Working Paper
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
(Ottawa, Canada, 28-30 October 2013)

Topic (i): New methods for protection of tabular data or for other types of results from table and analysis servers

Protecting confidentiality in statistical analysis outputs from a virtual data centre

Prepared by Christine M. O'Keefe, Mark Westcott, Adrien Ickowicz, Maree O'Sullivan and Tim Churches, Australia

Protecting confidentiality in statistical analysis outputs from a virtual data centre

Christine M. O'Keefe*, Mark Westcott*, Adrien Ickowicz**, Maree O'Sullivan** and Tim Churches***

* CSIRO Computational Informatics, GPO Box 664, Canberra ACT 2601 AUSTRALIA
Christine.OKeefe@csiro.au, Mark.Westcott@csiro.au
** CSIRO Computational Informatics, Locked Bag 17, North Ryde NSW 1670 AUSTRALIA
Adrien.Ickowicz@csiro.au, Maree.OSullivan@csiro.au
*** Sax Institute, PO Box K617, Haymarket NSW 1240 AUSTRALIA
Tim.Churches@saxinstitute.org.au

Abstract: In this paper we are concerned with protecting confidentiality in statistical analysis outputs from a virtual data centre. This is an increasingly popular approach in which data are held in a secure environment and are made available to a researcher over a secure internet connection. The researcher has unrestricted access to the data, which is regulated by applicable legislation and researcher agreements. Current systems generally rely on expert manual checking of analysis outputs and/or confidentiality requirements in the applicable legislation and researcher agreements.
We believe that a desirable, though potentially interim, solution is to train researchers to conduct the output confidentialisation themselves, while recognising that they will probably not be experts in confidentiality protection methods. Eventually automated systems for output confidentialisation may become available.
In this paper we describe a proposal for a two-stage process involving:

- Dataset preparation by the data custodian before loading the data into the virtual data centre
- Confidentialisation of the analysis outputs by the researcher on removal from the secure environment

The second stage makes use of a checklist developed to assist researchers. However, it would be essential to provide researchers with training in disclosure control as part of the virtual data centre researcher and project approval process.

1 Introduction

The challenge of balancing the competing objectives of allowing statistical analysis of confidential data and maintaining confidentiality is of great interest to national statistical agencies and other data custodians seeking to make their data available for research. Of particular interest is the accelerating trend for healthcare organisations to adopt and adapt information technologies to support an expanding array of activities designed to derive value from their growing administrative, clinical and molecular data archives, in terms of research leading to enhanced health outcomes.

Health information is generally considered to be amongst an individual's most sensitive and private information, and approaches to balancing research use with confidentiality protection usually involve a combination of:

- Compliance with privacy legislation and regulation

- Restrictions on access
- Restrictions on the amount and detail of data available
- Statistical disclosure control methods applied to the data before release
- Enabling access through secure physical and virtual data centres, as well as remote analysis systems. These approaches can involve restrictions on the range and nature of allowable analyses

In this paper we are concerned with the increasingly popular virtual data centres, in which data are held in a secure environment and are made available to a researcher over a secure internet connection. Normally in a virtual data centre the researcher has unrestricted access to the data, unlike in a remote analysis system where the researcher cannot view the data directly at all. A virtual data centre provides good confidentiality protection during researcher access and use. There are, however, still issues of confidentiality associated with statistical analysis outputs that researchers may wish to remove from the secure environment and publish in the academic literature, since such outputs cannot be assumed to be free from disclosure risk.

1.1 Conceptual Model of a Virtual Data Centre

The model has four stages, represented by four outer boxes in Figure 1.

- Dataset stage: The custodian agency prepares a dataset, and makes it available through the secure analysis laboratory.
- Data Transformations stage: The researcher may apply data transformations including data subsetting, new variable creation and variable transformations.
- Query stage: The researcher submits an analysis to be run on the dataset, and results are generated. The sub-stages of the Query stage are:
    - Analysis stage: The researcher submits an analysis request
    - Output stage: The analysis results are viewed on the computer screen
- Output for Publication stage: The output is prepared for inclusion in an academic publication, presentation, report or other dissemination channel.

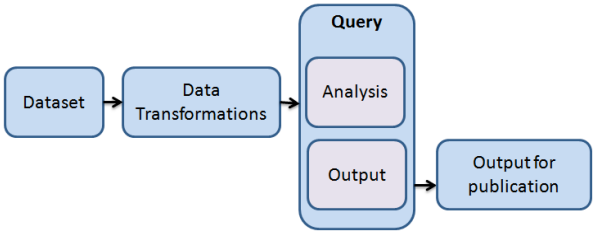

Figure 1. Conceptual model of a secure analysis laboratory (O'Keefe and Chipperfield 2013)

1.2 Examples of Virtual Data Centres

The UK Data Service (ukdataservice.ac.uk) provides a single point of researcher access to a wide range of anonymised secondary data including large-scale government surveys, international macrodata, business microdata, qualitative studies and census data from 1971 to 2011. The Service provides secure access to data of high sensitivity where specialised staff apply statistical disclosure control techniques to ensure the delivery of safe statistical results.

