eurostat
Methodologies and
Working papers

Work session on statistical data confidentiality

Manchester 17-19 December 2007

2009 edition

eurostat
EUROPEAN COMMISSION

eurostat
Methodologies and
Working papers

Work session on statistical data confidentiality

Manchester 17-19 December 2007

2009 edition

eurostat
EUROPEAN COMMISSION

How to obtain EU publications

Our priced publications are available from EU Bookshop (http://bookshop.europa.eu), where you can place an order with the sales agent of your choice.

The Publications Office has a worldwide network of sales agents. You can obtain their contact details by sending a fax to (352) 29 29-42758.

Europe Direct is a service to help you find answers to your questions about the European Union

Freephone number (*):

00 800 6 7 8 9 10 11

(*) Certain mobile telephone operators do not allow access to 00 800 numbers or these calls may be billed.

More information on the European Union is available on the Internet (http://europa.eu).

Luxembourg: Office for Official Publications of the European Communities, 2009

ISBN 978-92-79-12055-8
Cat. No. KS-78-09-723-EN-N

Theme: General and regional statictics
Collection: Methodologies and working papers

Table of contents

Foreword

Data collected for official statistics contain much more information than what is normally published by a statistical organisation. Collectors and producers of data face two substantial challenges: (1) How can information derived from vast streams of data on human beings be used without violating statistical confidentiality? (2) The use of sound scientific methods: how can we best provide and promote access to rich and sensitive data so that empirically significant results can be generated?

Academic researchers are key players in this work, often conducting policy-relevant research on behalf of governments but also providing an independent perspective on policy. Close partnerships between official statistics and research communities are therefore essential. A strong scientific research input is necessary in order to design best practices and inform discussions on legal issues at both the national and international levels. The United Nations Principles and Guidelines "Managing statistical confidentiality and microdata access" adopted in 2006 constituted a big step forward towards international agreement on common principles for dissemination of microdata.

This Work Session was an excellent opportunity for statisticians and researchers to exchange ideas and discuss methods and tools dealing with confidentiality at the national and international levels.

The agenda of the work session consisted of the following substantive topics:

- Microdata;
- Tabular data protection;
- Applications (including practical implementation of statistical disclosure control methods, actual issues within national statistical institutes, and software).

Papers presented under topic (i) focused on the following aspects of microdata: crypto methods for database privacy, survey-specific anonymisation, new methodology for microdata protection, disclosure risk assessment and comparisons of methods.

In topic (ii), the discussion focused on tabular data protection. The topics covered in the presentations within this session centred on frequency tables, magnitude tables and web access.

The presentations under topic (iii) focused on development of new procedures, software applications and their practical implementation in statistical organisations: methods and applications for disclosure control of micro- and tabular data as well as development of strategies for statistical disclosure control.

Debates were also organised, centred on the following themes:

- Panel discussion on microdata protection and remote access facilities;
- Panel discussion on balancing data quality and confidentiality.

The work session was a successful event allowing official statisticians and researchers to discuss various issues related to statistical confidentiality and exchange ideas in this important domain of statistics.

Pedro Díaz Muñoz
Director,
Statistical methods and tools; Dissemination
Eurostat

Heinrich Brüngger
Director,
Statistical Division
UNECE

Acknowledgements

The work session on statistical confidentiality was jointly organised by Mr. Mark Elliot for the University of Manchester, Ms. Jane Longhurst from the Office for National Statistics (ONS, United Kingdom), Ms. Maria João Santos from the Eurostat Unit for Methodology and Research and Mr. Juraj Riecan from the United Nations Economic Commission for Europe (UNECE).

UNECE and Eurostat gratefully acknowledge the valuable contributions of all the participants and particularly the session organisers and discussants, who were the following: Mr. Alistair Ulph (Dean Faculty of Humanities, University of Manchester), Mr. Frank Nolan (ONS), Mr. Rainer Muthmann (Eurostat), Mr. Juraj Riecan UNECE, Ms. Sarah Giessing (Germany), Ms. Jane Longhurst (United Kingdom), Mr. Josep Domingo-Ferrer (Spain), Mr. Lawrence H. Cox (United States of America), Mr. Eric Schulte Nordholt (Netherlands), Mr. Paul J. Jackson (United Kingdom), Ms. Luisa Franconi (Italy), Mr. Anco Hundepool (The Netherlands).

Special thanks to the University of Manchester (especially to Mr. Mark Elliot and Mr. Kingsley Purdam), and to the Office for National Statistics (especially Ms. Jane Longhurst) for hosting and organizing the event, and to Mr. Anco Hundepool for chairing the session.

The proceedings have been produced by Ms. Martine Peeters from the Eurostat Unit for Methodology and Research.

I

Microdata

Community Innovation Survey: comparable dissemination

Luisa Franconi and Daniela Ichim

Istituto Nazionale di Statistica, via C. Balbo 16, Rome, Italy.
(franconi@istat.it, ichim@istat.it)

Abstract. The European Union is facing the problem of releasing microdata in a multinational setting i.e. microdata stemming from twenty seven member states. Different laws, methodologies, practices and cultural approaches to confidentiality may severely limit the possibility of obtaining comprehensive anonymised data sets. We recall the approach adopted in Europe to design and manage harmonised statistical surveys and claim that such an approach could be successfully applied in the release phase of a survey as well. An application to the Community Innovation Survey is proposed.

1 Introduction

Survey information is disseminated by National Statistical Institutes by means of different products, e.g. indices, tabular data or microdata files. With respect to timeliness, accuracy, level of detail and other quality indicators, each of the previous products has its own features. Each product should correspond to some well-defined users needs. Moreover, for either dissemination strategy, the national statistical institute has to guarantee that respondents confidentiality cannot be breached. The risk of re-identification of each unit should be evaluated. If needed, protection methods are commonly applied to reduce the risk of re-identification of respondents. Information loss criteria are then used to select among several protection achieving a pre-defined acceptable level of re-identification risk.

Dissemination of business microdata files for research (MFR) is one of the most delicate challenges the NSIs are facing nowadays. This is mainly due to two special characteristics of the business surveys. The first one is related to the data sampling strategies. Even if the business surveys are conducted as sample surveys, for some strata, all the units are included in the sample. Consequently, a census is actually conducted in several sampling strata. Moreover, the sampling frame content is often very similar to some external, publicly available, business register. The second special feature of business surveys is represented by the characteristics of the variables usually registered. Some economic variables like turnover or exports generally have very skew distributions. This means that only a small number of units significantly contribute to the overall phenomena. These units would then be more identifiable then others because such concentration is generally publicly known.

There aren't so many examples around the world of release of enterprise microdata. The Regulation CE 831/2002 establishes a list of european business surveys

for whom access is granted for research purposes: the Community Innovation Survey, the Structure of Earnings Survey and the Continuing Vocational Training Survey. These surveys have undergone a complex process of harmonisation that inherently includes comparability as an important dimension of the quality framework. Comparability aims at measuring the impact of differences in applied statistical concepts and definitions on the comparison of statistics between geographical areas, non-geographical domains, or over time. The factors that may cause several statistical figures to lose comparability are attributes of the survey that produces them. Such features may be grouped into two broad categories: the first one relates to survey concepts and the second one relates to measurement and estimation methodologies. To address the problems deriving from the first type of attributes, the approach usually taken at European level is via a regulatory framework where all the concepts of the survey are clearly defined and harmonised. This common framework clearly defines the phenomenon under study, target population, statistical units to be surveyed and all possible metadata descriptions for all the variables involved so as to avoid "structural" non-comparability. As far as the second group of issues is concerned, guidelines on the suggested methodologies for every survey phase are given: sampling design, data collection, weight calculation, imputation and so on. To improve standardisation on all phases, routines are provided by Eurostat for the use of member states. However, "member states are in general free to use whatever methods they prefer as long as some quality thresholds are met", see [6]. These guidelines coupled with quality thresholds represent the core of methodological comparability among member states. In fact, a decision process that gives some guarantee of reaching comparability judgements that are neither ad hoc nor arbitrary is necessary. This involves an assessment of the effects of different practices on predefined statistics, a threshold for determining when action is necessary and a process for choosing acceptable practices. We claim that this general framework adopted at European level to design and carry out harmonised surveys should be implemented also for the microdata release phase. We support this point by analysing the Community Innovation Survey.

In section 2 a description of the survey is given together with examples of how comparability has already been used in other phases of the survey and how it could be implemented in the release of microdata. Section 3 presents current situation as far as anonymisation strategies are concerned and suggests future evolution based on the comparability framework. Finally, some conclusions are given in section 4.

2 Survey comparability

The Community Innovation Survey (CIS) collects information on the innovation tendency at firm level. On each statistical unit, the enterprise, CIS registers information on the economic activity, geographical location, number of employees, expenditure on innovation and research, etc.. The latter is decomposed with respect to factors like intramural/extramural research, acquisition of machinery, acquisition of external knowledge, personnel training, etc. Various facets of innovation are also investigated, e.g. factors that determine or hamper innovation, number of employees

Country	Strata criteria	N° N	N° S	N° R	Sample rate	Response rate
BE	N, S, R	4	5	3	32%	30%
DK	N, S	18	4		39%	30%
DE	N, S, R	21	3	2	12%	21%
EL	N, S, R	20	3	3	30%	62%
FR	N, S	Mixed	3 - 5		12%	82%
IT	N, S, R	46	5	20	20%	62%
LU	N, S	9	3		45%	72% - 73%
NL	N, S	41	3		43%	55%
AT	N, S, R	16	5	3	22%	43%
PT	N, S	42	3		19%	46%
FI	N, S	23	4		35%	50%
SE	N, S	37	6		27%	48%
IS	Census	Census	Census	Census	Census	93%*
NO	N, S	41	5		40%	94%

Table 1: CIS3: stratified sampling for each Member State. N = NACE, S = firm size, R = region. * for a pre-survey by phone.

with higher education, number of registered patents, etc.. A full survey description of the third wave of CIS (CIS3) is given in [4].

For the CIS in order to ensure what we have called "structural" comparability across countries, Eurostat, in close cooperation with the EU Member States, developed a standard core questionnaire, with an accompanying set of definitions and detailed metadata based on [16]. To address the type of incompatibilities due to issues of measurement and estimation, clear methodological recommendations were given at European level. If we take as an example the sampling design, the general methodological indications were to break down the target population into similar structured subgroups or strata which should be as homogeneous as possible and form mutually exclusive groups. The stratification variables to be used, i.e. the characteristics used to break down the sample into similarly structured groups, were: the economic activities (in accordance with NACE), enterprise size (given as the number of employees) and regional aspects. The sample selection should be based on random sampling techniques, with known selection probabilities, applied to strata. In Table 1 we replicate the table presented in [4] on the practices followed by several countries for CIS3.

It is clear that, although the sampling design adopted by member states was broadly a stratified random sampling one, the number of strata as well as the hierarchical level of the stratifying variables were different. This may be due to peculiar characteristics of the phenomenon and distribution of enterprises among NACE in each country as well as well current implemented practices in member states. In any case, the European regulation on innovation sets clear quality thresholds on predefined statistics in order to comply with the comparability dimension of the European data set. Whatever practice was adopted, to be considered comparable, the sample should be carried out in order to achieve a predefined level of precision

with regards to the following indicators: the percentage of innovators, the share of new or improved products in total turnover and the total turnover per employee. In particular, it is recommended that the 95% confidence interval for the first two indicators should be within ±5%. For the last indicator the confidence interval should be within ±10% of the estimated indicator. So the process can be summarised in three steps: development of general methodological guidelines, definition of benchmarking statistics and assessment of the effects of different practices on such statistics and, finally, the definition of a threshold for determining when an action is necessary.

The application of such framework in the context of anonymisation procedures for the release of microdata files for research would imply, to start with, the indication of the methodological paradigm of statistical disclosure as described for example in [9]. Such paradigm states the definition of a disclosure scenario, subsequent definition of risk, a measure to assess it, procedures to reduce the risk and finally, but absolutely crucial for the whole process, measures of data utility allowing the final users to judge how poor/good the results of his analysis on the anonymised microdata would be. Such utility measures represent the benchmarking statistics for comparability. In fact, as in the sampling example the aim was releasing estimates that exhibit certain characteristics, again in the anonymisation phase one main goal should be the production of anonymised data sets sharing certain statistics with the original microdata. The key of the whole process should then be the definition of protection methods that maintain such statistics or the customisation of existing procedures to guarantee pre-selected characteristics.

We now analyse in detail each step of the statistical disclosure paradigm.

2.1 Disclosure scenario

A disclosure scenario defines the users of the released microdata and describes possible ways a malicious user could try to re-identify a unit in the released file. It also examines which variables can be used for re-identification purposes leading to the definition of the so called identifying variables. The scenario highlights the possible disclosure content, too. For any MFR the possible user is a very well-defined one: a scientific researcher who signs a contract impeding him to disclose any individual information from the released file. Consequently, the bona fide of such user may be readily assumed. A researcher has an optimal knowledge about the studied phenomenon, not necessarily about each unit. Anyway, in a business framework, information on the greatest enterprises is well known. We believe that, in this context, the user (although not malicious) has some a priori knowledge of a few large and publicly known enterprises. The disclosure scenario should then prevent spontaneous identification.

With respect to the CIS data, spontaneous identification might be based on combinations of variables included in external business registers such as the main economic activity (NACE), the geographical location (NUTS), the number of employees (EMP) and the total turnover (TURN). For example, the enterprise with the maximum value of turnover in a given NACE code could be publicly known, independently on the exact value of TURN. All these variables are to be considered as identifying variables.

Moreover, due to their skew distributions, other economic numerical variables could be subject to spontaneous identification. For example, among a small group of innovating firms, the one having a dominant investment in research and development, could be known either. RTOT (total expenditure on research and development) and RRDINX (expenditure on intramural R&D) were indicated as variables possibly subject to spontaneous identification and therefore identifying. We notice that some key variables are categorical (NACE and EMP or indeed an indication of the enterprise size) whereas the others are continuous (turnover and expenditures on innovation).

2.2 A general definition of disclosure risk

If we assume a scenario of spontaneous identification then the units at risk will be those that cannot be mistaken for others taking into account some reasonable knowledge (for example economic activity and size). If a unit u belongs to a very dense cloud of units similar to it, it may be supposed that its re-identification wouldn't be of interest: the potential gain would be too little with respect to resources needed. Moreover, since there are other units similar to u, due to the existence of measurement errors, the re-identification would still be uncertain. Instead, if a unit is very isolated with respect to its closest neighbours, it would be more easy to recognise it with some confidence. The latter units should be then considered at risk of re-identification.

2.3 Risk assessment

Once a definition of risk has been given, a measure or estimate of it is necessary. As first step, a specification of the level of detail of the categorical identifying variable is needed as these act as stratifying variables, i.e. defining sub-domains for the analysis of the users. The minimal user requirements are: a 2-digit NACE code for the economic activity variable and indication of the firm size. So, in accordance to the sampling scheme and researcher needs, a variable was produced that groups the number of employees in three classes -49, 50-249, 250+. The geographical variable NUTS was recoded at national level since at macroregional level would have led to an unacceptable number of extremely identifiable enterprises. The definition of these possible sub-domains (main economic activity at 2-digit NACE code and size of enterprise) allows now to concentrate on the continuous identifying variables.

