On-line Access of Micro-Data at the Australian Bureau of Statistics – Challenges and Future Directions

Tam, Siu-Ming
Australian Bureau of Statistics
45 Benjamin Way,
Belconnen, ACT 2617
Australia
E-mail: Siu-Ming.Tam@abs.gov.au

Introduction

The Australian Bureau of Statistics' (ABS) mission is to "assist and encourage informed decision making, research and discussion within governments and the community, by leading a high quality, objective and responsive national statistical service".

Dissemination of micro- data for statistical research in the form of Unit Record Files (URFs), which provide the most detailed level of information available from the ABS, contributes significantly to this mission.

URFs are authorised for release for statistical purposes at the discretion of the Australian Statistician. The authority for such releases is vested under the Census and Statistics Act, 1905, and the relevant Statistics Determination. However, the Act requires that the release must be carried out in a manner that is not likely to enable identification of a particular person or organisation. Such files are referred to as Confidentialised Unit Record Files (CURFs).

Current environment for micro-data access

Consistent with the Act, CURFs are currently released in "Basic" format via CDROM, in Expanded format via the secure, Internet based, Remote Access Data Laboratory (RADL), and in Specialist format in data laboratories located in ABS premises (ABSDL).

Basic CURFs released in CDROM contain highly confidentialised data which may be accessed on users' desktops, whereas Specialist CURFs released in ABSDL contained detailed micro-data which, with spontaneous recognition of the unit record removed, can only be accessed in tightly ABS controlled environments.

On the other hand, the Expanded CURFs allow the user remotely perform analyses on CURFs which provide details at a level between basic and specialist CURFs, using an ABS sanctioned set of commands under SAS, SPSS or STATA. This provides an online service complementing the desktop and ABSDL access to micro-data.

In all cases, not only the unit records are de-identified with names and addresses information removed, they are made un-identifiable by applying confidentialising procedures. As well, the outputs from users' analyses are also required to be made un-identifiable.

These access approaches represent a good compromise between the two ends of the data protection spectrum, i.e. safe data and safe environment, and represent the level of disclosure risks acceptable to the ABS under these release arrangements.

The ABS cost of providing CURF access by the ABS is recovered by a marginal pricing regime through which ABS recovers the cost of establishing and supporting user access to CURFs. The cost of creating the files is, however, not cost recovered from users.

To ensure that CURFs are widely used for research, the ABS has provided CURFs to Australian university researchers through a series of Agreements with the universities' peak body, Universities Australia. This enables university academic staff and students unlimited access to the nearly 150 CURFs that have been released by the ABS.

User requirements

In 2009 and 2010, user consultations were undertaken to determine the future direction for micro-data access in Australia.

The consultations identified that there was strong support for a suite of micro-data access mechanisms to support statistical research.

These users place a high value on the current flexibility of the Basic CURFs with researchers using them for a broad range of uses from the production of simple descriptive statistics to sophisticated statistical modelling.

Users also indicated support for the development of an upgraded RADL service, comprising a table generation service and an analysis service. Flexibility, accessibility and interactivity were seen as highly desirable requirements by researchers.

The consultation identified the main analytical tools required by sophisticated data users, including linear and logistic regression, marginal effects analysis and multinomial models.

Alternative options such as synthetic data sets were of interest to users. However, users expressed doubts about the ability of these data sets to preserve statistical relationships to provide the necessary level of utility.

The consultations also found that about 60 per cent of the current remote micro-data access through RADL is for the generation of statistical tables. As well, main issues identified in relation to RADL include users:

- want to undertake all the analysis they would like to do without restriction of functionality by ABS on SAS, STATA and SPSS;
- may not have the skills to write code in these languages;
- want the ABS to provide the most recent versions of the statistical languages in RADL;
- want RADL turnaround time for jobs similar to what they can have with Basic CURFs;
- want access to: a broader range of datasets (including more business micro-data);
- want access to richer micro-data including longitudinal and linked datasets; and
- want the time between enumeration and release of micro-data to be much shorter.

