restores additivity and maintains the perturbed grand total.
A suite of products has been developed to allow access to the confidentialised Census data. These products include web-based systems that allow users to define and download automatically-protected tables.
Possible future directions include extending the methodology to protect weighted counts and magnitude data.

1 Introduction

The Australian Bureau of Statistics (ABS) has been given the authority by legislation to collect statistical information. This legislation requires that any data collected under this authority shall not be released in a manner that is likely to enable the identification of a particular person or organisation.

The ABS has developed a range of policies to assist it in meeting these legislative requirements. One aspect of this policy covers the release of tables that contain information about very small sub-populations. The ABS has policies specifying that table cells containing very small counts should not generally be released. Also, it should not be possible for users to derive the true values of these cells from other data released by the ABS.

The ABS conducts the Census of Population and Housing every five years. The data from this Census allow detailed analysis about small sub-populations and are a valuable source of information that is not available elsewhere. However, it is

important that any release of Census data conforms with the legislation and ABS policy.

For the 2006 Census of Population and Housing, the ABS has developed a method for automatically protecting tables of Census count data. This method was designed to protect against requests for similar tables, repeated requests for identical tables, and repeated requests for the same table cell within different tables. The method was intended to allow greater access to data for sub-populations, and to enable the development of a web-based system allowing users to define their own tables. The aims of this method will be described in greater detail in section 2.

The methodology was originally presented in Fraser and Wooton (2005). In section 3, this paper will give a summary of the method.

The proposed methodology has been implemented in several products that enable the automatic creation of confidentialised tables from the 2006 Census data. These products include two web-based systems that allow users to define and download tables. These products will be described in section 4.

The ABS is now considering possible extensions of the method. The aim would be to make statistical tables from other collections available using similar web-based systems. The ABS will explore ways of adapting the method to protect weighted count data and magnitude data. Possible future directions will be discussed in section 5.

2 The aims of the method

The method was specifically designed to be used within a web-based product that allows users to define their own tables. Compared with a system that only releases a predetermined set of confidentialised tables, a product that creates user-defined tables has additional confidentiality risks.

One risk is that users may be able to undo some of the protections by making repeated requests for the same table. If the confidentiality method uses a stochastic process to deliver a random amount of perturbation, then a user could obtain different versions of the same table. Comparing the cell values across these different versions may reveal some information about the original, unconfidentialised table. For example, averaging the cell values could reveal the original, unprotected value. A method for a web-based product would need to address this risk.

The second main risk is posed by "differencing". Differencing can occur if users are able to obtain tables for similar sub-populations and then take the difference between the two tables to find the data for a much smaller sub-population. The ABS's

intention was to give users flexible ways of defining sub-populations. Therefore, the method would also need to address the risk posed by differencing.

Also, to be useful in a web-based product, the method would have to allow flexibility. Many data items are collected in the Census, and ideally users should be able to specify tables related to any of these data items. The standard geographical structure also has many levels. Ideally, users would be allowed to specify tables at most of these levels, and to create their own geographical areas from combinations of standard geographical units.

The standard output from the Census includes counts of people, families and dwellings. The method would have to allow tables to be created at the three different levels.

The method should not change the data so much that the tables are significantly less useful for analytical purposes, and ideally the tables should provide a coherent picture of Australia's population.

Finally, the method would need to complete the confidentialisation in a reasonable amount of time, so that it would be practical to implement and use.

3 A brief description of the method

A paper describing the method was presented to the joint UNECE/Eurostat Work Session on Statistical Data Confidentiality in Geneva in 2005 (Fraser & Wooton, 2005). That paper presented details of the proposed method, and that proposal was essentially the method that was implemented by the ABS.

The method attaches an element of a finite group to each unit record. These elements are then combined and mapped to indices that are used to derive the perturbation that will be applied to each cell.

In the ABS's implementation of the method, the elements are permanent numeric values known as "record keys". When a table is created, the record keys are combined using modulo arithmetic to produce a cell-level key. This cell-level key is used to determine the perturbation that is applied to the cell, via a fixed look-up table. Zero cells have no contributing records, and do not receive any perturbation. Interior cells and marginal totals are perturbed independently, so that the perturbed tables are not necessarily additive.

This method ensures that each cell receives the same perturbation whenever it appears in any table. Repeated requests for the same table will therefore produce the same results. Also, if the same cell appears in different tables, it will be perturbed in the same way each time, protecting against attempts to use varying results to determine the original value.

