United Nations ECE/CES/2013/29

Economic and Social Council

Distr.: General
16 April 2013

Original: English

Economic Commission for Europe

Conference of European Statisticians

Sixty-first plenary session
Geneva, 10-12 June 2013
Item 4 (b) of the provisional agenda
How should national statistical offices respond - moving from risk avoidance to risk management

Innovative micro-data access - confidentiality on the fly

Note by the Australian Bureau of Statistics*

Summary

This paper addresses innovative approaches for the analysis of micro-data. It provides an overview of the key drivers that have led the Australian Bureau of Statistics to pursue these new techniques. It describes the suite of remote analysis servers and the 'on the fly' approach. The paper makes a brief overview of international approaches to the provision of access to micro-data and concludes with future research directions.

The Australian Bureau of Statistics has recognised the need to 'rethink' micro-data access to enable meeting user demand for access to more detailed unit record data, across a wider array of datasets. More flexible approaches are needed than just confidentialising the data prior to its release. Therefore the Australian Bureau of Statistics has developed a remote execution environment for micro-data access including a survey table builder and analysis service. Compared to traditional confidentialisation routines, the new technique minimises impacts on data quality by applying the necessary modifications only to the final output and not to the underlying data. All outputs are confidentialised 'on the fly' so as to protect data confidentiality while maintaining detail and quality.

* This document was submitted late due to delayed inputs from other sources.

Please recycle

I. Introduction

1. 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. The demand for gaining greater access to such micro-data has continued to grow since the 2003 Conference of European Statisticians (CES) session on the topic of confidentiality and micro-data access.

2. 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'. The ABS, like many other National Statistical Organisations (NSOs), has needed to 'rethink' micro-data access, faced with the challenge of balancing the trade-off between legal obligations to ensure that the likelihood of disclosing information about a particular person or organisation is unlikely, with releasing more detailed micro-data for informed decision making, research and discussion to benefit society.

3. 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). Given the increasing amount of available data via administrative and linked sources, facilitated through technological advances, the risk of disclosure for micro-data is arguably ever increasing.

4. There are a number of software products for confidentiality that are currently available. The majority of these are designed to be used by NSOs to confidentialise data before release. These can be either for tabular data (for example, Tau-Argus[1] and sdcTable[2]) or for micro-data (for example, the Special Uniques Detection Algorithm[3] and sdcMicro[4]). Many NSOs have also developed their own tailored processes and software specific to their legislative requirements.

5. The ABS has progressively developed a range of methods, processes and applications to produce a range of products to provide access to its statistical data including the release of statistic tabular data to the ABS's web site, the release of customised tabulations through an information referral service and through the analysis of pre-confidentialised micro-data files, referred to as Confidentialised Unit Record Files (CURFs). CURFs are made available either in the form of a Basic CURF for analysts to use in their own research environment, or as an Expanded CURF which is accessible through submitting queries remotely in the ABS Remote Access Data laboratory (RADL), or by visiting one of the on-site ABS Data Laboratories (ABSDL). Typically approved users

[1] Tau-Argus is a free software program designed to protect statistical tables, downloadable from: http://neon.vb.cbs.nl/casc/tau.htm.

[2] sdcTable is a free Statistical Data Limitation package for protecting tabular data, downloadable from: http://cran.r-project.org

[3] Special Uniques Detection Algorithm is a system for detecting and grading special uniques. This is needed for confidentialising datasets by first identifying all special unique records and either disguising or removing them.

[4] SDCMicro is free software for the generation of protected microdata for researchers and public use, downloadable from: http://cran.r-project.org/web/packages/sdcMicro/index.html

submit queries in SAS, STATA or SPSS[5]. In the case of RADL, the results of the queries are automatically checked and cleared outputs made available to users via their desktops. Additionally, a sample of queries is selected on an ongoing basis for manual inspection. For the ABSDL all outputs to be removed from the laboratory are cleared manually. The level of pre-confidentialisation applied to each CURF file is dependent on the method of dissemination, from the heavily confidentialised Basic CURF to the more detailed, less confidentialised Expanded CURFs. Considerable amounts of staff resources and time are required to produce a CURF in either form.