As NACE and size are deemed very reliable variables, they may represent the a priori knowledge of users on the structure of the economy. The risk of re-identification with respect to the numerical identifying variable turnover and innovation expenditures could then be estimated in each sub-domain i.e. for each combination of these two categorical key variables. Based on its own dissemination policy, each National Statistical Institute should define the minimum number of units to which a unit u should be close to in order to be considered safe.

2.4 Microdata protection

Dissemination of business microdata files for research should involve a dedicated anonymisation method in order to avoid the most obvious identifications. By modifying only the records at risk of re-identification and only for the key variables, an optimal trade-off between protection and information loss could be achieved. For microdata files for research purposes, the information quality constraints are much more tight with respect to other dissemination products. Researchers naturally require a high quality data in order to perform reliable analysis and derive correct conclusions. Even if some statistical disclosure limitation method is applied, preservation of the most important statistics is highly desirable. In sections 3.1 and 3.2 two current protections methods will be analysed in more detail.

Once the statistical disclosure paradigm has been set up, the definition of benchmarking statistics to check whether different disclosure procedures may be considered comparable from the final user point of view is necessary. Such benchmarking statistics should be related to user needs, i.e. data utility. In the next section we try to address this issue taking as a starting point several analyses carried out on CIS data in recent years.

2.5 Choosing benchmarking statistics: remarks on analyses performed on CIS data

Whatever protection method is applied, some information loss is *unavoidable*. Preservation of the most used analytical properties of the microdata file is possible only if the data protector is aware of the possible data usages. Even if some important statistics cannot be exactly maintained, the information loss with respect to these indicators should be at least quantified. For coherence reasons, from the point of view of the data producer, preservation of already published statistics (mainly tabular data) would be desirable. From the user point of view, data utility measures are essential in knowing the difference with original data.

In order to gain some insight on possible statistical usages of CIS data a brief review on the scientific literature based on such data was carried out. An example of such review is provided in [1]. Below few common characteristics of several analyses based on CIS data are given.

Analyses are commonly performed at NACE 2-digit level, using the data at national level. This proves the strategic importance of the economic variable. Consequently, dissemination of CIS data at a more aggregated level of the economic activity would be almost useless.

A relationship between companies economic performance and their innovation attitude is commonly investigated. The economic performance may be modeled, for example, through turnover, employment and their variations. Examples of studied statistics are the innovation intensity (expenditure per employee on innovation linked to employment growth, by internal or external innovation) or the share of turnover that is due to new or improved products (quantifying the economic relevance of innovations). Each registered component of the expenditure on innovation is equally used to analyse the innovation phenomenon. Correlations and ratios involving these

components and the ones expressing the economic performance seem to be particularly important. Such analyses may be found, for example, in [8, 12, 15, 14, 2].

As usual in survey statistics, weighted means are widely used. Besides being part of the already published tabular data, weighted means were found to be involved in the majority of analysis. For example, any share is expressed through the weighted totals. Consequently, preservation of such statistics is crucial.

As a result of this short overview we can cast a possible list of statistics to be used for benchmarking purposes in data utility. Ratios of innovation variables as a mean to analyse scaled quantities seems predominant. Also the change in turnover with respect to the first year of the reference period (a recorded variable) seems relevant. The next step in this approach is the definition of thresholds on such statistics in order to define comparability of anonymisation methods.

3 Towards comparable dissemination

The current situation regarding CIS microdata access at European level is quite varied. Eurostat has suggested, for CIS3, an individual ranking based anonymisation strategy for numerical variables, see [5]; such procedure has recently been reviewed for the CIS4, see [7], but maintaining the same approach. The idea of micro-aggregation and some features of the current implementation are commented in section 3.1. Some member states have accepted this proposal while other for different reasons haven't done so. Following a series of experiences on business microdata anonymisation, Istat is going to release microdata files for research stemming from both CIS3 and CIS4. The essential features of the procedure are described in [10, 11]; in section 3.2 the basic ideas and some results of its application are reported. In both approaches the categorical key variables are recoded following data utility reasonings. The applied recoding is generally given by the minimal user requirements, see section 2.3. However,the two approaches differ a lot on recognising these variables as structural information: the consequences of the different reasonings will be discussed in both section 3.1 and 3.2. In section 3.3 a view on how to gather the current approaches into a unique framework is presented and a way to proceed in order to reach comparable dissemination for CIS microdata is proposed.

3.1 Micro-aggregation based strategy

Micro-aggregation is a very well-known perturbation method introduced in [3]. It aims at creating at least k equal units with respect to the numerical key variables of a statistical unit. The records subject to a micro-aggregation process are clustered in groups of at least k similar units. Then, the value taken by a variable on a unit is replaced by the mean of the group to which the unit belongs to. Instead of the mean, other statistical indicators, like median or weighted mean, may be used.

The basic idea of micro-aggregation is removal of the re-identification risk by means of perturbation. The simplest way to use micro-aggregation is to indistinctly apply it to all numerical variable and to all units. The underline idea is that all values are changed (and there exists always at least k equal values in the released

file) so as to prevent of exact disclosure. However in microaggregation there is a clear lack of risk definition and assessment.

When micro-aggregation is applied in real case-studies there are several issues that ought to be discussed. Firstly, the choice of k should be derived from each NSI dissemination policy. In the following examples reporting results on the Italian CIS4 data, k was set equal to three. Secondly, the way in which the numerical variables are micro-aggregated has to be tackled. A possible strategy is to apply a multivariate micro-aggregation, i.e. all numerical identifying variables are simultaneously micro-aggregated. In practical situations, this strategy is not generally used because it was proved that it produces a significant information loss. An alternative would be the individual application of micro-aggregation on each numerical variable independently from the others. This approach is called individual ranking and was actually adopted by Eurostat. The final issue to be addressed in the application of micro-aggregation is the treatment of categorical identifying variable. These could be totally ignored in which case the numerical identifying variables could be readily perturbed by micro-aggregation, independently on such categorical variables, i.e. independently on the specification of important sub-domains. In practice, this means that all sample units are subject to the same micro-aggregation process which produces almost no information loss. Application of micro-aggregation irrespective of the structural economic variables (which are generally represented by the categorical identifying variables) was widely reported in statistical disclosure control. In [17], it was noted that micro-aggregation does not offer any degree of protection, even for higher values of k. A ranking approach for risk assessment of micro-aggregated CIS microdata was reported in [13]. For what attains the disclosure scenario described in section 2.1 we simply notice that individual ranking not stratified by NACE and EMP does implicitly ignores any intruder *a-priori* knowledge.

These issues may be observed in figure 1. The upper panel shows the correlations between RTOT and TURN, for each combination of NACE and EMP. The preservation of this statistic is quite remarkable and the same effect was observed for other statistics taken into consideration; little information loss should be reported. The lower panel proves the inefficiency of such micro-aggregation in protecting visible dominant values: an extreme isolated outlier maintains this property when individual ranking is applied without taking into consideration the categorical identifying variables NACE and EMP. Setting k equal to 5, didn't seem to improve too much the protection level achieved by this type of micro-aggregation. The weighted mean usage instead of the average, didn't significantly change the outcome of this simulation, too.

3.2 Stratified selective masking

An alternative approach has been followed in [10, 11] where the paradigm of statistical disclosure has been followed. In strata defined according to the disclosure scenario chosen the risk of re-identification has been estimated following a density approach. Indeed, the density of units around a unit u may be quantified by the local outlier factor. The latter is a relative measure of the degree of isolation of u. Then, by setting a threshold, the units at risk of re-identification in each stratum

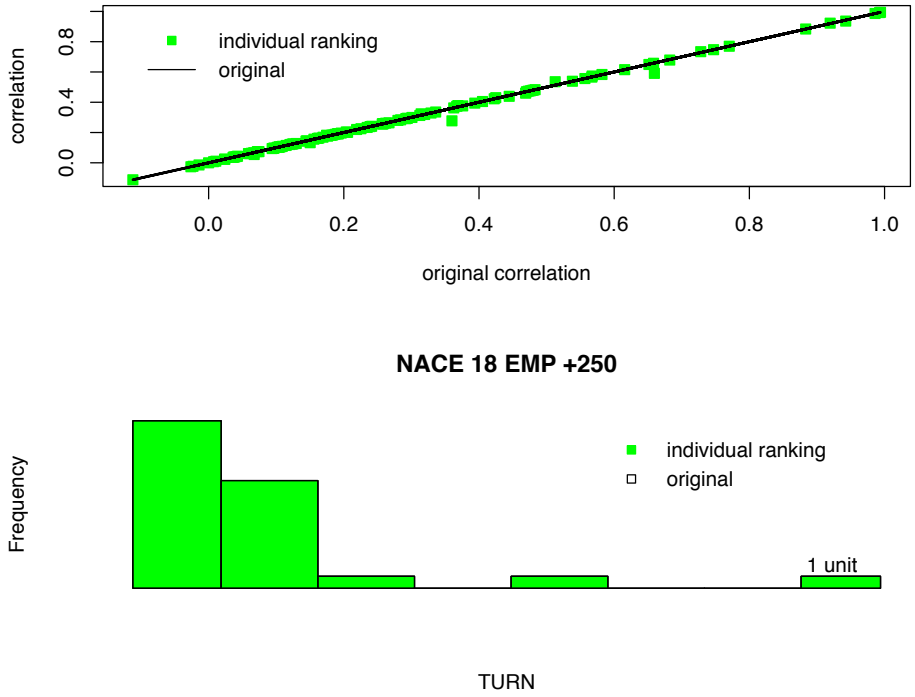


Figure 1: Results obtained when individual ranking is applied to TURN. $k = 3$.

may be singled out. In [11] an automatic method for this latter threshold setting is also discussed.

As far as protection is concerned different methods are proposed according to the nature of the identifying variables: categorical or continuous. For categorical identifying variables representing structural information of the phenomenon (NACE, size and Nuts) no perturbation is performed. If necessary, i.e. when only a very small number of enterprises are present in each combination of such variables, a recoding is suggested. The combinations of the levels of the resulting variables are the strata where the successive protection of the continuous identifying variables will be performed. For such variables a selective masking method based on the same uncertainty principle is used; moreover, by modifying only the records at risk of re-identification the information loss could be reduced. The selective masking proposed in [11] is based on the nearest neighbour imputation and micro-aggregation. The first perturbation method is applied only to those units at risk whose nearest neighbour is not at risk of re-identification. Then, micro-aggregation is applied to the remaining units at risk.

For the Italian CIS4, this protection method was applied to TURN for each combination of the categorical identifying variables. RTOT and RRDINX values for the units at risk of re-identification in the spontaneous identification were also perturbed. A change proportional to the change introduced in the corresponding TURN value was aimed. In order to preserve the relationship between variables, the other components of the expenditure in innovation and research were modified, too.

Several statistics were taken into account to perform a comparison between individual ranking and the uncertainty based selective protection method. For each combination of NACE and EMP, the preservation of TURN distribution was firstly addressed. The figure 2 shows a typical result. As expected, the individual ranking reduces the skewness. The comparison of variances and correlations is reported in figure 3. Generally, the selective protection method preserves better these statistical indicators. As mentioned in section 2, the expenditure in innovation and research variables are very concentrated in some units. Figure 3 also shows the performance of the two protection methods with respect to the Gini concentration coefficient. Since the ratios like RTOT/TURN are very much used in data analysis, the perturbation effect on such distributions was assessed; a typical outcome is presented in figure 3.

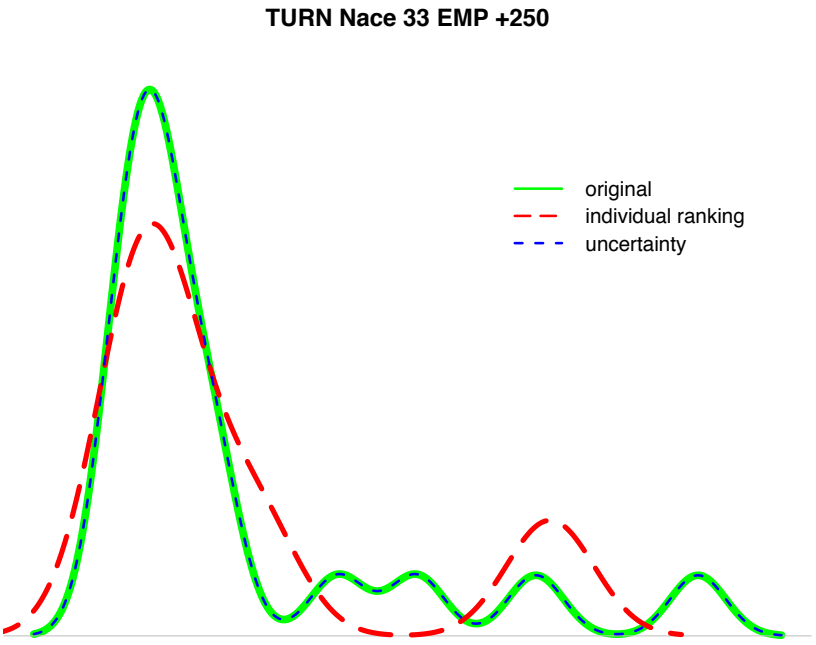


Figure 2: Comparison of individual ranking and uncertainty-based selective masking; both were applied for each combination of NACE and EMP.