Disclosure risks

When individual unit records are allowed to be accessed, there are the following disclosure risks to be mitigated against (Willenborg and de Waal, 2001; Chipperfield and Lucie, 2010):

- Spontaneous recognition risk, which may occur when the analyst recognises someone, or a business, they know – this may be addressed by re-moving identifying information from the unit records;
- Matching records risk to identify an individual. The risk may occur when the analyst uses a set of variables common to both the ABS unit record file and a database with identifying information held by the analyst for matching and then identification. This risk applies not only to unit record data, but also inappropriately confidentialised aggregate data as well. This risk may only be addressed by explicitly requiring analysts to sign an undertaking not to match, and by confidentialising the data in such a way to reduce the probability of successful matches;
- Inferential risk, which may occur when relationships in micro-data when considered together are strong enough to accurately make inference about individual persons or businesses. This may be addressed by confidentialising the unit records sufficiently to reduce, or minimise the number of unique records in the CURF; and
- Differencing risk, which applies to aggregate outputs and may occur, when the difference between the cells of one or more tables generated from the CURF allow the analysts to make accurate inference about an individual. Mitigation of this risk requires tools that will still have perturbation remaining in the counts after differencing, e.g. the ABS TableBuilder developed for the 2006 Census.

Future strategy for micro-data access

The user consultation showed there is no single solution that will meet the requirements of all micro-data users and ABS will therefore continue to provide a range of modes of access to micro-data.

The future ABS strategy (Tam, Farley-Larmour and Gare, 2010) is to:

i. continue to produce and release Basic CURFs for use in the user's environment;

ii. progressively replace RADL with a remote execution environment for micro-data (REEM) consisting of two different but complementary services and of providing access to un-confidentialised URFs, and

iii. increase the use of the ABSDL for complex analysis of micro-data, including providing access to longitudinal and linked datasets.

The two key components of the REEM are the development of a TableBuilder covering census and survey data and an Analysis Service (for statistical analysis).

A key element of this future strategy is that REEM will use internationally recognised standards for the exchange of data and metadata. Such standards comprise Data Documentation Initiative (DDI) and Statistical Data and Metadata Exchange (SDMX), and machine to machine interfaces.

Survey TableBuilder

In 2009, the ABS released an interactive on-line product, Census TableBuilder, which allows authorised users to create customised tabular output from the full 2006 population census file (Fraser and Wooten, 2005).

The obvious extension of the technology is to extend it to survey datasets, given that over 60% of the outputs from RADL are statistical tables.

The underlying data source for Census TableBuilder is the raw de-identified unit record file. Primary identifiers, such as names and addresses, have been removed, but no further modifications have been made to the file.

To ensure respondent information is protected, TableBuilder includes a routine that dynamically confidentialises tabular output using perturbation techniques prior to it being returned to the user. This routine ensures the end results are unlikely to enable the identification of any respondents on the original file from which the results were generated.

The primary advantage of this technique is that it protects the confidentiality of individual’s information against the risk of identification through differencing multiple tables, As well, the impact on overall data quality of the confidentialisation routine is minimised by applying the necessary modifications only to the final output and not to the underlying data source (Wooten, 2006).

This method also ensures that each cell receives the same perturbation whenever it appears in a 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. Users running the same query again will also receive the same results.

Each cell in each table has a chance of being perturbed, including the marginal cells of a table. 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. An additivity module then restores additivity to the tables, in a way that preserves the perturbed values of the marginal totals in the table as much as is feasible. The additivity routine, unlike the perturbation routine, does not guarantee consistent estimates are produced if the same cell appears in a different table as additivity is conditional on the marginal and grand total of the specific table.

This method differs to the controlled rounding perturbation technique incorporated in Tau-Argus. The controlled rounding method rounds all cells in a table to a multiple of a chosen number, referred to as the base number, while preserving the marginal totals as much as possible.