6. In response to a number of drivers of change, including the increased complexity of datasets where traditional approaches are likely to break down, the ABS has recently invested in the development of a suite of research applications that enable registered analysts to submit queries via the internet that are executed against the underlying micro-data and confidentialised in real-time 'on the fly' with confidentialised output returned to the analyst. Internationally, such applications are commonly referred to as Remote Analysis Servers (RASs). They provide users with control over the particular outputs they want to extract from a dataset. The challenge for the statistical offices is to provide SDC for the different possible outputs.

7. Section II of this paper provides an overview of the key drivers that have led the ABS to pursue new innovative approaches for the analysis of micro-data. Section III introduces the ABS suite of Remote Analysis Servers and the 'on the fly' approach to SDC. Section IV provides a brief overview of different International approaches to the provision of access to micro-data and Section V concludes with an overview of future research directions and opportunities for international collaboration.

II. Drivers for change

8. Many NSOs, including the ABS, are changing the way in which micro-data are disseminated to researchers. Key drivers for this change include:

(a) Increasing demands for better, more flexible and timelier access to detailed micro-data;

(b) Enhancing user experience through the provision of user friendly, menu-driven interfaces that are not reliant on users having statistical programming skills;

(c) Reducing costs associated with existing manual and resource intensive approaches to disseminating micro-data, such as the process of creating CURFs;

(d) Increasing the timeliness of access to micro-data (CURFs are currently available up to 6 months later than static tabulations are disseminated);

(e) Mitigating the increasing risk of disclosure as a result of increased computing power (both hardware and software), the increased volume of outputs disseminated and the increased accessibility of large datasets;

(f) Facilitating analysis of emerging sources of data (including transactional, administrative and integrated data) where traditional approaches to SDC are not sufficient to mitigate the increased identification risk;

[5] These general-purpose statistical software solutions are used in statistical production and analysis: SAS (Statistical Analysis System), STATA and SPSS (Statistical Package for the Social Sciences)

(g) A growing recognition that not all essential statistical assets are held by the NSOs, hence the need to develop methods and infrastructure that can be utilised by other organisations; and

(h) The changing model of the typical data analyst to organisations looking for increased accessibility of outputs through more efficient machine to machine querying, such as through the use of SDMX (Statistical Data and Metadata Exchange) Web services.

9. These drivers for change have led many NSOs, including the ABS, to commence development of sophisticated Remote Analysis Servers.

III. Australian Bureau of Statistics remote analysis servers – confidentiality on the fly

10. The suite of Remote analysis servers at the ABS enable the ABS to make the full detail of the dataset available while minimising the loss of utility through the application of SDC tailored to each specific output. From a user's perspective, they have control over the particular outputs they want to extract from a dataset. This is a fundamental shift in the process, from the traditional paradigm where the ABS would decide all the outputs that would be released to one where users can specify what they require when and as they need it.

11. There is a real risk of disclosure from the dissemination of tabular and analytical outputs that needs to be mitigated. A number of papers have been published on this subject, including proposed approaches to manage the disclosure risk. In respect to analysis output see Gomatam et. al (2005), Bleninger et. al (2011) 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, including graphics and model diagnostics, to reconstruct attributes for one or more records which, if successful, could be used to attempt identification. The challenge for the ABS is to provide SDC for the different possible outputs 'on the fly'.

12. A very simplistic model for a remote analysis server is:

(a) The agency makes available a micro-data file for researchers. It is held securely by the agency. The sensitive micro-data is typically not observable to the analyst;

(b) An analyst submits a query, via the internet, to the agencies analysis server which is processed against the sensitive micro-data;

(c) The statistical output (e.g. regression coefficients or tabulation) from the query is modified by a tailored confidentiality method specific to that analysis for the purpose of SDC;

(d) The analysis server sends the modified output, via the internet to the analyst. Some outputs may be restricted on the basis that they could allow an analyst to reconstruct the attributes of an arbitrary record.