3.3 Reaching comparability

An interesting feature of the anonymistion procedure outlined in 3.2 is that for extreme choice of the parameters in the risk assessment phase the protection process is equivalent to individual ranking. In fact, if a degenerate distance is considered for the categorical identifying variables, the method would be applied irrespective of NACE and EMP. Additionally, if an extreme threshold (zero) were used, all units

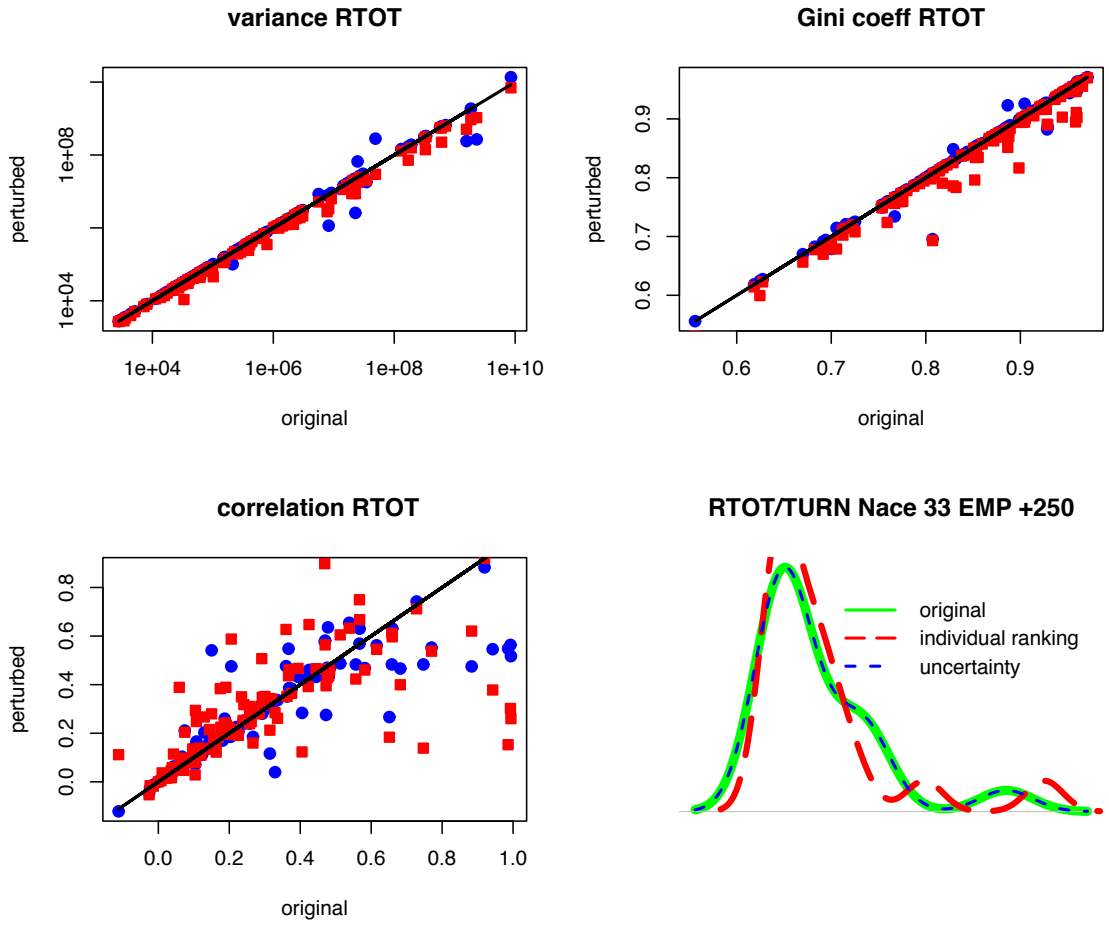


Figure 3: Comparison of some analytical properties. The red squares represent the individual ranking; the blue circles represent the selective uncertainty-based method; the black lines indicate the original values.

would be considered at risk of re-identification. According to the procedure, all these units will be protected using micro-aggregation. An evolution of the current situation could see selective masking as possible framework for choosing different degrees of anonymisation. Eurostat could provide member states guidelines on the release of CIS microdata along the lines with what is currently done for the other phase of the survey (eg. sampling). CIS experts and users could define further benchmarking statistics useful to measure relevant data utility and set thresholds to guarantee a common *baseline* quality for anonymous microdata. Careful definition and tuning of benchmarking statistics coupled with clear threshold setting would allow comparability of analysis among different methods and different parameter choices in different member states. After a period of training of member states and the preparation of suitable routines implementing different methods and evaluating the benchmarking statistics it would be possible to move towards a complete harmonisation on this subject.

4 Conclusions and further research

To obtain comparable and high quality survey data, a great effort was already made at EU level to harmonize CIS data collection and processing. Ideally, dissemination of microdata would have to be harmonized, too. Unfortunately, depending on the national situation, e.g. law restrictions or availability and quality of external business registers, each National Statistical Institute must face its own problems. Nonetheless, harmonisation of the anonymisation methodology may be achieved twofold. Firstly, different intensity parameters may be used to reach some required, pre-defined, protection levels. Secondly, the analytical issues should be put in evidence.

In this paper we claim that the implementation of the classical statistical disclosure paradigm for enterprise microdata is indeed possible. Moreover, we propose a flexible framework for developing different anonymisation procedures suitable for different member states, guaranteeing the final users on data quality. This is achieved through the use of the comparability concept. Careful definition of relevant statistics for the type of data under analysis is a key issue for defining data utility measures. Given the assurance of a pre-defined acceptable re-identification risk level, preservation of benchmarking statistics should then be the primary objective, independently on the anonymisation methodology. Further research will be devoted into developing guidelines, setting the list of key statistics and implementing the whole framework investigating the degree of flexibility of the protection methodology to reach the required level for the data utility indicators. Cooperation between survey experts and methodologists is strategic for improving the current situation by both increasing the number of microdata offered and improving the quality of the released data.

Acknowledgment

The views expressed in the paper are those of the authors and do not necessarily reflect Istat policies. The authors thank Valeria Mastrostefano for helpful suggestions.

References

[1] Arundel, A. and Bordoy, C. (2005) "The 4^{th} Community Innovation Survey: Final Questionnaire, Supporting Documentation, and the State-of-Art for the Design of the CIS", working paper, available on request.

[2] Belderbos, R., Carree, M. and Lokshin, B. (2004) "Cooperative R&D and Firm Performance", *Conference on Industrial dynamics, innovation and development.*

[3] Defays, D. and Anwar M.N. (1998), "Masking Microdata Using Micro-Aggregation", *Journal of Official Statistics*, **14 (4)**, 449–461.

[4] Eurostat (2004) "Innovation in Europe: Results for the EU, Iceland and Norway", *Panorama of the European Union, Theme 9 Sciences and technologies, European Communities.*

[5] Eurostat (2005) "The Third Community Innovation Survey. Methodology of Anonymisation."

[6] Eurostat (2006) " The Fourth Community Innovation Survey (CIS4). Methodological recommendations", Doc. Eurostat/F4/STI/CIS/2b.

[7] Eurostat (2007) "The CIS4 micro-data anonymisation method".

[8] Evangelista, R. and Mastrostefano, V. (2006) "Firm size, sectors and countries as sources of variety in innovation", *Economics of Innovation and New Technology*, **15 (3)**, 247–270.

[9] Hundepool, A. *et. al.* (2006) "Handbook on Statistical disclosure control", available at http://neon.vb.cbs.nl/casc/.

[10] Ichim, D. (2007) "Microdata anonymisation of the Community Innovation Survey data: a density based clustering approach for risk assessment", *Documenti Istat*, 2, available at www.istat.it.

[11] Ichim, D. (2007) "Disclosure control for business microdata: a density-based approach", submitted.

[12] Klomp, L. and van Leeuwen, G. (2001) "Linking innovation and firm performance: a new approach", *International Journal of the Economics of Business*, **8(3)**, 343–364.

[13] Leppälahti, A. and Teikari, I. (2007) "Problems with micro-data from small countries", *32nd CEIES Seminar Innovation Indicators - more than technology.*

[14] Loof, H. and Heshmati, A. (2002) "Knowledge capital and performance heterogeneity: a firm level innovation study", *International Journal of Production Economics*, **76(1)**, 61–85.

[15] Mastrostefano, V. and Pianta, M. (2007) "Innovation Dynamics and Employment Effects", submitted.

[16] OECD and Eurostat (1997) "Oslo-Manual, Proposed Guidelines for Collecting and Interpreting Technological Innovation Data", *Organisation for Economic Co-Operation and Development*, Paris.

[17] Winkler, W. (2004) "Re-identification methods for masked microdata." In *Privacy in Statistical Databases.*, Eds. J. Domingo-Ferrer and V. Torra, 216–230.

Microdata sharing via pseudonymization

David Galindo*, Eric R. Verheul**,***

* Department of Computer Science, University of Malaga, Spain. (dgalindo@lcc.uma.es)

** Institute for Computing and Information Sciences, Radboud University Nijmegen, The Netherlands. (eric.verheul@cs.ru.nl)

*** PricewaterhouseCoopers Advisory, The Netherlands. (eric.verheul@nl.pwc.com)

Abstract. Individual data records are essential for empirical research, and yet due to the very precious information they contain, their release poses a problem to the confidentiality of the individuals concerned. In this paper we give a high level description of a privacy-preserving microdata sharing system wherein subjects identifiers are replaced by cryptographic pseudonyms. The resulting system facilitates information sharing between organizations that typically are not allowed to exchange the microdata they own.

1 Introduction

Individual data records are essential for empirical research, and yet due to the very precious information they contain, their direct release thwarts the confidentiality of the individuals concerned. The fact that research is interested in collective features rather than individual distinctiveness, makes it possible to reconcile data utility and individual confidentiality: data identifiers can be removed or encoded and data fields can be modified by means of statistical disclosure controls, while overall the collective features of the resulting de-identified data are preserved.

Microdata comes from heterogenous sources, such as statistical offices, hospitals or insurance companies to name a few. There are a number of parties, named as Researchers, interested in getting access to this data for economical or research purposes. In the case of national statistical offices, Researchers face in general two modes of accesses: either access to the microdata is granted in the premises of the national statistic authorities; or the microdata is anonymized and released to Researchers under certain conditions. In both cases, the original data has been modified to preclude the direct identification of the subjects.

The aim of this paper is to describe privacy-preserving microdata sharing systems obtained by replacing subjects identifiers with pseudonyms with special mathematical and cryptographic properties. The pseudonymizing system is controlled by a Trusted Third Party (TTP), and no party in the scheme (except the TTP) can re-identify the individuals from the pseudonyms. Still, natural set operations between

different pseudonymized databases, like database union and intersection are supported. These operations allow for flexible research of personal data of individuals residing at different organizations that typically do not share information.

2 Pseudonymous data sharing

Consider a *database* consisting of entries of the form $(id, D(id))$, where id is the identifier field (also called identity) and $D(id)$ is the data field. A *pseudonymized database* is obtained by replacing the identity id in the database entries by a blinded identifier $P(id, O)$, called *pseudonym*. The blinded identifier $P(id, O)$ does ideally not leak any information on the identity id. The individual with identity id is only known to the Organization O by its pseudonym $P(id, O)$, and the key property is that the organization O is not able to link together $P(id, O)$ and id (under certain cryptographic assumptions). This property is called *pseudonymity*.

Pseudonymized databases with the above properties provide a virtually unexplored tool for building privacy preserving information sharing systems. Roughly speaking, an information sharing system is called *privacy-preserving* if no information is leaked on individuals identities. We stress the latter is interpreted from a strict cryptographic point of view, that is, the qualification privacy-preserving refers to the cryptographic techniques used for pseudonymization and related operations, since from a global point of view privacy-preserving pseudonymized microdata sharing systems likely do not exist. The reason is simple, even though the data is pseudonymized, there is the risk that the characteristics of the data singles out a person, e.g. by a combination of profession, age and place of residence. This risk of *indirect identification*, cf. [6, 3], becomes even larger when linking several pseudonymized databases, which is one of our targets. The issue of indirect identification is outside the scope of this paper and is covered by an abundant literature [1]. Although out of scope, it is our position that indirect identification should be an important point of attention in deciding which data Researchers are provided with; at the very least a Researcher should only get the information required for his Research and nothing more.

An example is illustrative. Suppose a Researcher wants to find out the correlation between certain pharmacy usage and traffic accidents, e.g. with the objective to provide for better warnings on the usage of certain pharmacy in traffic. However privacy laws prevent the Suppliers holding the data from releasing this information. Let us assume that the representation of identities of individuals is unique, e.g. takes the form of a Social Security Number. We can achieve the Researcher's desired functionality while circumventing the Suppliers concerns by providing the Researcher with two kinds of data: pseudonymized drug usage of individuals from pharmacies

[1] The interested reader is referred to [6] for an introduction to this topic, and to [7] for a state of the art.

and pseudonymized traffic accidents data from insurance companies.

With the pseudonymized data received, the Researcher can easily compute its target correlation, since an individual that occurs in the non-pseudonymized databases of pharmacies and insurers leads to the same pseudonym $P(id, R)$ in the Researcher's database. We name this information sharing technique as *pseudonymous data sharing*. Obviously, a malicious Researcher is tempted to learn the identities of the individuals involved. Still, a misbehaving Researcher is prevented from learning information on the identities thanks to the use of pseudonyms. Additionally and depending on the application, one might require that two Researchers R_i and R_j should not be able to match their pseudonymized databases, and thus the pseudonyms $P(id, R_i)$ and $P(id, R_j)$ must necessarily be different and unlinkable.

This in principle complies with the Recommendation 83 (10) of the Council of Europe [10, 12] on the protection of personal data collected and processed for statistical purposes, since the recommendation states that

> An individual should not be regarded as 'identifiable' if the identification requires an unreasonable amount of time, cost and manpower.

3 Applications

There are several contexts where this technique is useful. Consider the case of several organizations working with different responsibilities on the same set of (potential) individuals. These responsibilities prevent those organizations from sharing information unless certain concurrences occur. Alas, one cannot detect concurrences without sharing information, thus ending in a catch-22 situation. By sharing on a pseudonymous basis one can break through this situation. This catch-22 situation does not come without consequences. Indeed, it was one of the reasons the 9/11 tragedy was not timely foreseen; the different agencies (law enforcement, secret service) did not share their information in a timely fashion, see the 9/11 report [5, p.416]. It appears that this situation is also present in the context of child-abuse. According to [4], if social workers, nurses, midwifes, psychiatrists, police share their information in a timely fashion then serious child abuse can be prevented. In both situations one can use our techniques to develop a Reference Registry containing all individuals on a pseudonymous basis; as soon as a concurrence occur, a special procedure is taken to identify the responsible individual.

Other applications include the integration of separated clinical and biological data (e.g. genomic data) in common databases [2, 13]. In case relevant diagnostic findings are revealed in the join of the databases, only a trusted party (to be introduced later) can re-identify the patient from the pseudonym. If the results gained in the research can positively influence the patient's therapy, the patient can be contacted and be told the results. This and the former example show that

pseudonymization has advantages over anonymization, since pseudonyms enable re-identification by a trusted authority, whereas anonymization does not.

4 Our framework

The framework for pseudonymized data sharing we propose is based on three main considerations. In the first place, we observe that the main task of Supplier organizations is very often different from that of supplying information to Researchers. Following our example, this would be the case for pharmacies or insurers, while a governmental agency taking care of traffic matters falls probably outside of this category. In our view it is important to keep Suppliers workload as low as possible in a practical information sharing system. To this aim we introduce in our model an intermediary organization called *Accumulator*. These are buffers between Suppliers and Researchers which are responsible for keeping pseudonymized versions of Suppliers databases and feeding Researchers with data. From time to time, Suppliers hand over their data in pseudonymized form to one or more Accumulators that each have their own sets of pseudonyms $P(id, A_i)$. Researchers are supplied pseudonymized data by Accumulators under pseudonyms $P(id, R_j)$. Accumulators exchange data in a sensible manner too. Suppliers have their workload alleviated since it is supposed they hand over data to Accumulators rarely, for instance only when important database updates need to reflected. In contrast, Researchers will ask Accumulators for data much more frequently.

