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 (iv): The trade-off between quality, utility and privacy

Consistency of Output from Remote Access Servers, Microdata, and Tabular Data Release: The Need for Common Performance Metrics

Prepared by Krish Muralidhar[*], Rathindra Sarathy[**], and Mario Trottini[***]

[*] Krish Muralidhar, University of Kentucky, Lexington KY, USA; krishm@uky.edu
[**] Rathindra Sarathy, Oklahoma State University, Stillwater OK, USA; rathin.sarathy@okstate.edu
[***] Mario Trottini, University of Alicante, Alicante, Spain; mario.trottini@ua.es

Consistency of Output from Remote Access Servers, Microdata, and Tabular Data Release: The Need for Common Performance Metrics

Krish Muralidhar[*], Rathindra Sarathy[**], and Mario Trottini[***]

[*] Krish Muralidhar, University of Kentucky, Lexington KY, USA; krishm@uky.edu
[**] Rathindra Sarathy, Oklahoma State University, Stillwater OK, USA; rathin.sarathy@okstate.edu
[***] Mario Trottini, University of Alicante, Alicante, Spain; mario.trottini@ua.es

1 Introduction

National statistical and other government agencies gather extensive data regarding individuals and organizations. While the primary purpose of this is typically to assist policy and decision makers within the government, this data also serves as an important source of information to the general public. In the past, agencies typically provided data primarily in two forms: microdata and tabular data. In recent years, there has also been considerable increase in the demand from the public for greater access to data residing with agencies. This has led to a rise in the implementation of what are often referred to as Remote Analysis Systems (RAS) where users are permitted to analyze the data residing in a secure environment and the results of the analysis are provided to the user (O'Keefe and Chipperfield, in press, Simard 2011).

Given several different modes of access to the data, the need for maintaining consistency between the outputs generated by the three systems is obvious. From the agency perspective, the first reason to maintain consistency between the two types of data is that data is released to ensure that the data released in one mode does not result in compromising confidential information released in a different mode. From a statistical disclosure limitation perspective, it is possible that the different methods employed to protect the data are indeed the best methods *individually*. But, as we discuss later, what is adequate to protect tabular data release may not necessarily be adequate for microdata release, and vice versa. RAS only makes matters more difficult since it is likely that users will be allowed to issue multiple queries to the system. Given microdata, an intelligent user may be able to compromise individual records by using targeted queries to the RAS.

Second, maintaining consistency is necessary in order to convince the public that the statistical disclosure methods employed are effective both in providing meaningful data for analysis purposes and in preventing disclosure of confidential information. It is understandable that government agencies focus on disclosure as their primary consideration, but is also understandable that users will focus primary on data utility. Many government agencies may soon be using all three modes of presentation. And the same result could be obtained in all three modes of access. If different methods

are used to protect the data in each mode of presentation, then it is possible that a comparison of the two releases could lead to inconsistencies in the released data. In this case, the public is likely to mistrust the entire data release. Hence, there is a need to ensure that when different forms of the same data are released, that the entire data release is consistent when analyzed. Several articles also comment on the relative advantages and dis-advantages of the two approaches (Massell et al. 2006, Massell and Funk 2007, Trottini et al. 2012, Zayatz 2007). Giessing (2011) also describes an experiment by Hohne (2008) along these lines as well.

2 Important Issues to Consider

2.1 Disclosure scenario, and the balance between disclosure risk and information loss

A precise definition of the disclosure scenario is a critical aspect when the agency plans to use tabular, microdata, and responses to analytical queries releases for a collected data set. What should be considered disclosure as well as the intruder's prior information should be independent of the type of the release (although both can depend on the type of access). Failing to consider this leads to inconsistencies that are difficult to justify in practice.

Even assuming that a common disclosure scenario has been defined, still the specification of global measures of information loss and disclosure risk remains a very difficult problem. From the disclosure risk side, for a given interpretation of disclosure we should be able to define: (i) Risk measures both for microdata and tabular release that are consistent with the assumed notion of disclosure; and (ii) A global measure of disclosure, which in addition to the components defined in (i), should take into account the possibility that an intruder might combine the information in the published microdata, with the information in the published tables, and responses to analytical queries.