Given the popularity of the Census TableBuilder, ABS is developing a survey version of this Builder to extend the tool for the creation of tables from de-identified micro-data from ABS surveys.

New analysis service

The ABS RADL will be replaced by a new analysis service. The new service will, like the current TableBuilder, utilise de-identified detailed micro-data, with confidentiality routines built into the outputs generated to ensure that they are confidentialised in line with ABS legislative requirements and that can be released as public use outputs (that is, they can be published and shared with others without restrictions) (Elazar, 2010).

This project is still a work in progress. However, for the class of general linear models that are fitted to count data, Chipperfield and Lucie (2010) have:

- used the TableBuilder methodology to perturb the counts data in the estimation equation for estimating the model parameters;
- used a combination of the ”delete-a-group” Jacknife and Bootstrap to estimate the variance of the estimated model parameters, to account for the additional variance due to Table Build perturbation;
- developed a methodology for making inference on the model parameters;
- applied the method to the estimation of model parameters for three logistic models using the 2008 National Health Survey, and shown that the increase in standardised measure of bias, and the increase in mean squared error were all between 0.1% to 1.1%, with predominately most measures at the lower end of the interval.

Future ABSDL

The ABSDL is seen by the ABS as the ideal highly secure environment in which complex analysis of micro-data and analysis of longitudinal and linked datasets can be undertaken. However, cost and timeliness of access have hitherto acted as barriers to use of the ABSDL.

Client consultations in 2009 and 2010 indicated that the existing ABSDL and Specialist CURFs are unlikely to form a significant part of our clients' future micro-data access solutions unless cost is reduced and timeliness of access increased.

ABS is therefore challenged to increase the value proposition of the ABSDL for clients. It is likely that this value proposition lies not in seeking mass use of the ABSDL, but in ABS accepting and embracing the concept of the ABSDL providing “vital data for vital projects”. Such “vital projects” may be small in number overall but their importance to government, business or the community will outweigh the inconvenience of visiting an ABS Office, especially given the access to ABS expertise that on-site micro-data analysis makes possible.

Conclusion

ABS has a mature program for provision of micro-data products to its clients and while researchers have always sought access to more datasets and more functionality, for the most part they have been broadly satisfied with access to the range of data available and the modes of access provided. This situation is challenged, however, by newer demands for more flexible and interactive access to micro-data and for access to richer micro-data including longitudinal and linked datasets.

Concurrently ABS is concerned about new challenges to confidentiality as a result of the growing range of other data sets available to researchers and by the growing sophistication of new data matching technologies.

We are therefore very actively exploring new and innovative ways to provide micro-data access to researchers, and believe TableBuilders and the new Analytic Service will provide part of the solution package for more innovative way of access.

Making micro-data available in a manner such that individuals cannot be likely identified to support statistical research is key to the relevance of the ABS. We are therefore keen to share our expertise and are committed to international collaboration to developing new micro-data remote access and processing methods and systems. An OECD Expert Group, of which ABS is a member, has recently been established to provide a forum for national statistical offices to share their experience and expertise, and to collaborate to develop a vision for cross border access to micro-data.

REFERENCES

Elazar, D, 2010 – User analytical functionality for the REEM analysis service. Unpublished ABS research paper.

Chipperfield, J, and Lucie, S, 2010 – Analysis of micro-data: Controlling the *risk of disclosure.* Unpublished ABS research paper.

Fraser, B and Wooton, J, 2005 – A proposed method for confidentialising tabular output to protect against differencing, paper presented at the November 2005 UNECE Work Session on statistical data confidentiality

Tam, S, Farley-Larmour, K and Gare, M, 2010 – Supporting research and protecting confidentiality – Current strategies and future directions for ABS micro data access, Statistical Journal of the International Association of Official Statistics, Vol 27, p. 65 -72

Willenborg, I, and de Wall, T, 2001. Elements of Statistical Disclosure Control. Springer.

Wooton, J, 2006 – Measuring and correcting for information loss in confidentalising census counts. Unpublished ABS research paper