Secondly, the allowance of these protocols and the type of data that is sent along with the protocols is governed by the Regulatory Privacy Body (RPB) from a functional perspective. We envision that a strict licensing infrastructure for protocols will be enforced by the RPB, describing:

1. Which parties are allowed to perform what protocols with each other.
2. What kind of data may be sent along with the protocols.
3. What kind of subsets of identities (Suppliers) or pseudonyms (Accumulators) are allowed as input to the protocols.

From a technical perspective the execution of all protocols depends on cryptographic keys governed by a Trusted Third Party (TTP). We want the involvement of this TTP to be very low, at least from a computational and availability point of view.

Thirdly, we deliberately choose to not support protocols between Researchers as they are assumed malicious and might deviate from protocols, e.g., by sending along disallowed data with the pseudonyms.

Supported operations. Let us now outline the operations enabled by a pseudonymous data sharing system. The execution of all protocols requires cryptographic

keys governed by a Trusted Third Party (TTP); the parties in our system need to be handed in advance certain cryptographic keys by the TTP (see Figure 1).

Supplier Pseudonymization The result of the operation $S_i \rightarrow_P A_j$ is that a (selected) list of identities id in Supplier S_i database is provided to Accumulator A_j under his pseudonyms $P(id, A_j)$. This operation typically needs to be performed periodically to reflect updates of the Supplier's database. By sending along data related to these identities, Accumulator A_j gets a pseudonymized version of S_i's database (or part thereof). After this operation has been performed, Supplier S_i can not link together id and $P(id, A_j)$.

Accumulator Exchange The operations $\rightarrow_\cup$ and $\rightarrow_\cap$ (pseudonyms union and intersection respectively) enable two Accumulators to perform the two atomic set operations union and intersection on their pseudonymized data collections.

operation $\rightarrow_\cup$: After the operation $A_i \rightarrow_\cup A_j$ the Accumulator A_j possesses pseudonyms of the form $P(id, A_j)$ for every id that was available in pseudonymized form in either A_i's or A_j's database. By sending along data related to these identities, Accumulator A_j gets a pseudonymized join with A_i's database.

operation $\rightarrow_\cap$: After the operation $A_i \rightarrow_\cap A_j$ the Accumulator A_j knows which of his pseudonyms also occur (as peers) in Accumulator A_i's database. If A_i relates his data to these pseudonyms then a specific encryption feature in the $A_i \rightarrow_\cap A_j$ operation facilitates that A_j can only decrypt the data for pseudonyms in the intersection.

Researcher Provisioning As the result of the data provisioning operation $A_i \dashrightarrow_D R_j$ between an Accumulator A_i and Researcher R_j a list of pseudonyms in an Accumulator's database is provided to the Researcher under his pseudonyms $P(id, R_j)$. By sending along data related to these pseudonyms, Researcher R_j gets a pseudonymized version of A_i's database.

5 Security properties

The TTP is assumed honest, i.e. it will not deviate from protocols or try to deduce secret information from the information it gets as part of protocol execution. Suppliers and Accumulators are assumed honest but curious. That is, they will not deviate from protocols but might try to deduce secret information from the information they get. In practice one should try to ensure that Suppliers and Accumulators are honest; the fact that our model allows them to be curious should be seen as an indication that there is some "room for error". Finally and most importantly, we assume that Researchers are malicious. They are willing to deviate from protocols

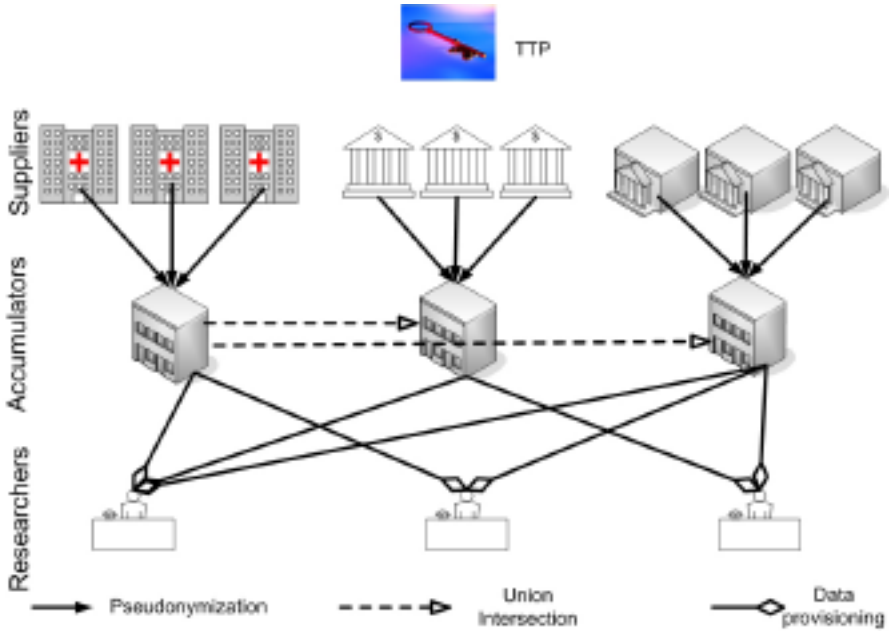


Figure 1: Overview

and to share their cryptographic keys with other Researchers in order to deduce secret information; most notably to relate pseudonyms of other parties or to realize pseudonymity removal.

We require two basic security properties. First and foremost is the prevention of pseudonymity removal: no agent in the scheme (Suppliers, Accumulators or Researchers) should be able to relate through cryptanalysis a given pseudonym with a given identity even if this party was able to do so for many pseudonyms by other means (most notably through indirect identification). This property is called *pseudonymity*.

The second property deals with the prevention of pseudonym matching of different parties. If a party X in the scheme gets hold of a copy of a database of another party Y, then party X should not be able to relate his database with that of Y without the collaboration of Y, even if X and Y previously run the pseudonymized intersection protocol and X is able to relate the databases partially by non-cryptographic means. Additionally, the sender in a protocol should not be able from engaging in the protocol to deduce information about which pseudonyms are already in the receiving party database. These properties are named as *mutual separation*. This property considers the scenario where X's pseudonymized database is renewed periodically, and where the fact that Y was allowed to match its database to that of X does not imply this matching is allowed at a later time. Notice that the mutual separation property excludes the possibility that the pseudonymous data sharing operations are

implemented by the parties through a transition table between pseudonyms, that is, a table of pseudonym pairs $\{(P(id, A_s), P(id, A_d))\}$ or $\{(P(id, A_s), P(id, R_u))\}$.

6 Designing a pseudonymous data sharing system

We start by outlining a straightforward implementation of pseudonymous data sharing system based on symmetric encryption. This implementation enjoys high efficiency due to the use of symmetric encryption, but suffers from requiring an on-line TTP to perform pseudonymization and database operations, as well as scalability and flexibility drawbacks.

In this implementation, the TTP selects a blockcipher $\big(\mathrm{Enc}_{K_i}(\cdot), \mathrm{Dec}_{K_i}(\cdot)\big)$ and generates for each Accumulator A_i a secret key K_i which is unknown to A_i. The database of Supplier S_j is denoted as database rows $\big(id, D(id)\big)$. Supplier S_j sends the datablocks $D(id)$ directly to A_i, i.e. the identity field is removed. The identities id are sent (in the same order as the datablocks were sent to A_i) to the TTP. The TTP encrypts the identities using K_i, leading to $P(id, R_i) = \mathrm{Enc}_{K_i}(id)$ and sends them to A_i; these constitute A_i's pseudonyms. The Accumulator A_i joins the information received from the Supplier and the TTP in the same order, leading to the pseudonymized database $\big(P(id, A_i), D(id, i)\big)$. The union operation $A_i \rightarrow_{\cup} A_m$ would be as follows: Accumulator A_i sends the data blocks $D(id, i)$ of his choosing directly to A_m and sends his pseudonyms to the TTP in the same order. On receipt the TTP does a decrypt operation $id = \mathrm{Dec}_{K_i}\big(P(id, A_i)\big)$ and an encrypt operation $\mathrm{Enc}_{K_m}(id)$ (leading to the pseudonyms of A_m) and sends these to A_m in the same order as received. On receipt R_m joins the information received from the Researcher R_m and the TTP in the same order, leading to the pseudonymized database $\big(P(id, A_m), D(id, i)\big)$. Finally, R_m joins the latter with its own database $\big(P(id, A_m), D(id, m)\big)$. The rest of the operations are implemented analogously.

This implementation enjoys efficiency and high security properties, but requires an on-line TTP, which is a burden to the system. It turns out that an efficient pseudonymous data sharing system with low involvement of the TTP (only in charge of distributing at once cryptographic keys) is possible by using more sophisticated asymmetric cryptography techniques. The details can be found in [9].

7 Related work

Although the framework of our pseudonymous data sharing scheme is actually found in practice [13, 3], it seems that it has not received much attention in the cryptographic literature. What has received considerable attention is the case in which the holders of the private databases (Suppliers in our framework) are themselves interested in sharing information while preserving privacy. Examples of that are the data mining protocols using secure multiparty computation in [11], or works like [1, 14, 8] focusing on concrete operations like database union or intersection. However, none

of these approaches seem to be directly applicable to our framework. Applying these tools in our setting would imply that if a Researcher R_j wants to perform operations on the databases owned by S_l and S_t, these Suppliers must run themselves the protocol and deliver the results to the Researcher. This is not acceptable here, as we want to avoid Suppliers doing the bulk of the work. Additionally, we would like to note that none of these works provide non-interactive protocols for implementing database intersection, in contrast to our concrete scheme [9].

References

[1] Rakesh Agrawal, Alexandre V. Evfimievski, and Ramakrishnan Srikant, *Information sharing across private databases.*, SIGMOD Conference (Alon Y. Halevy, Zachary G. Ives, and AnHai Doan, eds.), ACM, 2003, pp. 86–97.

[2] Russ B. Altman and Teri E. Klein, *Challenges for biomedical informatics and pharmacogenomics*, Annual Review of Pharmacology and Toxicology **42** (2002), 113–133.

[3] Dutch Data Protection Authority, *Landelijke zorgregistraties (national healthcare registrations)*, www.dutchdpa.nl, 2005.

[4] Marian Brandon, Jane Dodsworth, and Daphne Rumball, *Serious case reviews: learning to use expertise*, 2005, pp. 160–176.

[5] 9/11 commission, *The 9/11 report*, www.9-11commission.gov/report/911Report.pdf, 2004.

[6] Josep Domingo-Ferrer (ed.), *Inference control in statistical databases, from theory to practice*, Lecture Notes in Computer Science, vol. 2316, Springer, 2002.

[7] Josep Domingo-Ferrer and Luisa Franconi (eds.), *Privacy in statistical databases, cenex-sdc project international conference, PSD 2006, rome, italy, december 13-15, 2006, proceedings*, Lecture Notes in Computer Science, vol. 4302, Springer, 2006.

[8] Michael J. Freedman, Kobbi Nissim, and Benny Pinkas, *Efficient private matching and set intersection*, Advances in Cryptology - EUROCRYPT 2004, Lecture Notes in Computer Science, vol. 3027, Springer, 2004, pp. 1–19.

[9] David Galindo and Eric R. Verheul, *Pseudonymous data sharing*, 2007, Manuscript.

[10] Franz Kraus, *Data protection and access to official microdata for european research*, NESSIE SRoundtable 4 Access to Quality Comparative Data for European Comparative Socio-Economic Research, 2004.

[11] Yehuda Lindell and Benny Pinkas, *Privacy preserving data mining.*, J. Cryptology **15** (2002), no. 3, 177–206.

[12] Council of Europe, *Recommendation no.r(97) 18 on the protection of personal data collected and processed for statistical purposes*, 1997.

[13] Klaus Pommerening and Michael Reng, *Medical and care compunetics 1*, ch. Secondary Use of the EHR via Pseudonymisation, pp. 441–446, IOS Press, 2004.

[14] Alberto Maria Segre, Andrew Wildenberg, Veronica Vieland, and Ying Zhang, *Privacy-preserving data set union.*, Privacy in Statistical Databases, 2006, pp. 266–276.

Secret Sharing vs. Encryption-based Techniques For Privacy Preserving Data Mining [1]

Thomas B. Pedersen, Yücel Saygın, Erkay Savaş

Faculty of Engineering and Natural Sciences
Sabanci University, Istanbul, TURKEY

Abstract. Privacy preserving querying and data publishing has been studied in the context of statistical databases and statistical disclosure control. Recently, large-scale data collection and integration efforts increased privacy concerns which motivated data mining researchers to investigate privacy implications of data mining and how data mining can be performed without violating privacy. In this paper, we first provide an overview of privacy preserving data mining focusing on distributed data sources, then we compare two technologies used in privacy preserving data mining. The first technology is encryption-based, and it is used in earlier approaches. The second technology is secret-sharing which is recently being considered as a more efficient approach.

1 Introduction

With the popularity of Internet, it is now extremely easy to collect person-specific data which can also be linked to other data sets. Ubiquitous devices such as RFID tags and readers and GPS equipped mobile phones increased the privacy concerns as well, since they made it possible to collect location information about people. But the turning point was the 9-11 events after which US government increased its nation-wide data collection and integration efforts in the name of "fight against terrorism." In fact, some of the largest airline companies in the US, including American, United and Northwest, turned over millions of passenger records to the FBI according to NY Times. Such scandals proved that privacy risks are real. Privacy breaches could also be accidental as in the case of the AOL scandal. AOL released the "de-identified" search logs of its 650.000 customers over a 3 month period in August 2006. AOL realized its mistake and removed the data, but is was already downloaded by many users, and in fact an individual was being identified from her query logs.

Data mining is motivated by the large-scale data collection efforts by companies and government organizations with the aim of turning massive amounts of raw

[1]This work was partially funded by the Information Society Technologies Programme of the European Commission, Future and Emerging Technologies under IST-014915 GeoPKDD project.

data into useful information. Machine learning, Artificial Intelligence, Statistics, and Databases are utilized in data mining in order to come up with data-centric techniques for extracting models from massive data collections. The extracted models could be in many forms, such as rules, patterns, or decision trees. Most of the profitable applications of data mining concerns humans. Therefore a considerable proportion of the collected data is about people and their activities. This is why data mining and privacy discussions are inseparable now. In fact some data mining projects were not funded due to privacy concerns. According to Computer World [21], "The chairman of the House Committee on Homeland Security has asked Department of Homeland Security Secretary Michael Chertoff to provide a detailed listing of all IT programs that have been canceled, discontinued or modified because of privacy concerns." In response to such privacy concerns, data mining researchers started working on methods for preserving privacy when doing data mining. Techniques were developed for different data mining models, starting from classification models, then association rules and clustering for distributed scenarios.

The research efforts on privacy preserving data mining at European level were supported by two large scale projects funded by Future and Emerging Technologies of Information Society Technologies under the 6th Framework. One of the projects is Geographic Privacy-aware Knowledge Discovery and Delivery (GeoPKDD), and the other is Knowledge Discovery in Ubiquitous Computing (KdUbiq). These projects have different purposes but they are both concentrated on new data mining technologies and their privacy implications. GeoPKDD is a research project concentrating on spatio-temporal knowledge discovery, and privacy issues in spatio-temporal knowledge discovery. KdUbiq aims at creating a community in the area of ubiquitous knowledge discovery which will define the area and research directions. One of the working groups is privacy and security in ubiquitous computing.