The definition of risk measures for microdata, tabular release, and responses to analytical queries that are consistent with the assumed notion of disclosure is not a trivial problem. With respect to information loss, a global measure should take into account not only the extent to which the masked microdata, tabular release, and response to analytical queries "differ" from their corresponding versions based on the original data but also the extent to which the information on the two data release is consistent.

2.2 Over-protection

Trottini et al (2012) showed that when disclosure risk assessment is performed from the traditional tabular data perspective, the masked microdata, in general, performs well in preventing disclosure of sensitive cells. The results however, are not quite as encouraging for information loss. When information loss is evaluated from the

traditional tabular data perspective, our results indicate that model based microdata masking techniques tend to protect even safe cells resulting in information loss. Even for sensitive cells, in some cases, the level of masking is quite high, resulting in relatively high information loss. Similar results were found by Massel et al. (2006) using additive noise as masking method for the underlying microdata. Massell and Funk (2007) described a possible modification of the additive noise technique called *Balanced Noise* aimed at reducing the β error (and thus increasing data utility). Research on the same line should be done to design modified version of data shuffling (and more in general of model based masking methods) that could preserve the good performance of these microdata masking in terms of disclosure risk while reducing the information loss associated with the tabular release. The inclusion of RAS that may provide the ability to perform many different types of analyses complicates the issue immensely.

2.3 Distributional Characteristics of the Data

A third issue that is of importance deals with the distributional characteristics of the data. It is well understood that this is of critical importance for microdata since changes in distributional characteristics can have a detrimental impact on statistical analyses performed on microdata. For magnitude tabular data however, information loss is often characterized only by the change in the magnitude of the table value that is released. Giessing (2011) has recognized that distributional characteristics can have a significant impact on magnitude tabular data. With RAS, the distributional characteristics of the data will dictate the type of analysis that can be performed. For example, when a particular variable in the data set is heavily skewed, it would be necessary to consider transformations of the variable when performing regression analysis. If this is the case, then microdata that may be released should be consistent with such transformations as well.

2.4 Input versus Output Masking

In the most general terms, the question of microdata masking and masking the output from tabular results and responses to analytical queries can be classified under the more general issue of input versus output masking mechanisms. Input masking mechanisms modify the individual data records and respond to all queries using the masked data. With output masking mechanisms, the response to the query is computed using the original data and the response is modified to generate the actual output to the user. There is a need to investigate the relatively benefits of both approaches. This issue gains even more importance for RAS.

3 Flexibility

In this section, we briefly attempt to identify the most flexible approaches for access and masking. We acknowledge that our evaluation was based only on flexibility.

3.1 Access Mode

In terms of access mode, there is little question that RAS is the most flexible. Using RAS, it would be possible to provide users with the ability to access analytical results for any allowable analysis. Hence, tabular results simply represent just one type of analysis and can be incorporated in RAS easily. In terms of microdata, there is nothing which prevents the administrator from providing the source microdata to the users. Obviously, providing the source microdata has considerable disclosure risk implications, especially if the source data has not been modified. From the viewpoint of being able to provide access to the data however, the RAS can be easily designed to provide such access.

Using RAS as the general system offers an additional advantage. Model based masking approaches provide excellent data utility characteristics for specific types of analyses. But they are often criticized for the fact that when users perform analyze the data using techniques that are not appropriate for the assumptions of the model based masking approach, the results of such analyses can be misleading (if not completely incorrect). The underlying distribution of the original data may also restrict the types of analyses that can be performed. Using RAS would mean that the data administrator would be able to limit the types of analyses that can be performed on the data. Recent research has also identified how the data administrator can protect the data for output from RAS systems using simple rules for the types of analyses, data sub-setting, and limitations on the output. For a comprehensive summary of these restrictions, see O'Keefe and Chipperfield (in press). Implementing these rules provides the data administrator with the ability to further limit disclosure in addition to masking techniques.