The Secure Anonymised Information Linkage (SAIL) Databank (www.saildatabank.com) links together a wide range of person-based data using robust anonymisation techniques and makes it available for research. All statistical outputs are manually reviewed by experts for potential disclosures.

The National Opinion Research Center (NORC) hosts a secure Data Enclave (www.dataenclave.com) offering secure, remote researcher access services. Researchers cannot move files in or out of the secure environment without review approval by NORC statisticians.

The Secure Unified Research Environment (SURE) (www.sure.org.au) is a remote-access computing environment that allows researchers to access and analyse linked health-related data files for approved studies in Australia. It is part of the Population Health Research Network initiative (www.phrn.org.au). Outbound files intended for use outside of SURE are reviewed by the study's chief investigator or an alternative senior investigator before being released.

1.3 Assumptions for a Virtual Data Centre

In order to clarify the difference between a virtual data centre and a remote analysis system, we propose some assumptions.

A.1 Data custodians prepare the datasets to be compliant with all applicable legislation, regulation and assurances given to data provider organisations.

A.2 The researcher will comply with the applicable researcher agreements. Thus, we do not need to protect the dataset records from the researcher. We assume that the researcher is fully authorised to view the dataset records, and we only need to consider confidentiality in the context of readers of the academic literature. Also, we do not need to consider malicious confidentiality attacks by the researcher, such as massively repeated regressions or regressions designed to reveal response variable values, but only confidentiality risks due to the release of outputs from genuine queries.

A.3 The researcher can apply unrestricted data transformations and analyses.

1.4 Case Study: Survival Analysis Outputs

In this section we give a case study to further demonstrate the difference between the operational models of a virtual data centre and a remote analysis system, and in particular the consequences of assumption A.2.

Our case study starts with the detailed discussion of confidentialising the output from a survival analysis (O'Keefe et al, 2012), and adapts the confidentiality protection measures to the context of a virtual data centre.

To confidentialise Kaplan-Meier Output in a remote analysis system:

- Suppress the symbols on the survival plot indicating study censoring events. These are event dates which could be used to identify individuals when linked to other databases such as surgery rosters.
- Smooth the survival plot and the confidence interval limit plots, in order to conceal death times. The times reveal dates which could be used to identify individuals when linked to hospital death records.

Adaptation: These are reasonable precautions under either operational model, since they protect confidentiality in statistical research outputs.

To confidentialise Model Selection in a remote analysis system:

- Conduct each analysis on a random sample of 95% of the observations.
- Allow a factor to be included in the model only if each level is observed for at least a minimum threshold value of data items.

Adaptation: The first point is unnecessary in a virtual data center, since it aims to protect against attacks exploiting algebraic relationships in the data, which generally require a large number of related analyses. The second point is a specific instance of well-known issues with small counts, and is applicable to statistical research outputs in either scenario.

2 Proposed Approach

As seen in Section 1.2, currently virtual data centres rely on manual checking for confidentiality protection. This solution is expensive and time-consuming, and may not be feasible given the long term trend of rising researcher demand. On the other hand, Duncan et al (2012) remark that "…developing valid output checking processes that are automated is an open research question". An intermediate solution is to train researchers to conduct the output confidentialisation themselves, while recognising that they will probably not be discipline experts. This is the approach proposed by SURE.

When addressing confidentialisation of the output of statistical analysis in preparation for publication, arguably the most helpful available sources are the sets of guidelines developed for expert manual checking of outputs of statistical analysis conducted in on-site and virtual data laboratories, including:

- Statistics New Zealand Data Lab Output Guide, Version 3.0 (StatsNZ) (Statistics New Zealand 2011).
- ESSNet SDC (EuroStat) Guidelines for the checking of output based on microdata research (ESSNet) (see Hundepool et al 2012).

The Statistics New Zealand Guide specifically addresses confidentialisation of outputs for publication and presentation.

The information in these guidelines can be augmented with confidentiality protection measures from the remote analysis literature, see the recent summary by O'Keefe and Chipperfield (2013) including Gomatam et al (2005) and Sparks et al (2008).

The confidentiality protection measures currently suggested for virtual data centres and remote analysis system almost universally include a data preparation stage and an output confidentialisation stage. Remote analysis systems required further stages.