In this paper we are going to concentrate on privacy preserving data mining in distributed environments and discuss two classes of techniques, namely the encryption based and recently introduced secret sharing based techniques for privacy preserving data mining.

2 Overview and State-of-the-art

In this section we summarize the pioneering work on privacy preserving data mining which shaped the research in this field. Two concurrent papers published in 2000 by researchers from different groups used the title "Privacy Preserving Data Mining"[3, 14]. Although the titles were the same, their approach and problem setting was different. The authors of [3] assumed that the data mining effort will be outsourced to a third party. Before the data could be handed over to a third party the confidential values in the database, such as the salary of employees, needed to be perturbed in a way that the original probability distribution could be estimated

from the perturbed data but not the original data values. This way, a decision tree can still be constructed from the perturbed data within a certain error margin that the authors approve. The authors in [14] assume that there are two parties with private data sources who would like to do data mining without seeing each other's data and propose cryptographic techniques to achieve that. They also demonstrated their approach on decision tree construction.

In distributed data mining one or more *data holders* provide the input data for the data mining, while one or more *data miners* cooperate in performing the data mining. A simple way to perform distributed data mining is to send all data to one data miner, and let the data miner perform the data mining. If the data contains private information such an approach clearly violates privacy, since the data miner will have full access to all data. In such a distributed environment it is thus not enough to ensure that the result of the data mining preserves privacy, we must also guarantee that privacy is not breached during the computation itself. Kantarcioglu and Clifton developed protocols to privately mine for association rules in a distributed environment[12]. The authors considered a case in which there are multiple parties that have their own confidential local databases that they do not want to share with others. The assumption is that the data is distributed horizontally, i.e., the database schema is the same for all the parties. The individual association rules together and their statistical properties were assumed to be confidential. Another assumption is that the involved parties are honest but curious, i.e., they follow the protocol but they would like to get as much information as possible from the data they receive. Under these assumptions, secure multi-part computation base protocols were employed based on commutative encryption schemes to make sure that the confidential association rules are circulated among the participating companies in encrypted form. The resulting global association rules can then be obtained in a private manner without each company knowing which rule belongs to which local database.

When data is going to be published, or handed over to a third party, it needs to be sanitized by removing sensitive information. This sensitive information could be some data values or it could be in the form of data mining models. In [2], the authors propose an approach for privacy preserving data mining which maps the original data set into a new anonymized data set preserving the correlations among the different dimensions.

Security of random perturbation methods against partial disclosure through successive querying of the database by snoopers is studied in [17]. The effect of high dimensionality in randomisation was studied by Aggarwal in [1]. In the work by Liu *et al.* the authors point out that perturbation techniques which preserve distance between data objects can be attacked if the attacker knows a small set of data selected according to the same probability distribution as the original data set[15, 16]. The attack applies principal component analysis to the perturbed data and tries to

fit it to the known data set. Liu *et al.* also propose an alternative transformation where the objects in the original data set are projected onto a subspace in a way that distance is preserved with high probability. They point out that the alternative approach is secure against the identified attack, but may not be secure against other attacks.

3 Encryption-based Techniques

Research in privacy preserving data mining started after 2000, but the cryptographic background dates back to Yao's definition and solution to the "millionaires problem" in 1982[22]. In Yao's millionaires problem two millionaires want to find out who is richer, but without revealing their wealth to the each other. In fact, the ability to compare numerical data is crucial in most data mining tasks. Yao's work initiated research in *secure multi-party computation* which is the study of the class of functions which two or more players can securely compute on their joint inputs. This is done in a way that nothing but the final result of the computation will be revealed to the parties. In particular, no party will know the inputs of the other parties. Yao later on demonstrated that any problem which can be described by a polynomial size boolean circuit of logarithmic depth can be solved securely[23]. Today we know that any computation which can be done in polynomial time by one party can be done securely by multiple parties[5]. The only ingredient needed in these generic protocols is encryption.

An important issue in secure multi-party computation is the definition of security. What does it mean that "the protocol for computing function f does not reveal too much"? Intuitively we would like our protocols to be as if all players send their inputs to an *honest third party*, which performs the computation and returns the results to the players. This perfect protocol is clearly secure, since no player sees anything else than its own inputs and outputs (in multi-party computation we do not address the privacy issues of the input itself). A formalisation of this idea is the standard definition of secure computation used today[10]. The definition requires the existence of a *simulator* which can generate the state of any (possibly dishonest) party at each step of the protocol when given the inputs and outputs of the party in the protocol. The simulated state should be such that it is not computationally feasible to tell the difference between the simulated state of the dishonest participant, and the state of the participant in a real invocation of the protocol.

Some of the well-known *public-key encryption schemes* are RSA and ElGamal. RSA encrypts messages of approximately 1024 bits in ciphertexts of 1024 bits. ElGamal is an elliptic curve based encryption which can handle messages (typically around 160 bits) that are much smaller than what RSA can handle. Public key encryption schemes are easier to use and administrate, but are slower than the so called *symmetric key encryption* schemes. DES and AES are well known symmetric

key encryption schemes. They encrypt messages of 64 and 128 bits respectively, and generate ciphertexts of the same length.

3.1 Circuit Evaluation

Many of the protocols based on encryption use the idea introduced by Yao[23]. In Yao's protocol one of the parties compute a scrambled version of a boolean circuit for evaluating the desired function. The scrambled circuit consists of encryptions of all possible bit values on all possible wires in the circuit. The number of encryptions is approximately $4m$, where m is the number of gates in the circuit. The encryptions can be symmetric key encryption, which has a typical ciphertext-length of 64 bits. The scrambled circuit is sent to the other party, which can then evaluate the circuit to get the final result. These approaches are, in general, expensive since they require complicated encryptions for each individual bit.

Many privacy preserving data mining protocols use the idea of scrambled circuits, but in order to limit the overhead of scrambled circuits, they only use scrambled circuits as sub-protocols to compute certain simple functions[14, 20].

3.2 Homomorphic Encryption

A powerful tool in computing a wide range of functions with computational security is *homomorphic encryption*. With homomorphic encryption we can avoid the bit-wise encryption from the scrambled circuits described in Section 3.1. Homomorphic encryption schemes are a special class of public key encryption schemes. The first homomorphic cryptosystem, called the Goldwasser-Micali (GM) cryptosystem, was proposed in 1984[11]. Due to its prohibitive message expansion during encryption (i.e. each bit of plaintext is encrypted as a ciphertext of eat least 1024 bits), it is not practical for data mining applications. The natural extension of the GM cryptosystem is the Benaloh cryptosystem [21], which allows the encryption of larger block sizes at a time. Although the message expansion is not as bad as in the GM cryptosystem, it is still not suitable for data mining applications. Furthermore, the fact that the decryption is based on exhaustive search over all possible plain-texts also makes the Benaloh cryptosystem unpractical for privacy preserving data mining. A more recent scheme is the Paillier cryptosystem [18], which avoids many of the drawbacks of the earlier homomorphic cryptosystems. The Paillier cryptosystem provides fast encryption and decryption algorithms, and it encrypts 1024-bit messages in ciphertexts of at least 2048-bits, which is reasonable if we work with large plaintexts.

Homomorphic encryption enables us to compute certain functions more efficiently compared to scrambled circuits. The authors of [9] use homomorphic encryption for computing secure scalar products used in privacy preserving data mining. The protocol is shown in Fig. 1, where two players, A and B, compute the scalar product of vectors $\overline{v} = (v_1, \ldots, v_d)$, and $\overline{w} = (w_1, \ldots, w_d)$, such that only B learns the scalar

product, and A learns nothing at all.

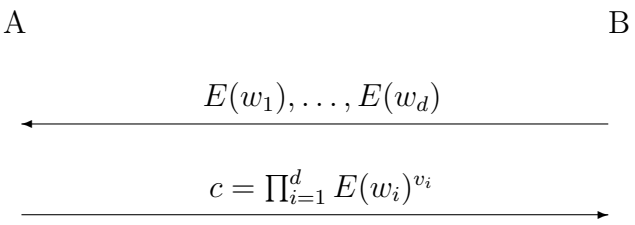


Figure 1: Computationally secure scalar product protocol ($D(c) = \overline{v} \cdot \overline{w}$).

4 Secret Sharing

Secret sharing was introduced independently by Shamir[19] and Blakley[7] in 1979. The idea is that one party has a secret which it distributes among n other parties in a way that none of the n parties alone can recover the secret. As a matter of fact the secret is shared in a way that the information of at least t of the n parties is needed to recover the secret, where t is a predefined threshold. Any attempt by less than t parties to recover the secret will fail and they will not learn *anything* about the secret.

A (t, n) *secret sharing scheme* is a set of two functions S and R. The function S is a *sharing function* and takes a *secret* s as input and creates n *secret shares*: $S(s) = (s_1, \ldots, s_n)$. The two functions are selected in a way that, for any set $I \subseteq \{1, \ldots, n\}$ of t indices $R(I, s_{I_1}, \ldots, s_{I_t}) = s$. Furthermore we require that it is *impossible* to recover s from a set of $t - 1$ secret shares. A secret sharing scheme is *additively homomorphic* if $R(I, s_{I_1} + s'_{I_1}, \ldots, s_{I_t} + s'_{I_t}) = s + s'$.

A very simple (n, n) additive secret sharing scheme is $S(s) = (r_1, \ldots, r_{n-1}, r)$, where r_i is random for $i \in \{1, \ldots, n-1\}$, and $r = s - \sum_{i=1}^{n-1} r_i$. To recover s all secret shares are added: $s = r + \sum_{i=1}^{n-1} r_i$. If even one secret share is missing nothing is known about s.

Shamir secret sharing was used by Ben-Or *et al.* [6] in 1988 to show that any function of n inputs can be computed by n parties such that no coalition of less than $n/3$ of the parties can gain *any* additional information about the honest parties inputs (even if they do not behave according to the prescribed algorithm). If we assume that all parties behave *semi-honestly* (i.e. they follow the protocol), then no coalition of less than $n/2$ of the parties can gain any information about the inputs of the honest parties. The protocol uses the additively homomorphic property of Shamir secret sharing. The idea is that addition and multiplication together is enough to evaluate any function (in particular addition and multiplication over Z_2 is a universal set of boolean operations). The bottleneck of the algorithm in [6] is multiplication. Since Shamir secret sharing is not multiplicatively homomorphic, in

order to perform a multiplication, a special "degree reduction step" has to be performed. This degree reduction requires that all parties secret share a new number (a total of n^2 new messages for each multiplication). For most data-mining applications this degree reduction step is too costly, since a vast number of multiplications is common. Another limitation of the generic multi party computation based on secret sharing is when some parties may behave dishonestly. They may try to gain extra information by deviating from the prescribed protocol. To avoid this, a special variant of secret sharing is used. This variant, called *verifiable secret sharing* adds extra information to each secret share, such that any set of players, at any time in the protocol, can verify that the shares they have are consistent. Both the extra information in the secret shares, and the interaction required to verify a secret sharing adds extra communication overhead to the protocols.

4.1 The Coopetative Model

Data holders that participate in distributed data mining have naturally an interest in the result of the data mining. They are, however, understandably reluctant to share their private data with others to either protect their interests or meet privacy requirements imposed by authorities and/or clients. Data holders, in other words, are ready to *cooperate* with each other to extract useful information from combined data while *competition* among them dictates that individual data is not revealed to others.

The term *coopetation* is used in economics to refer to cooperation between competing entities to improve the overall value of their market. This is quite similar to the distributed data mining scenario where data holders behave with similar motivations. In the coopetative model data holders provide inputs to a relatively small set of data mining servers or so called third parties, which are assumed semi-honest (i.e. they are honest but curious; they follow the protocol steps, but are interested in any leaked information). Some of the data holders can actively participate in the distributed data mining playing the role of third parties. The non-collusion property must be satisfied by certain sets of third parties. At the end of the protocol the data miner, which can be either a separate entity or one of the data holders, will have the outcome as an output.

Some of the benefits of the coopetative model are:

- Very efficient data mining protocols can be constructed.
- The major workload can be put on a small set of dedicated servers which are better protected and regulated.
- Only these small sets of servers need to posses the necessary hardware, software and know-how to perform data mining.

- Encryption is avoided, thus key distribution and other problems related with encryption are no longer an issue.

The basic version of the coopetative model [13] requires two dedicated third parties and a miner. Data holders secret shares their data and send each share to one of the third parties. For sake of simplicity we can assume that the private input of each data holder is an integer x and the data holder creates two shares r and $x - r$ where r is a randomly selected integer. The share r is sent to the first third party and the share $x - r$ is sent to the other. Clearly both shares are random when observed alone and no single entity (adversary, third party, or miner) can obtain any information about the private input x. The private input can only be revealed when two shares are put together, which never happens in the coopetative model.

The third parties work on the individual shares and compute algebraic operations such as numerical difference and comparison on the shares, which are the fundamental operations in many data mining applications (e.g. constructing decision trees, association rule mining and clustering). The result of these operations are the shares of the final outcome of the computation, which can be obtained only by the data miner.

In order for the third parties to work on shares, they need to employ secret sharing schemes which are homomorphic with respect to the operations they perform. For instance, additive secret sharing described above is homomorphic with respect to addition (and subtraction): adding shares pairwise gives an additive sharing of the sum of the secrets. Therefore, the additive secret sharing scheme can directly be used in numerical difference operations in clustering algorithms.

5 Discussion and Comparison

When comparing the efficiency of two alternative protocols for a specific task there are three factors to consider: the computation cost, communication cost, and the number of rounds in the protocols. We will compare encryption-based techniques and secret sharing with respect to these three factors in the following.

Public key encryption schemes are (by definition) based on computationally difficult problems, and thus require expensive operations such as modular exponentiation of large numbers (in the order of 1000 bits). In contrast it is very efficient to compute secret shares when using e.g. Shamir secret sharing or the simple additive secret sharing described in Section 4. Sharing a secret with Shamir secret sharing consists in choosing a random polynomial and evaluating it in n points. The polynomial is chosen over the same field as the secret, which means that usually all computations are done with ordinary integers.

As mentioned in Section 3 public key encryption schemes create ciphertexts of at least 1024 bits (with the exception of Elliptic curve based encryption schemes). If we want to use the homomorphic properties of an encryption scheme we have to

encrypt each input in it's own ciphertext. Often this will mean that we encrypt 32-bit numbers in 1024 bits (giving an overhead of 32). If we use circuit evaluation techniques we are forced to encrypt each bit as at least 160 bits if we use elliptic curve cryptography. In contrast, secret sharing creates n shares of each input, where each share is of the same size as the secret. We thus always have an overhead of n. Note, however, that some secret sharing schemes, like the Asmuth Bloom scheme[4], create shares which are larger than the secret.