The look-up table contains the value of the perturbation that will be applied to the cell, based on the original cell value and the cell-level key. The cell-level key determines the row of the look-up table, and the cell value determines the column of the look-up table. For larger cell values, modulo arithmetic is used to determine the column. The distribution of perturbation values within the look-up table determines the range of the perturbation that can be applied, and the average magnitude of perturbation that is applied to non-zero cells.

With the exception of zero cells, every cell in each table has a chance of being perturbed. The maximum amount of perturbation that any cell can receive is fixed, which means that larger cell values will receive a proportionally smaller amount of perturbation.

Since all cells have a chance of being perturbed, the difference between any two tables will also be perturbed. This prevents small cell values being rederived by differencing. If small cells can be calculated by differencing, these cells will receive, on average, a proportionally large amount of perturbation. The method does not guarantee that the difference will be different from the true value, but the variance around the true value gives sufficient uncertainty to protect individuals from being identified.

The method does not depend on the data items or geographical areas that users choose for their tables. While the method can be applied to different levels of data, it does not guarantee exact consistency across these levels. For example, the count of dwellings and the count of persons in the same geographic area may be perturbed in different directions.

For all these reasons, the ABS decided that the proposed method would be appropriate for a web-based table builder product. There were a number of details that had to be finalised before the method could be implemented. In particular, there were investigations into:

1. the distribution used to generate the record keys;
2. the best way to combine the record keys to create the cell-level keys;
3. the distribution of the perturbation values held in the look-up table;
4. the additivity algorithm.

The choices for some of these parameters have a significant impact on the disclosure risk. In particular, the distribution of perturbation values in the look-up table determines the amount of protection applied to the tables, and the properties of this distribution were chosen to satisfy the requirements of the ABS's legislation and policy. As described in Fraser and Wooton (2005), the keys and the look-up table were designed to produce integer-value perturbations whose distribution satisfies the following criteria:

1. the mean is zero;
2. the perturbations will not create negative cell values or very small positive cell values;
3. the perturbations have a fixed variance;
4. the absolute value of any perturbation is less than a fixed positive integer.

As proposed by Fraser and Wooton (2005), an additivity module was also incorporated into the overall method. The purpose of this module was to restore additivity to the perturbed tables, so that the tables would make more sense to users and have greater utility. The additivity module uses iterative methods to restore additivity, under various constraints. These constraints specify that:

1. the resulting table will contain non-negative integer values;
2. very small non-zero cell values will not appear in the additive table;
3. the perturbed grand total of the table will be preserved;
4. the changes to cell values will be minimised.

The additivity module does not share all the properties of the perturbation method. For example, if the same cell appears in two different tables, the cell will receive the same perturbation, but it may be altered in different ways by the additivity module. This is not considered to be a significant confidentiality risk. It may be possible for a user to undo the changes applied by the additivity module, for example by averaging cell values over a large number of table requests. However, the user will only be able to recover the underlying perturbed table, not the original data.

Also, zero cells may be altered by the additivity module for a very small number of unusual tables. This is not necessarily ideal, especially if there are cells that logically should not contain any contributors. However, it was necessary to allow the module to alter some zero cells, particularly in sparse tables, to meet the other constraints.

The additional changes introduced by the additivity module are usually small in comparison to those applied by the perturbation method. It is generally only with very sparse tables that the additivity module has a greater impact.

4 The web-based systems

Using the Census confidentiality method, the ABS has developed several systems for generating tables from Census data.

All tabular output from the 2006 Census is protected using the same method, including tables created by ABS staff for publications. For this reason, there are internal systems for applying the method.

There are also two web-based products available to users outside the ABS: CDATA Online and the Census TableBuilder. These products were jointly developed by the ABS and Space-Time Research Pty Ltd.

CDATA Online is available free of charge to all users with access to the Internet and one of a range of supported web browsers. CDATA Online contains a large collection of underlying topic-based data cubes, from which users can create customised tables, maps and graphs. While the data cubes are used to define the tables that users can request, the data cubes are not used directly in the confidentialisation process. That is, the tables are not confidentialised by aggregating confidentialised counts in the data cubes. Instead, each table created using CDATA Online is confidentialised using the standard method. Once users have requested tables, maps or graphs, these outputs can then be exported or downloaded. CDATA Online can be accessed from www.abs.gov.au/CDataOnline. This webpage also contains a link to the user manual.