We first developed a list of the main ways that disclosures can occur in statistical analysis output (Section 2.1). We then synthesised the applicable confidentiality protection measures into data preparation guidelines for custodians (Section 2.2), as well as a "Checklist" designed to be used by researchers who are not expert in statistical disclosure control (Section 2.3). The Checklist is designed to enable a researcher to assess disclosure risks in their analysis outputs, and apply confidentialisation treatments to reduce the risks to acceptable levels.

2.1 Main Disclosure Risks in Statistical Output

The main ways in which disclosure can potentially occur in statistical analysis outputs are:

- Individual data: Individual data values are always potentially disclosive. These can be quoted directly or implied by some other output; for example, the jump points in an empirical CDF plot.
- Threshold: A statistic computed on a small number of records is always potentially disclosive. This includes the familiar case of small cell counts, but also includes other statistics such as means, modes or regression coefficients.
- Dominance: A statistic computed on a number of records where one is dominant is always potentially disclosive. This includes small cell counts, but also includes other statistics such as means, modes or regression coefficients.

- Differencing: Comparison of values of the same statistic on two samples which differ in a small number of records is always potentially disclosive.
- Linear (or algebraic) relationships: These can be exploited for disclosure.
- Precision: More significant figures/decimal places raise disclosure risk.

In each case, the presence of a potential disclosure risk does not always lead to a disclosure. For example, consider a statistic which is computed on a small number of counts. This is certainly a potential disclosure risk, however it would be important to consider the actual formula for the statistic, the variables involved and the data environment (what other datasets would be likely to be available) before assessing the likelihood that the potential disclosure risk would be realised in a disclosure.

2.2 Dataset Preparation by Custodians

Although the aim is to supply the researcher with the most detailed and unmodified data possible, it is still usually necessary to apply some basic confidentialisation measures to the dataset before making it available through a remote analysis server. A recent review of the literature (O'Keefe and Chipperfield 2013) has found a good degree of consistency about confidentiality protection in the dataset preparation stage, namely:

- Removing obvious identifiers including names, addresses, dates, email addresses, licence numbers, as well as biometric identifiers and IP addresses
- Ensuring datasets have sufficient records
- Ensuring published datasets differ by sufficiently many records
- Ensuring variables and combinations of variables have sufficient records
- Reducing detail in data (especially dates and locations) using data aggregation
- Applying other statistical disclosure control methods, such as data swapping or adding noise

2.3 Statistical Analysis Output Confidentialisation

In the case of confidentialisation of the output of statistical analysis in preparation for publication, none of the approaches described in the literature are directly applicable for researchers to use in virtual data centres. We used the summary of the main ways in which disclosure can occur in statistical analysis outputs (Section 2.1 above) to develop a list of the most common confidentiality risks, as follows:

- Individual value: an individual data value is directly revealed
- Threshold *n*: A cell or statistic is calculated on fewer than *n* data values

- Threshold *p%*: A cell contains more than *p%* of the values in a table margin
- Dominance *(n,k)*: Amongst the records used to calculate a cell value or statistic, the *n* largest account for at least *k%* of the value
- Dominance *p%*: Amongst the records used to calculate a cell value or statistic, the total minus the two largest values is less than *p%* of the largest value
- Differencing: Two cells or statistics are calculated on populations that differ in fewer than *n* records
- Relationships: The statistic involves linear or other algebraic relationships
- Precision: The output involves a high level of precision in terms of significant figures and/or decimal places
- Degrees of Freedom: The model output has fewer than *n* degrees of freedom.

We then developed a checklist which assists the researcher to identify potentially disclosive output for each class of statistical outputs, through the application of the relevant tests. The checklist also includes suggested treatments designed to reduce the identified disclosure risk.

Unfortunately the space is too limited here to present the full checklist, which will be included in a forthcoming publication. Table 1 shows excerpts from the checklist for some common published outputs of statistical analysis on linked health data.

Note that some rules have parameters associated with them (eg *n* for threshold rules), particularly those for tables. There has been a range of values of the various parameters suggested in the literature, and appropriate values should be chosen by the custodian and virtual data centre administrator.

An important benefit of this checklist approach is that researchers can ensure that the confidentialisation measures applied do not adversely impact the statistical inferences and conclusions drawn.

3 Conclusion

In this paper we are concerned with the increasingly popular virtual data centres, in which data are held in a secure environment and are made available to a researcher over a secure internet connection. Normally in a virtual data centre the researcher has unrestricted access to the data, unlike in a remote anlaysis system where the researcher cannot view the data directly at all.