3.2 Masking Mechanism

In terms of the masking mechanism, input masking (or masking the source data) offers the greatest level of flexibility to the data administrator. Given the type of analyses that can be performed on the data, the data administrator can evaluate the results of such analyses and determine the risk of disclosure that could occur from the data. If such an evaluation does not satisfy the disclosure risk requirements for a particular type of analyses, the data administrator has the ability to modify the masking to produce a new data set. Obviously, it would be impossible for the data administrator to identify every type of analysis on every possible subset of the data. However, source masking allows the data administrator to identify worst case disclosure risk for any type of analysis. For instance, the recent results by Dinur and Nissim (2003) has identified that responding to queries using the original data could potentially allow an adversary to accurately reconstruct the database. If source masking is used and all responses are provided from the masked data, then the best the adversary can hope to do is to reconstruct the masked data (Sarathy and Muralidhar 2011). In this case, the data administrator has the ability to ensure that the masked data is different enough from the original data so that even if the

adversary is able to reconstruct the masked data, the disclosure is limited to some acceptable level.

4 Conclusions

In this study we have attempted to emphasize the importance of maintaining consistency between types of access modes. Failure to do so may result in disclosure of confidential information and loss of confidence among the users. Neither is desirable. Unfortunately, maintaining consistency is not an easy problem. We identify important issues relating to maintaining consistency among the output. Finally, we recommend the RAS as the most flexible access mode and source masking as the most flexible masking mechanism. A combination of the two may provide the data administrator with the greatest flexibility to ensure appropriate level of data utility and confidentiality protection. We do realize that in many cases this decision is based on policies and, at least in some cases, legal requirements. We hope that policy makers at the agencies will have a serious discussion on the implications of the different access modes and masking mechanisms. In the long run, we believe that an integrated single system with a single data source offers the best solution to protect the confidentiality of the sensitive data while simultaneously providing the users with meaningful results.

References

Dinur, I. and K. Nissim (2003) Revealing Information While Preserving Privacy, PODS 2003, June 9-12, San Diego, CA.

Giessing, S. (2011) Post-tabular Stochastic Noise to Protect Skewed Business Data, UNECE Work session on Statistical Disclosure Limitation, October 26-28, Tarragona, Spain.

Hohne, J. (2008) Anonymisierungsverfahren f¨ur Paneldaten. In: Wirtschafts- und Sozialstatistisches Archiv., Bd. 2, pp. 259–275. Springer-Verlag Berlin Heidelberg.

Massell, P., Funk, J. (2007) Protecting the Confidentiality of Tables by Adding Noise to the Underlying Microdata. In: Proceedings of the 2007 Third International Conference on Establishment Surveys (ICES-III), Montreal, Canada, June 18-21.

Massel, P., L. Zayatz and J. Funk (2006) Protecting the confidentiality of tables by adding noise to the underlying Microdata: Application to the Commodity Flow Survey, J. Domingo-Ferrer and L. Franconi (Eds.): PSD 2006, LNCS 4302, pp. 304 – 317, Springer-Verlag Berlin Heidelberg.

Sarathy, R. and K. Muralidhar (2011) “Evaluating the Characteristics of Input and Output Perturbation in the Context of Dinur-Nissim Disclosure Results,” 4th IAB

Workshop on Confidentiality and Disclosure - SDC for Microdata, Nuremberg, Germany.

O'Keefe, C.M. and Chipperfield, J.O. (in press) A Summary of Attack Methods and Options for protective measures for Fully Automated Remote Analysis Systems, International Statistical Review.

Simard, M. (2011) Progress with Real Time Remote Access, UNECE work session on Statistical Disclosure Limitation, October 26-28, Tarragona, Spain.

Trottini, M. and K. Muralidhar, and R. Sarathy (2012) An Investigation of Model-Based Microdata Masking for Magnitude Tabular Data Release, J. Domingo-Ferrer and I. Tinnirello (Eds.): PSD 2012, LNCS 7556, pp. 47–62, Springer-Verlag Berlin Heidelberg.

Zayatz (2007) New implementations of Noise for tabular Magnitude Data, Synthetic tabular frequencies and Microdata, and a Remote Microdata Analysis System, Statistics#2007-17, Research Report Series, US Census Bureau.