If Shamir secret sharing is used as described in Section 4 each multiplication requires that each pair of parties exchange information. Having to wait for the transmission of these messages at each multiplication clearly slows down a protocol, in comparison Yao's circuit evaluation only requires one round of communication. Some work has been done to minimise the number of rounds needed by secret sharing based techniques[8], though they do not give constant round complexity as in the case of Yao's protocol. It is still an open problem to fully classify the problems which can be solved with a constant number of rounds with unconditional security.

We should note that not all problems can be solved with unconditional security. A very important fact, from a data mining point of view, is that unconditionally secure scalar products between two parties is impossible. Any two-party data mining algorithm which applies scalar products (between secret vectors held by the two parties) has to rely on either encryption based techniques, or external parties.

6 Conclusion

Privacy preserving data mining is an ongoing research area and there are a lot of issues that needs to be addressed. First of all, the databases that are collected for mining are huge, and scalable techniques for privacy preserving data mining are needed to handle these data sources. Secret sharing based methods can be considered one step forward in scalable privacy preserving data mining. The degree of data distribution could also be a problem when we consider a grid, peer-to-peer, or ubiquitous computing environments. Techniques which minimize the amount of computation and data transfer are needed in highly distributed environments. New data types such as spatio-temporal data collected by location-based services, and other mobile service providers pose new types of threats to privacy, and existing techniques for privacy preserving data mining may not be adequate to handle these types of data.

References

[1] C. C. Aggarwal. On randomization, public information and the curse of dimensionality. In *ICDE*, pages 136–145, 2007.

[2] C. C. Aggarwal and P. S. Yu. A condensation approach to privacy preserving

data mining. In *EDBT*, pages 183–199, 2004.

[3] R. Agrawal and R. Srikant. Privacy-preserving data mining. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, May 16-18, 2000, Dallas, Texas, USA*, pages 439–450. ACM, 2000.

[4] C. Asmuth and J. Bloom. A modular approach to key safeguarding. *IEEE Transactions on Information Theory*, 29(2):208–210, March 1983.

[5] D. Beaver, S. Micali, and P. Rogaway. The round complexity of secure protocols. In *STOC '90: Proceedings of the twenty-second annual ACM symposium on Theory of computing*, pages 503–513. ACM, 1990.

[6] M. Ben-Or, S. Goldwasser, and A. Wigderson. Completeness theorems for non-cryptographic fault-tolerant distributed computation. In *Proceedings of the nineteenth annual ACM conference on Theory of computing (STOC)*, pages 1–10. ACM Press, 1988.

[7] G. R. Blakley. Safeguarding cryptographic keys. In *Proceedings of AFIPS 1979 National Computer Conference*, pages 313–317, June 1979.

[8] R. Cramer and I. Damgård. Secure distributed linear algebra in a constant number of rounds. In *Advances in Cryptology - CRYPTO 2001: 21st Annual International Cryptology Conference*, pages 119–138, 2001.

[9] B. Goethals, S. Laur, H. Lipmaa, and T. Mielikäinen. On private scalar product computation for privacy-preserving data mining. In *Information Security and Cryptology — ICISC 2004*, volume 3506 of *Lecture Notes in Computer Science*, pages 104–120. Springer-Verlag, 2005.

[10] O. Goldreich. *The Foundations of Cryptography — Volume 2, Basic Applications*. Cambridge University Press, May 2004.

[11] S. Goldwasser and S. Micali. Probabilistic encryption. *J. COMP. SYST. SCI.*, 28(2):270–299, March 1984.

[12] M. Kantarcioglu and C. Clifton. Privacy-preserving distributed mining of association rules on horizontally partitioned data. *IEEE Trans. Knowl. Data Eng.*, 16(9):1026–1037, 2004.

[13] S. V. Kaya, T. B. Pedersen, E. Savaş, and Y. Saygin. Efficient privacy preserving distributed clustering based on secret sharing. In *LNAI 4819 PAKDD 2007*, pages 280–291. Springer, 2007.

[14] Y. Lindell and B. Pinkas. Privacy preserving data mining. In *Advances in Cryptology (CRYPTO'00)*, volume 1880 of *Lecture Notes in Computer Science*, pages 36–53. Springer-Verlag, 2000.

[15] K. Liu, C. Giannella, and H. Kargupta. An attacker's view of distance preserving maps for privacy preserving data mining. In *Knowledge Discovery in Databases: PKDD 2006, 10th European Conference on Principles and Practice of Knowledge Discovery in Databases*, volume 4213 of *Lecture Notes in Computer Science*, pages 297–308. Springer, 2006.

[16] K. Liu, H. Kargupta, and J. Ryan. Random projection-based multiplicative data perturbation for privacy preserving distributed data mining. *IEEE Trans. Knowl. Data Eng.*, 18(1):92–106, 2006.

[17] K. Muralidhar and R. Sarathy. Security of random data perturbation methods. *ACM Trans. Database Syst.*, 24(4):487–493, 1999.

[18] P. Paillier. Public-key cryptosystems based on composite degree residuosity classes. In *Advances in Cryptology EUROCRYPT'99, LNCS 1592*, pages 223–238. Springer, 1999.

[19] A. Shamir. How to share a secret. *Communications of the ACM*, 22(11):612–613, November 1979.

[20] J. Vaidya and C. Clifton. Privacy-preserving k-means clustering over vertically partitioned data. In *KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 206–215, New York, NY, USA, 2003. ACM Press.

[21] J. Vijayan. House committee chair wants info on cancelled dhs data-mining programs. *Computer World*, September 18, 2007.

[22] A. C. Yao. Protocols for secure computations (extended abstract). In *23rd Annual Symposium on Foundations of Computer Science*. IEEE, 1982.

[23] A. C. Yao. How to generate and exchange secrets. In *Proceedings of the twenty-seventh annual IEEE Symposium on Foundations of Computer Science*, pages 162–167. IEEE Computer Society, 1986.

Numerical Data Masking Techniques for Maintaining Sub-Domain Characteristics

Krish Muralidhar
Gatton Research Professor
University Of Kentucky
Lexington KY 40513

Rathindra Sarathy
Ardmore Professor
Oklahoma State University
Stillwater OK 74078

1. Introduction

In a recent paper, Muralidhar and Sarathy (2007a) showed that data shuffling and sufficiency-based linear models performed better than other techniques for masking numerical data. This conclusion was based on assessment of both data utility and disclosure risk. Specifically, in terms of data utility, data shuffling (Muralidhar and Sarathy 2006) was shown to: (a) maintain the marginal distribution of all the confidential variables to be the same after data masking, and (b) maintain monotonic relationships between the variables. The sufficiency-based linear models approach (Burridge 2003, Muralidhar and Sarathy 2007b) was shown to maintain the mean vector and covariance matrix of the masked data to be identical to that of the original data. In terms of disclosure risk, both methods were shown to minimize disclosure by ensuring that, given the non-confidential variables, the original and masked data were independent.

In addition to the traditionally used measures for assessing data utility, one of the desirable properties of masking techniques is that they maintain the characteristics of the data not just in the overall data set, but also in sub-domains of data (Winkler 2006). For numerical variables, it is possible to generate an infinite number of possible sub-groups and it becomes difficult to evaluate all possible sub-groups. However, categorical non-confidential variables usually result in a finite number of sub-domains. Furthermore, data consisting of both categorical and numerical variables are very common in practice. Hence, we evaluate the performance of the two selected techniques (sufficiency-based linear models and data shuffling) on sub-domains. For each sub-domain, we evaluate the extent of information loss and disclosure risk resulting from the masked data.

2. A Brief Introduction to Sufficiency-based Linear Models and Data Shuffling

2.1. Sufficiency-based Linear Models

Masking approaches based on linear models generate the perturbed values **Y** using some variation of the following model:

$$\mathbf{Y} = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1\mathbf{S} + \boldsymbol{\beta}_2\mathbf{X} + \boldsymbol{\varepsilon}, \qquad (1)$$

where **S** represents a set of (categorical and numerical) non-confidential variables, **X** represents a set of confidential variables, and **e** represents the noise term. Muralidhar et al. (1999) originally proposed a model of the form as shown in (1), but with the requirements that the covariance matrix of the released data (**S** and **Y**) be the same as that of (**S** and **X**). This specification imposes a specific structure on the covariance of **ε**. In order to improve the disclosure risk characteristics they proposed a modified model of the form

$$\mathbf{Y} = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1\mathbf{S} + \boldsymbol{\varepsilon}. \qquad (2)$$

In equation (2), $\boldsymbol{\beta}_1 = \boldsymbol{\Sigma}_{XS}\boldsymbol{\Sigma}_{SS}^{-1}$, $\boldsymbol{\beta}_0 = \boldsymbol{\mu}_Y - \boldsymbol{\Sigma}_{XS}\boldsymbol{\Sigma}_{SS}^{-1}\boldsymbol{\mu}_S$, and $\boldsymbol{\Sigma}_{ee} = \left(\boldsymbol{\Sigma}_{XX} - \boldsymbol{\Sigma}_{XS}\boldsymbol{\Sigma}_{SS}^{-1}\boldsymbol{\Sigma}_{SX}\right)$, where $\boldsymbol{\Sigma}_{XX}$ is the covariance of **X**, $\boldsymbol{\Sigma}_{SS}$ is the covariance of **S**, $\boldsymbol{\Sigma}_{XS}$ is the covariance between (**X** and **S**), and $\boldsymbol{\Sigma}_{ee}$ is the covariance of the noise term **e**. With these specifications, for large data sets, the mean vector and covariance matrix of the released data (**S** and **Y**) will be the same as that of the original data (**S** and **X**). However, there is some information loss in estimates of the covariance matrix, due to sampling error in smaller data sets. Since **Y** is generated as a function of **S** and **ε** this procedure also minimizes the risk of both identity and value disclosure.

An important variant of the linear model was suggested by Burridge (2003). In this approach, by appropriately generating the values of **ε**, it is possible to ensure that the mean vector and covariance matrix of the released data are *identical* to that of the original data. Hence, for all statistical analyses for which the mean vector and covariance matrix are sufficient statistics, the results of the analysis using the masked data will yield *identical results* to that using the original data. That is, the (statistical) information loss will be *zero*. Note that for most traditional statistical analyses (including, but not limited to, comparison of means, ANOVA, regression analysis, and even such multivariate procedures such as canonical correlation analysis), the mean vector and covariance matrix serve as sufficient statistics. Hence, if this procedure is employed to mask the data, a user who analyzes the masked data will get exactly the same results as using the original unmasked data. In addition, this procedure also minimizes disclosure risk.

Muralidhar and Sarathy (2007) recently proposed a further modification of this linear model in equation (1) with the following restrictions:

$$\boldsymbol{\beta}_0 = (\mathbf{I} - \boldsymbol{\beta}_2)\boldsymbol{\mu}_X - \boldsymbol{\beta}_1\boldsymbol{\mu}_S, \tag{3}$$

$$\boldsymbol{\beta}_1 = (\mathbf{I} - \boldsymbol{\beta}_2)\boldsymbol{\Sigma}_{XS}\boldsymbol{\Sigma}_{SS}^{-1}, \text{ and} \tag{4}$$

$$\boldsymbol{\Sigma}_{\varepsilon\varepsilon} = \left(\boldsymbol{\Sigma}_{XX} - \boldsymbol{\Sigma}_{XS}\boldsymbol{\Sigma}_{SS}^{-1}\boldsymbol{\Sigma}_{SX}\right) - \boldsymbol{\beta}_2\left(\boldsymbol{\Sigma}_{XX} - \boldsymbol{\Sigma}_{XS}\boldsymbol{\Sigma}_{SS}^{-1}\boldsymbol{\Sigma}_{SX}\right)\boldsymbol{\beta}_2^T. \tag{5}$$

With the above specifications and appropriately selecting the value of $\boldsymbol{\beta}_2$ it is possible to ensure that the masked data (**Y** and **S**) has exactly the same mean vector and covariance matrix as the original data (**X** and **S**). That is, this approach preserves sufficient statistics underlying linear relationships. Consequently, there is zero information loss in estimating any of the linear relationships among the different variables. When $\boldsymbol{\beta}_2$ is zero, this model reduces to that shown in equation (2). For non-zero $\boldsymbol{\beta}_2$ the resulting masked variables *do not* provide the lowest possible level of disclosure risk. However, these masked values may have greater acceptance among users who may have reservations using the more "synthetic" data generated from the model in equation (2).

While the sufficiency-based linear models (SBLM) procedure provides significant advantages over other perturbation procedures, it is not without problems. First and foremost, this procedure results in information loss in the marginal distribution of the masked variable (**Y**). The exact form of the distribution of **Y** would depend on the selection of the distribution for the error term **ε**. However, unless **X** was normally distributed, the marginal distribution of **Y** will be different from that of **X**. One common problem that arises as a consequence is that the masked values may consist of negative values whereas the original data may be all positive. In addition, while this procedure maintains linear relationships among all variables, non-linear relationships are not preserved in the masked data. Thus, while this procedure is a *complete solution* to the masking problem when the joint distribution of (**X** and **S**) is multivariate normal, resulting in zero information loss and zero incremental disclosure risk, in other cases, it has some shortcomings.

2.2. Data Shuffling

Data shuffling is a new patented procedure (US Patent # 7200757) developed by Muralidhar and Sarathy (2006). It is a hybrid procedure where the original variables are first perturbed using the copula based perturbation approach (Sarathy et al. 2002). The resulting perturbed values are then reverse-mapped on to the original values, resulting in the shuffled data set. Superficially, data shuffling can be considered to be a multivariate version of data swapping since it is performed on the entire data set rather than on a variable by variable basis.

Data shuffling has the following desirable properties. First and foremost, the perturbed values are generated independent of **X** (given **S**) and hence have no incremental disclosure risk. Second, like data swapping, the shuffled values are actually the original values of the confidential variables assigned to a different observation. Hence, the marginal distribution of the masked data is identical to the marginal distribution of the original data. Third, the use of the copula-based perturbation approach enables data shuffling to maintain the rank order correlation of the masked data to be the same as that of the original data. This implies that data shuffling results in minimal information loss in linear and monotonic non-linear relationships among variables. It does not maintain non-monotonic non-linear relationships.

3. Empirical Assessment

We performed an empirical assessment of the two masking techniques using two data sets. The first masking technique used was data shuffling that does not require any parameter specifications. The second masking technique was the SBLM procedure with the requirement that $\boldsymbol{\beta}_2$ be a diagonal matrix with the value d ($0 \leq d \leq 1$) in the diagonal and 0 in the off-diagonal terms. This simple specification implies that when d = 0, the resulting model is the one shown in equation (6) and when d = 1, the entire data set is released unmodified. Thus, the selection of d directly influences the extent to which the original values are used in the masking. Note that when d > 0, this method does not provide minimum security.