We have argued that a desirable, though potentially interim, solution is to train researchers to conduct the output confidentialisation themselves, while recognising that they will probably not be discipline experts. Eventually automated systems for

Statistic	Confidentiality Test	Treatment
Number	Threshold *n*	• Suppress number • Try to get more data
Ratio, percentage	Individual value	• Suppress number • Round or perturb the number
	Threshold *n*	• Recode variables • Amalgamate groups over which statistic is calculated • Suppress value
	Threshold *p%*	
	Dominance *(n,k)*	
	Dominance *p*%	
	Differencing	• Redefine one or both populations
	Relationships	• Round or perturb reported values
	Precision	• Reduce precision of reported values
Mean	Threshold *n*	• Do not report denominator • Amalgamate groups over which mean is taken • Suppress value
	Dominance *(n,k)*	
	Dominance *p*%	
	Differencing	• Redefine one or both populations
Odds ratio, Relative risk	Threshold *n*	• Recode variables • Amalgamate groups
	Precision	• Reduce precision of reported value
Confidence Interval	Degrees of freedom	• Change model or data groups to increase degrees of freedom
	Threshold *n*	• Recode variables • Amalgamate groups
	Precision	• Reduce precision of reported values
p-value	Precision	• Reduce precision of reported value
Kaplan-Meier plot	Individual value	• Do not show individual values, for example by smoothing the plot • Recode variables • Amalgamate groups or ranges

Table 1. Excerpts from the researchers' disclosure control checklist for use in a Virtual Data Centre. The full checklist will be presented in a forthcoming publication.

output confidentialisation may become available. This is unlike the current situation in most existing on-site and virtual data centres which use manual output checking by disclosure control experts.

After presenting a conceptual model for a virtual data centre and briefly indicating some examples, we described assumptions and presented a case study of confidentialising survival analysis output.

We then presented a proposal for a two-stage process involving:

- Dataset preparation conducted by the data custodian before loading the data into the virtual data centre

- Disclosure risk assessment and confidentialisation of the statistical analysis outputs by the researcher before removal from the secure environment

We introduced the concept of a checklist developed to assist researchers in assessing disclosure risk and confidentialising statistical analysis outputs. An important benefit of this checklist approach is that researchers can ensure that the confidentialisation measures applied do not adversely impact the statistical inferences and conclusions drawn. It would be essential to provide researchers with training in disclosure control as part of the virtual data centre researcher and project approval process.

We remark that public health and medical journals are increasingly encouraging authors to submit supplementary materials including datasets and additional analysis results for online publication. The ease of online publication is likely to increase both the volume and the detail of analysis results publicly available, which in turn is likely to increase disclosure risk. We plan to consider the applicability of the checklist approach to confidentiality issues in this less-controlled scenario.

Acknowledgements:

This work is the result of a collaborative project between the Sax Institute and CSIRO, to examine approaches to confidentialising statistical analysis outputs in the Secure Unified Research Environment (SURE). We thank the Australian Government Education Investment Fund (EIF) Super Science Initiative for providing part funding through the Population Health Research Network.
The authors acknowledge the valuable comments and suggestions of Joanna Khoo (Sax Institute), Chris Okugami (CSIRO) and Daniel Elazar and Joseph Chien (Australian Bureau of Statistics).

References

Duncan, G.T., Elliot, M. and Salazar-González, J.-J. (2012) Statistical Confidentiality. Principles and Practice. Springer: New York

Gomatam, S., Karr, A., Reiter, J. and Sanil, A. (2005) Data dissemination and disclosure limitation in a world without microdata: A risk-utility framework for remote access systems, Stat Sci, 20, 163-177.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E., Spicer, K. and de Wolf, P.-P. (2012) Statistical Disclosure Control, Wiley Series in Survey Methodology, United Kingdom: John Wiley & Sons.

O'Keefe, C.M. and Chipperfield, J.O. (2013) A summary of attack methods and options for protective measures for fully automated remote analysis systems, Int Stat Rev, in press.

O'Keefe, C.M., Sparks, R., McAullay, D. and Loong, B. (2012) Confidentialising survival analysis output in a remote data access system, J Priv Confid,3 4, 127-154.

Sparks, R., Carter, C., Donnelly, J., O'Keefe, C., Duncan, J., Keighley, T. and McAullay, D. (2008) Remote Access Methods for Exploratory Data Analysis and Statistical Modelling: Privacy-Preserving AnalyticsTM", Comput Meth Prog Bio, 91, 208-222.

Statistics New Zealand (2011) Data Lab Output Guide Version 3.0. Wellington: Statistics New Zealand. http://www.stats.govt.nz/tools_and_services/services/microdata-access/~/media/Statistics_/services_/microdata-access/data-lab/datalab-output-guide-v3.pdf