13. The next generation remote analysis server at the ABS comprises three applications – Census TableBuilder, ABS TableBuilder and ABS DataAnalyser. Each of these is described in the following paragraphs. At no time does the analyst see the underlying micro-data.

14. For the 2006 Census of Population and Housing, the ABS jointly developed **Census TableBuilder** with Space-Time Research Pty Ltd. This is a web-based product that is available to analysts outside the ABS. Census TableBuilder incorporates a perturbation method (Fraser and Wooton (2005)) for automatically protecting tables of Census count

data. This method was designed to mitigate disclosure risks from 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 subpopulations, and to enable the development of web-based systems allowing users to define their own tables. 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 as well as the web-based Census TableBuilder for analysts.

15. Expanding from this basis, the ABS has continued to develop a range of customised confidentiality algorithms and protections that run on the fly and minimise utility loss, while ensuring that the outputs produced are not likely to enable the identification of a respondent. Each confidentiality method is tailored to each specific output, reducing the level of confidentialisation required. In comparison, the existing CURF approaches require heavier confidentiality to protect against the full range of potential outputs that could be produced.

16. **ABS TableBuilder**, the successor of Census TableBuilder, also developed jointly with Space Time Research Pty Ltd, incorporates dynamic confidentiality routines for weighted survey data that have expanded beyond population counts to key summary statistics from magnitude data (such as custom ranges, totals, means, medians and quantiles). In addition to the perturbation, a number of in-built protections/restrictions have also been incorporated. These include restricting combinations of data items from being tabulated together, restricting the output of sparse tables where there are a large number of small cells, and requiring a minimum population size for the calculation of medians and quantiles. ABS TableBuilder also facilitates machine to machine querying of the micro-data via an SDMX web service.

17. **ABS DataAnalyser** has been developed to facilitate exploratory data analysis and regression modelling. The ABS DataAnalyser is a secure menu based system for conducting statistical analysis through a remote user interface. The system enables users to remotely estimate the parameters of the statistical models fitted to ABS data while protecting the confidentiality of providers. All statistical outputs that can be viewed by the user are automatically confidentialised using various disclosure control methods, including the perturbation of the estimating equation. This perturbation alone is not enough and a set of restrictions and protections against specific attacks have also been incorporated in the system, including confidentialised graphical displays to assist users to diagnose model goodness of fit. A summary of the perturbation approach and additional protections is described in Chipperfield, Gare and Yu (2011). The initial release of ABS DataAnalyser enables users to undertake data transformation and manipulation, tabulation, exploratory data analysis and Linear Robust, Logistic, Probit, Poisson or Multinomial modelling. The system is currently being trialled with users and a full production release is planned for 3rd quarter 2013. Chart 1 provides an illustration of ABS Data Analyser.

Chart 1
Australian Bureau of Statistics DataAnalyser incorporating on the fly confidentiality

[redacted]

18. Advantages of the ABS approach to remote analysis servers include:

(a) The analysis is undertaken on the real micro-data, retaining complex relationships in the data;

(b) Statistical output is modified to a degree that is tailored specifically to the type of analysis being undertaken as well as the level to minimise information loss;

(c) Once the server is set up, it can process multiple analyses in real time;

(d) All submitted programs can be logged and audited and if an attempt at disclosure is identified the analyst's access to the server can be revoked; and

(e) The point and click menu interface means that the user requires little training and does not have to learn a new software language.

19. Disadvantages of the ABS approach include:

(a) The analyst is restricted to use only analysis techniques, data transformations and manipulations supported by the server;

(b) Analysis through the remote servers may take longer than if the micro-data were available on the analyst's personal computer; and

(c) The substantial investment of time and money required to develop confidentialisation software routines for each new analytical functionality.

IV. International approaches to providing access to micro-data

20. In terms of access to micro-data, National Statistical Organisations have taken very different approaches, largely driven by the differing legislative requirements. A number of NSOs release public use files. These files are heavily confidentialised for general use.