Census TableBuilder is similar to CDATA Online, but allows more flexibility. Instead of containing underlying topic-based data cubes, Census TableBuilder allows users to create tables directly from the full range of Census data items. This means that users can include most combinations of Census data items in their tables. As with CDATA Online, users can create maps and graphs as well as tables. Users of Census TableBuilder need to register before they are allowed to use the product. There is more information about Census TableBuilder, and a user manual, at www.abs.gov.au/TableBuilder.

5 Future directions

The ABS intends to use the same method to confidentialise data from the 2011 Census, and does not expect major changes to the web-based products. However, there is still scope for future work.

To make informed decisions based on the confidentialised tables, users need some indication of the quality of the data. Ideally, it would be possible to allow users to adjust for the effect of the confidentialisation on their analyses, without revealing information that may compromise the protection of the data. The ABS has carried out some research into measures of information loss, including the chi-square test of association (Wooton, 2006), and work is continuing in this area. The aim is to develop measures that could be included in an information paper for users of CDATA Online and the Census TableBuilder. This paper would give more

information about the method than is currently available. It may also contain information about the average variance associated with the perturbation and the additivity, or some other measure of the quality of the tabular output.

The ABS is currently exploring collaborative approaches for improving access to data internationally. This may include making some aspects of the Census TableBuilder product available to other statistical organisations across the world, so that these organisations can load their own Census data into the system and make it available to users. There are certain parameters within the method that cannot be released externally without compromising the confidentiality of ABS data. However, the ABS may be able to provide advice to other organisations about how to set these parameters in a way that meets specific confidentiality requirements. Other organisations adopting the Census TableBuilder method will also need to ensure that the system is compatible with their metadata requirements.

The methodology was designed to meet the requirements of the ABS's legislation and policy, and the expected needs of users. Other statistical organisations may have different requirements due to their legislation and policy, or the needs of their users. For this reason, some organisations may prefer to alter aspects of the ABS's methodology. For example, the original proposal for the method (Fraser & Wooton 2005) included the possibility of assigning record keys in a way that preserves certain pre-specified totals exactly. The ABS chose not to follow this path, but other organisations may decide that preserving exact totals is a requirement for their Census data. Shlomo and Young (2008) have tested an approach that incorporates features of the ABS's Census methodology, but which is also designed to preserve marginal totals.

The ABS is also interested in extending the methodology and systems to allow the automatic confidentialisation of tables from sampled survey data. One of the first steps in the development of this Survey Table Builder will be to determine the best way of incorporating sampling weights into the method. The method will need to provide enough protection to meet the ABS's requirements, without introducing so much perturbation that the tables no longer contain useful information.

Depending on the types of survey data that will be made available in the Survey Table Builder, there will be additional methodological challenges.

If the ABS releases tables from repeated surveys, it may be necessary to explore methods of preserving the time series properties of sets of estimates from different points in time.

Also, the current methodology cannot be effectively applied to continuous data items such as profit and expenditure from business survey data. There are a number of promising approaches that involve perturbation. The amount of perturbation needed to protect continuous data items is likely to be proportionally larger than the perturbation needed for counts. This means that information and utility loss will

probably be of greater concern for continuous data items. A number of the possible approaches could be adapted to preserve certain statistical properties. However, it will be necessary to determine which statistical properties should be given higher priority. These constraints could then be incorporated into the method.

As with the Census method, the ABS is interested in exploring collaborative approaches for developing methods and systems for allowing increased access to survey data.

References

Fraser, B. & Wooton, J. (2005). 'A proposed method for confidentialising tabular output to protect against differencing'. Paper presented to the Joint UNECE/Eurostat work session on statistical data confidentiality, Geneva, Switzerland, 9-11 November.

Shlomo, N. & Young, C. (2008). 'Invariant post-tabular protection of Census frequency counts'. *Proceedings of Privacy in Statistical Databases 2008*.

Wooton, J. (2006). *Measuring and correcting for information loss in confidentialised Census counts.* 1352.0.55.083 - Research Paper, Australian Bureau of Statistics, <http://www.abs.gov.au/AUSSTATS/abs@.nsf/ProductsbyCatalogue/5CE405371612A81ACA257300001ADE40?OpenDocument>.