3.1. Experimental Assessment Using Simulated Data Set

The first data set was simulated and consisted of 50000 observations. The data consisted of 3 categorical non-confidential variables Gender (male or female), Marital Status (married or other), and Age group (1 to 6). The 3 confidential numerical variables (Home value, Mortgage balance, Total net value of assets) were generated using the NORTA approach for generating related multivariate non-normal variables. Of the three confidential variables, two (Home value and Mortgage balance) had non-normal marginal distributions, while the third had a normal distribution. The relationship between the last two variables was linear while the other relationships were non-linear. Twenty four sub-groups were formed as a combination of the Gender × Martial status × Age group. Data shuffling was applied to the entire data set. In addition, 3 different levels of masking were applied for linear model approach (d = 0.00, 0.50, 0.90). As indicated earlier, when d = 0.00, given the non-confidential variables, the perturbed variables are independent of the original variables and are sometimes considered synthetic data.

3.1.1. Assessment of Disclosure Risk. As indicated earlier, the first step in the assessment of the masking techniques was to compute the risk of identity disclosure for each sub-domain. Table 1 provides the results of the *identity disclosure* (or *re-identification risk*) assessment performed using the procedure suggested by Fuller (1993). There are many approaches for assessing identity disclosure and we could use any one of these procedures. However, the primary objective of this assessment is to compare the different methods rather than assess the extent of disclosure. While the specific results of using another procedure for assessing identity disclosure may be different, the relative performance of the different methods will be the same. Table 2 provides, for each sub-group defined by the categorical variables (a total of 24 sub-groups), the number of observations in each sub-group and the number of observations that were re-identified. As indicated earlier, when shuffling and perturbation with d = 0.00 are used to mask the variables, within a given sub-group, the original and masked variables are independent. Hence, the probability of re-identification within a sub-group is $(1/n_k)$ where n_k is the size of the sub-group. The results in Table 2 clearly show that this is indeed the case. The probability of re-identification is much higher for the other perturbed values, with the higher re-identification occurring when d = 0.90. Thus, in terms of disclosure risk, it is easy to see that the data shuffling and perturbation with d = 0.00 provide the best results, with re-identification occurring by chance alone.

It is also easy to assess the risk of *value disclosure*. As indicated earlier, for a given sub-domain, the shuffled data and perturbed data with d = 0.00 are independent of the original data. This implies that the covariance between the original and masked data are close to zero for shuffled data and exactly 0.00 for the perturbed data with d = 0.00. Hence, the correlation between the original and masked data for these two methods will be 0.00, resulting in no predictive ability. By contrast, for the other two approaches, the correlation between the original and masked variables will be d and the intruder would be able to explain d^2 proportion of the variability in the values of the original variables using the masked variables.

As an illustration, consider the sub-group Gender = 0, Marital = 0, and Age = 1. The mean and standard deviation of the Home value variable in this sub-group are 2.872 and 8.643, respectively. With only this information, for any observation in this sub-set, the best prediction of a 99% interval estimate of the true value of the Home value variable would have an interval of approximately (3 × 8.643). Now assume that the shuffled data is released. The correlation between the original and the shuffled home values is 0.03. Hence, if we perform regression analysis to predict the original value of the confidential variable using the shuffled values, the resulting R^2 would be 0.0009 resulting in a standard error of 8.642. Using this information, a simple 99% confidence interval would have an interval of approximately (3 × 8.642), which for all practical purposes is almost exactly the same as the interval constructed without access to the masked data. In other words, releasing the shuffled data does not allow the intruder to estimate the value of the confidential variable with any greater level of security. Similar results will be observed for the perturbed data when d = 0.00.

Gender	Marital	Age	Total number of observations	Number of Observations Identified			
				Shuffled	Perturbed (0.00)	Perturbed (0.50)	Perturbed (0.90)
0	0	1	1220	3	1	5	40
		2	1181	0	1	13	47
		3	1193	1	1	8	42
		4	1162	3	1	4	39
		5	1159	2	1	5	29
		6	1181	0	1	4	42
	1	1	4672	2	1	7	56
		2	4723	0	1	12	73
		3	4671	1	1	9	54
		4	4719	2	1	5	48
		5	4635	1	1	6	61
		6	4650	2	1	7	58
1	0	1	515	2	1	5	25
		2	468	2	1	6	33
		3	502	3	1	1	30
		4	511	0	1	3	24
		5	503	2	1	2	34
		6	464	0	1	2	21
	1	1	2019	2	1	7	59
		2	1968	0	1	6	43
		3	2044	0	1	3	49
		4	1940	0	1	4	35
		5	1960	1	1	5	52
		6	1940	0	1	4	50

Table 1. Risk of Identity Disclosure (Simulated Data)

The above result does not hold for the other two perturbation parameters (d = 0.50, 0.90). When d = 0.50, if we perform a regression analysis to predict the original Home value variable using the perturbed values, the resulting standard error is 7.485. A 99% confidence interval estimate would have an interval of approximately (3 × 7.485). This implies that the intruder is able to gain a more accurate estimate compared to not having the perturbed values. When d = 0.90, the resulting standard error from the regression analysis is 3.767. If we construct a 99% confidence interval using this information, it results in an interval of approximately (3 × 3.767). Compared to the original interval, the width of this interval is less than 50% of the original width. This allows the intruder to gain a far more accurate estimate of the value of the confidential variable.

Thus, an intruder would have a much better estimate of the original values when the data is masked using the perturbation approach with d = 0.50 and 0.90. In conclusion, when considering disclosure risk, because of their inherent property of conditional independence, data shuffling and perturbation with d = 0.00 perform better than perturbation with d = 0.50 and 0.90. If disclosure risk were the only criterion, data shuffling and perturbation with d = 0.00 would be the preferred methods.

5.1.2. Assessment of Information Loss. In assessing information loss, we focus our attention on sub-domain performance, rather than on the entire data set. We know that, for the entire data set, *data shuffling maintains the marginal distribution of the masked variables to be exactly the same as that of the original variables*. By contrast, the SBLM approach is capable of maintaining the marginal distribution of the variables only when the variable has a normal distribution.

One of the attractive features of the data shuffling procedures is that *the marginal distribution of the shuffled data within any sub-group defined by the non-confidential categorical variables is exactly the same as that of the original variable*. The marginal distribution of the perturbed data are different from that of the original data for sub-groups. To illustrate this, consider the case for the sub-group where Gender = 0, Marital Status = 0, and Age = 1. Figure 1 provides the marginal distribution of the original and the 3 perturbed data sets for the Home value variable. We do not provide the shuffled data since it will coincide exactly with the original data.

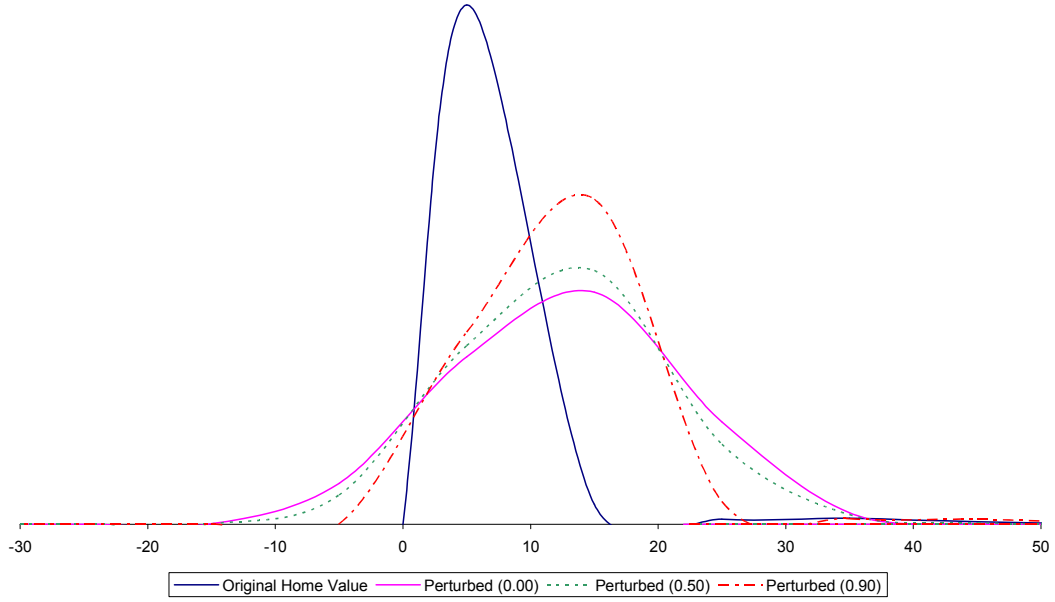


Figure 1. Sub-domain Marginal Distribution of Home Value (Simulated Data)

As can be seen from this example, the marginal distribution of the perturbed data differs considerably from the original data even when d = 0.90. Thus, the "addition of noise" results in a marginal distribution that is closer to normality than the original data. Note that, for the Net Assets variable, the marginal distribution of all the masked variables for all the sub-groups will be similar since the original variable was normally distributed. As discussed earlier, one other attractive feature of all the methods considered in this study is that the mean and variance of all the variables in every sub-group defined by the non-confidential categorical variables will be exactly the same as that of the original data. Hence, we have not provided this data. However, in addition to maintaining the mean and variance, the *shuffled data maintains all the univariate marginal characteristics of the masked data to be the same as that of the original data*.

To assess the extent to which the methods maintain relationships among variables, we computed the product moment correlation between the variables in each sub-group. The results of this analysis are provided in Table 2. As expected, *the product moment correlations of the original data and those of the perturbed data are exactly the same for the data set as a whole and for every sub-group*. The shuffled data does not provide exactly the same results, but *the product moment correlations of the shuffled data and those of the original data are very similar for the data set as a whole and for every sub-group*. Thus, in terms of maintaining product moment correlation, the SBLM approach seems to perform better than the shuffling approach. This is expected since the SBLM approach is intended to maintain first and second order moments (and consequently correlation) among the variables. However, this does not necessarily mean that it is superior to data shuffling as the following discussion shows.

Consider the relationship between the variables Mortgage balance and Net asset value. Figure 2.a provides a scatter plot of the original values with Net asset values on the X-axis and Mortgage balance on Y-axis. It is clear from this figure that the relationship between the two variables is non-linear. In cases where the relationship is non-linear, product moment correlation which measures only the linear relationship is not an appropriate measure. The product moment correlation for these two variables in the data set is 0.719 and all three perturbed values maintain this correlation. By contrast, the correlation between the corresponding shuffled variables slightly different (0.718). Now consider a plot of the perturbed values of Mortgage balance and Net asset values (with d = 0.00) overlaid on top of the original scatter plot (Figure 2.b). Figure 2.b clearly indicates that the perturbation approach has considerably modified the relationship between the variables; the original relationship was non-linear while the perturbed data is almost linear. A plot of the shuffled values of Mortgage balance and Net asset values overlaid on the original data is shown in Figure 2.c. This figure clearly indicates that the shuffled data maintains the (monotonic) non-linear relationship between the two variables better than the perturbed data. Thus, although the SBLM approach maintains the product moment correlation exactly, it does not necessarily maintain non-linear relationships between the variables. By contrast, while the shuffled data does not maintain the product moment correlation exactly, it is capable of maintaining monotonic non-linear correlations much better than the perturbed data.

Sub-Group	Correlation between Home Value and Mortgage Balance					Correlation between Home Value and Net Assets					Correlation between Mortgage Balance and Net Assets				
	O	S	P1	P2	P3	O	S	P1	P2	P3	O	S	P1	P2	P3
1	0.338	0.363	0.338	0.338	0.338	0.373	0.328	0.373	0.373	0.373	0.693	0.701	0.693	0.693	0.693
2	0.216	0.245	0.216	0.216	0.216	0.214	0.274	0.214	0.214	0.214	0.700	0.707	0.700	0.700	0.700
3	0.402	0.306	0.402	0.402	0.402	0.375	0.319	0.375	0.375	0.375	0.707	0.702	0.707	0.707	0.707
4	0.294	0.338	0.294	0.294	0.294	0.282	0.277	0.282	0.282	0.282	0.705	0.698	0.705	0.705	0.705
5	0.201	0.251	0.201	0.201	0.201	0.250	0.267	0.250	0.250	0.250	0.707	0.716	0.707	0.707	0.707
6	0.320	0.355	0.320	0.320	0.320	0.337	0.344	0.337	0.337	0.337	0.746	0.755	0.746	0.746	0.746
7	0.266	0.229	0.266	0.266	0.266	0.230	0.222	0.230	0.230	0.230	0.698	0.695	0.698	0.698	0.698
8	0.313	0.318	0.313	0.313	0.313	0.281	0.292	0.281	0.281	0.281	0.695	0.705	0.695	0.695	0.695
9	0.276	0.253	0.276	0.276	0.276	0.264	0.264	0.264	0.264	0.264	0.694	0.697	0.694	0.694	0.694
10	0.195	0.210	0.195	0.195	0.195	0.179	0.196	0.179	0.179	0.179	0.708	0.708	0.708	0.708	0.708
11	0.284	0.285	0.284	0.284	0.284	0.274	0.261	0.274	0.274	0.274	0.710	0.700	0.710	0.710	0.710
12	0.259	0.250	0.259	0.259	0.259	0.262	0.243	0.262	0.262	0.262	0.728	0.723	0.728	0.728	0.728
13	0.288	0.243	0.288	0.288	0.288	0.244	0.256	0.244	0.244	0.244	0.698	0.712	0.698	0.698	0.698
14	0.321	0.294	0.321	0.321	0.321	0.310	0.351	0.310	0.310	0.310	0.718	0.730	0.718	0.718	0.718
15	0.356	0.371	0.356	0.356	0.356	0.364	0.376	0.364	0.364	0.364	0.705	0.751	0.705	0.705	0.705
16	0.386	0.354	0.386	0.386	0.386	0.329	0.325	0.329	0.329	0.329	0.694	0.664	0.694	0.694	0.694
17	0.387	0.352	0.387	0.387	0.387	0.393	0.418	0.393	0.393	0.393	0.707	0.714	0.707	0.707	0.707
18	0.195	0.208	0.195	0.195	0.195	0.193	0.228	0.193	0.193	0.193	0.737	0.692	0.737	0.737	0.737
19	0.320	0.389	0.320	0.320	0.320	0.338	0.356	0.338	0.338	0.338	0.676	0.662	0.676	0.676	0.676
20	0.349	0.264	0.349	0.349	0.349	0.288	0.259	0.288	0.288	0.288	0.692	0.663	0.692	0.692	0.692
21	0.357	0.303	0.357	0.357	0.357	0.348	0.318	0.348	0.348	0.348	0.703	0.714	0.703	0.703	0.703
22	0.239	0.236	0.239	0.239	0.239	0.216	0.227	0.216	0.216	0.216	0.701	0.710	0.701	0.701	0.701
23	0.289	0.328	0.289	0.289	0.289	0.272	0.292	0.272	0.272	0.272	0.707	0.717	0.707	0.707	0.707
24	0.307	0.263	0.307	0.307	0.307	0.275	0.253	0.275	0.275	0.275	0.721	0.727	0.721	0.721	0.721
Entire Data	0.223	0.220	0.223	0.223	0.223	0.201	0.197	0.201	0.201	0.201	0.719	0.718	0.719	0.719	0.719