21. NSOs also make extensive use of Research Data Centres, similar to the ABSDL, for the analysis of detailed micro-data. Disadvantages of these approaches are that the outputs removed from these centres must generally be checked manually. This is resource intensive, leading to restrictions on the number of researchers that access can be provided too. The ABS seeks to provide a solution to facilitate as wide access as possible.

22. Some research centres are on-site while others are utilising technology advances to create a virtual research centre that can be accessed from dumb terminals installed in other organisations. The level of detail accessed in also varies greatly depending on the

legislation of the NSO. In some cases authorised or 'trusted' researches are provided with the same access to the detailed micro-data as the employees of the NSO. The ABS legislation prevents this in Australia.

23. In comparison, the ABS's web-based systems are designed for external users to access remotely. The approach does not require extensive confidentialisation of the micro-data prior to analysis, but utilises confidentiality methods that are applied in real-time. In recognition of the advantages of remote analysis servers that incorporate on the fly confidentiality, a number of NSOs have commenced research and development programs. Two fairly advanced developments worth noting are Morpheus (Höninger (2011)) developed by the State Statistical Institute Berlin-Brandenburg and the Microdata Analysis System (Lucero et. al (2011)) under development by the US Census Bureau.

V. Future directions and opportunities for international collaboration

24. The ABS remote analysis servers offer the advantage of providing high quality outputs derived from micro-data files, as well as the convenience of access by users, without compromising the confidentiality of the data. Many challenges still remain.

25. Current research is focusing on assessing the existing approaches and their efficacy for linked datasets, which pose a higher disclosure risk. Future research work will focus on the development of on the fly confidentiality methods for the dissemination of business and longitudinal datasets via remote analysis servers and the provision of utility loss measures to researchers (Marley and Leaver 2011).

26. The ABS is also very interested in exploring the potential of synthetic micro-data, even if only to allow analysts to test and review their models prior to running those on the real data held securely within the ABS DataAnalyser.

27. The ABS would be interested to work with the international statistical community to carry out research to address the challenges from these datasets. The ABS would also be willing to provide the statistical methods and algorithms behind those methods to the international statistical community if they want to incorporate these methods in their applications.

VI. References

Bleninger, P., Drechsler, J. and Ronning, G. 2011, ‘Remote Data Access and the Risk of Disclosure from Linear Regression: An Empirical Study’, *Privacy in Statistical Databases*, Springer. See http://www.idescat.cat/sort/sortspecial2011/DataPrivacy.1.bleninger-etal.pdf),

Chipperfield, J., Gare, M. & Yu, F. 2011, ‘Providing access to microdata for statistical purposes - experiences of the Australian Bureau of Statistics with Remote Analysis Servers’, paper presented to the Statistics Canada 2011 Methodology Symposium, Ottawa, Canada, 1-4 November.

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.

Gomatam, S., Karr, A. F., Reiter, J. P., & Sanil, A. P. 2005, ‘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.

Höninger, J. 2011, ‘An Innovative Approach to Remote Data Access’, 58th International Statistical Institute World Statistics Congress, Dublin, Ireland, 21-26 Aug 2011

Lucero, J., Zayatz, L., Singh, L., You, J., DePersio, M. and Freiman, M. 2011, ‘The Current Stage of the Microdata Analysis System at the U.S. Census Bureau’, 58th International Statistical Institute World Statistics Congress, Dublin, Ireland, 21-26 Aug 2011

Marley, J. & Leaver, V. 2011, ‘A method for confidentialising user-defined tables: statistical properties and a risk-utility analysis’, paper presented to the International Statistical Institute session, Dublin, Republic of Ireland, 22-26 August.

Shlomo, N. 2007, ‘Statistical disclosure control methods for census frequency tables’, International Statistical Review, 75, (2), 199-217.

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.

Willenborg. L. and de Waal, T. 2001, ‘Elements of Disclosure Control’, *Lecture Notes in Statistics*, Vol 155, ISBN 978-0-387-95121-8, Springer.