Legend: O = Original Data; S = Shuffled Data; P1 = Perturbed (0.00); P2 = Perturbed (0.50); P3 = Perturbed (0.90)

Table 2. Original and Masked Product Moment Correlation (Simulated Data)

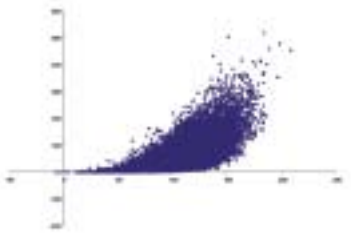


Figure 2.a. Scatter Plot of Net Asset Value and Mortgage Balance (Original Data)

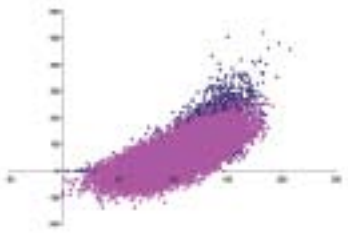


Figure 2.b. Scatter Plot of Net Asset Value and Mortgage Balance (Original and Perturbed)

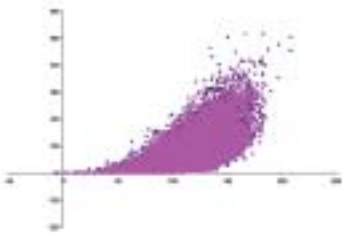


Figure 2.c. Scatter Plot of Net Asset Value and Mortgage Balance (Original and Shuffled)

In situations where the relationship is non-linear, in place of product moment correlation, rank order correlation is used to measure the strength of the relationship. The rank order correlation results, provided in Table 3, indicate that the shuffled data maintain rank order correlation better than the perturbed data. This is to be expected since the shuffling procedure attempts to maintain all monotonic relationships, while the SBLM approach only deals with linear relationships. Note that the shuffling procedure is able to maintain the rank order correlation of the masked data to be very close to that of the original data both at the

overall and sub-group level. This is a significant advantage of the shuffling approach over the SBLM approach. It is also important to note that *any approach based on linear models* (simple additive noise, Kim's method, multiple imputation, among others) are susceptible to the same "linearization" of non-linear relationships. Currently, only data shuffling offers the ability to maintain monotonic relationships among variables.

Sub-Group	Correlation between Home Value and Mortgage Balance					Correlation between Home Value and Net Assets					Correlation between Mortgage Balance and Net Assets				
	O	S	P1	P2	P3	O	S	P1	P2	P3	O	S	P1	P2	P3
1	0.553	0.575	0.329	0.340	0.365	0.624	0.648	0.368	0.368	0.376	0.762	0.786	0.680	0.679	0.708
2	0.527	0.535	0.211	0.216	0.234	0.608	0.625	0.214	0.224	0.259	0.772	0.768	0.680	0.682	0.722
3	0.536	0.539	0.371	0.379	0.388	0.617	0.605	0.357	0.360	0.396	0.764	0.743	0.692	0.696	0.731
4	0.516	0.535	0.275	0.282	0.290	0.603	0.615	0.255	0.255	0.291	0.741	0.744	0.688	0.689	0.718
5	0.540	0.554	0.201	0.205	0.255	0.622	0.647	0.237	0.240	0.285	0.763	0.757	0.692	0.691	0.723
6	0.566	0.566	0.301	0.305	0.315	0.646	0.663	0.322	0.316	0.336	0.792	0.792	0.735	0.737	0.770
7	0.529	0.531	0.252	0.258	0.266	0.623	0.613	0.216	0.224	0.247	0.761	0.767	0.674	0.682	0.707
8	0.523	0.503	0.296	0.300	0.307	0.604	0.591	0.266	0.272	0.297	0.768	0.770	0.677	0.684	0.718
9	0.535	0.532	0.264	0.268	0.282	0.608	0.605	0.253	0.257	0.292	0.759	0.761	0.676	0.679	0.714
10	0.538	0.525	0.183	0.195	0.220	0.613	0.612	0.172	0.183	0.224	0.761	0.759	0.693	0.693	0.724
11	0.539	0.540	0.279	0.285	0.301	0.619	0.621	0.263	0.267	0.297	0.752	0.756	0.692	0.699	0.723
12	0.538	0.532	0.242	0.247	0.273	0.623	0.632	0.245	0.256	0.291	0.768	0.764	0.719	0.718	0.741
13	0.590	0.606	0.267	0.302	0.330	0.669	0.672	0.257	0.273	0.315	0.798	0.816	0.685	0.683	0.705
14	0.510	0.550	0.304	0.310	0.310	0.640	0.652	0.295	0.305	0.321	0.764	0.788	0.704	0.698	0.723
15	0.539	0.552	0.351	0.339	0.342	0.633	0.617	0.349	0.336	0.357	0.752	0.783	0.706	0.689	0.712
16	0.560	0.564	0.378	0.370	0.366	0.598	0.631	0.337	0.317	0.322	0.732	0.742	0.670	0.682	0.692
17	0.523	0.530	0.374	0.365	0.370	0.597	0.603	0.410	0.399	0.412	0.754	0.782	0.678	0.677	0.707
18	0.569	0.542	0.161	0.168	0.217	0.650	0.663	0.170	0.175	0.244	0.763	0.732	0.710	0.713	0.747
19	0.536	0.562	0.314	0.328	0.353	0.638	0.648	0.316	0.332	0.376	0.762	0.758	0.656	0.665	0.697
20	0.538	0.530	0.332	0.330	0.323	0.622	0.607	0.273	0.273	0.284	0.755	0.751	0.676	0.680	0.698
21	0.521	0.525	0.351	0.358	0.364	0.601	0.609	0.328	0.341	0.365	0.759	0.761	0.680	0.685	0.719
22	0.553	0.573	0.228	0.233	0.251	0.638	0.641	0.223	0.226	0.261	0.760	0.766	0.688	0.696	0.727
23	0.521	0.541	0.279	0.291	0.314	0.598	0.610	0.262	0.274	0.316	0.762	0.773	0.696	0.697	0.715
24	0.554	0.566	0.297	0.300	0.294	0.628	0.637	0.261	0.272	0.292	0.767	0.769	0.708	0.715	0.740
Entire Data	0.582	0.583	0.262	0.265	0.283	0.681	0.682	0.255	0.258	0.292	0.782	0.783	0.707	0.711	0.740

Legend: O = Original Data; S = Shuffled Data; P1 = Perturbed (0.00); P2 = Perturbed (0.50); P3 = Perturbed (0.90)

Table 3. Original and Masked Rank Order Correlation (Simulated Data)

5.2. Experimental Assessment Using Census Data

In the previous example, we used a simulated data to highlight the strengths and weaknesses of the two procedures. In this section, we illustrate the applicability of the two procedures to any data set by considering the often used "Census Data". The original Census Data consists of 13 variables and 1080 observations. Of the 13 variables, the variable called PEARNVAL (Total personal earnings) equals PTOTVAL (Personal total income) – POTHVAL (Total other person's income). Hence, rather than using all 3 variables, we only used PEARNVAL in the analysis. Since all 13 of the variables were numerical, in order to illustrate the performance of these procedures for sub-groups, we converted 3 variables (AFLNWGT – Final weight, EMCOMTRB – Employer contribution, and PEARNVAL – Total personal earnings) to categorical variables. For each observation, if the value of each of these variables was less than the average for the entire data set, the value of the corresponding categorical variable was specified as 0 otherwise as 1. This resulted in a total of 8 possible combinations (sub-groups). We used shuffling and perturbation (d = 0.00, 0.50, and 0.90) to mask the data.

5.2.1. Assessment of Disclosure Risk. As before, we assessed identity disclosure risk using the procedure described in Fuller (1993). The results of this assessment are provided in Table 4. As with the simulated data set, it is easy to see that the shuffled data and perturbed data (d = 0.00) provide the lowest risk of identity disclosure, with just one or two records being

identified in each sub-group. The other two perturbed data sets do not fare quite as well. Using the perturbed data with d = 0.50, an intruder could identify a greater proportion of individuals in each sub-group. With the perturbed data with d = 0.90, the level of identity disclosure is extremely high.

Categorical Variable 1	Categorical Variable 2	Categorical Variable 3	Total number of observations	Number of Observations Identified			
				Shuffled	Perturbed (0.00)	Perturbed (0.50)	Perturbed (0.90)
0	0	0	156	4	1	8	83
		1	89	2	1	9	57
	1	0	57	4	1	5	35
		1	156	2	1	7	68
1	0	0	203	4	1	8	90
		1	103	4	1	13	47
	1	0	96	2	1	10	52
		1	220	3	1	10	82

Table 4. Risk of Identity Disclosure (Census Data)

In terms of value disclosure, the width of the confidence interval estimate for the perturbed data with d = 0.00 is exactly 100% of the original width. For the shuffled data, the width of the confidence interval is very close to 100% of the original data. For perturbed data with d = 0.50, the width of the confidence interval is 86.6% $[(1 - 0.5^2)^{0.5}]$ of the width of the original interval. The width of the confidence interval for the perturbed data with d = 0.90 is only 43.6% $[(1 - 0.9^2)^{0.5}]$ of the original width. Thus, the shuffled data and perturbed data with d = 0.00 minimize the risk of value disclosure. The perturbed data with d = 0.50 results in value disclosure which may be considered acceptable. The value disclosure risk resulting from the perturbed data with d = 0.90 is very high and allows the intruder to estimate the values of the confidential variables with much greater accuracy than without access to the data.

5.2.2. Assessment of Information Loss. Figure 3 shows the marginal distribution of the original and perturbed data for the INTVAL variable for the first sub-group (when the value of all the categorical variables is zero). The marginal distribution of the perturbed values are very different from the original values. As with the previous example, there are many negative values while the original variable does not consist of any negative values. While we have limited our discussion to this particular variable for one sub-group, this behavior is observed for practically all variables in all sub-groups. In our opinion, this is a significant problem with the perturbation approach. We also experimented with using alternative distributions for the noise term. The results however are similar to those observed in these cases.

One major advantage of the shuffling approach is that for all variables and all sub-groups, the shuffled data have exactly the same marginal distribution as the original variables. When the data is shuffled, users will be able to analyze individual variables within sub-groups without any information loss. SBLM at least maintains the mean and variance of the variables within the sub-groups. The other procedures (simple additive noise, Kim's method, multiple imputation, micro-aggregation, and swapping) do not typically maintain even mean and variance. Thus, from the perspective of univariate analysis of the masked data for the complete data set and sub-groups, data shuffling provides the best alternative among existing procedures.

As in the previous example, we analyzed both product moment and rank order correlation among the variables. In this case, with as many as 8 variables, there are a total of 21 different correlations to be considered for each of the 8 sub-groups and 4 methods. For the sake of brevity, we did not reproduce the entire set of results. Instead, Table 5 provides the product moment correlation of FICA (Social security deduction) and WSALVAL (Annual total wage and salary). We selected this particular example because of the fact that in one of the sub-groups, the correlation among the two variables is exactly 1.0. The results in Table 5 are similar to those observed for the simulated data. The perturbed data maintains the product moment correlation to be exactly the same for the overall data set and for each sub-group. The product moment correlation of the shuffled data, while very close to the original data, is not exactly the same. As observed earlier however, we do not believe that the product moment correlation is the best method for assessing the relationship among these variables. Hence, we computed the rank order correlation among the variables as an additional measure of information loss. Hence we computed the rank order correlation which is provided in Table 5.

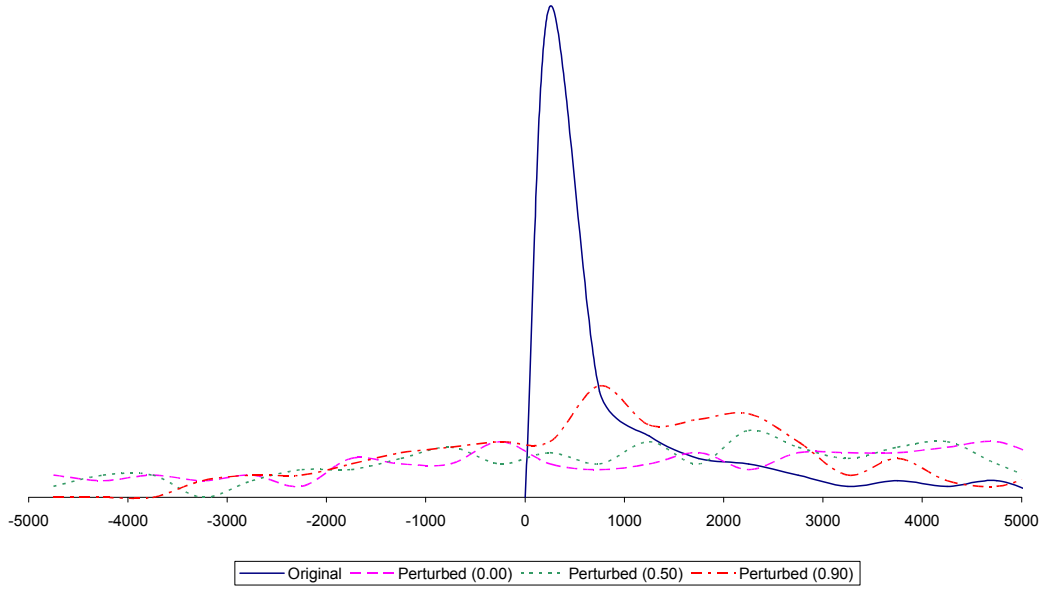


Figure 3. Marginal Distribution of INTVAL Variable for Sub-Domain 1 (Census Data)

Group	Original	Shuffled	Perturbed (0.00)	Perturbed (0.50)	Perturbed (0.90)
1	0.642	0.800	0.642	0.642	0.642
2	1.000	1.000	1.000	1.000	1.000
3	0.817	0.899	0.817	0.817	0.817
4	0.863	0.915	0.863	0.863	0.863
5	0.529	0.734	0.529	0.529	0.529
6	0.988	0.971	0.988	0.988	0.988
7	0.766	0.885	0.766	0.766	0.766
8	0.929	0.943	0.929	0.929	0.929
All Observations	0.910	0.946	0.910	0.910	0.910

Table 5. Product Moment Correlation between FICA and WSALVAL (Census Data)

The results in Table 7 clearly indicate that the shuffled data maintains the rank order correlation among these two variables better than the perturbed data for the overall data set as well as for practically every sub-group. Note that the data shuffling procedure is able to maintain the perfect correlation among the variables in sub-group 2 as does the perturbed data with d = 0.00. Thus, in addition to maintaining linear correlation, data shuffling performs better in maintaining non-linear relationships among variables while the perturbed data do not.

Group	Original	Shuffled	Perturbed (0.00)	Perturbed (0.50)	Perturbed (0.90)
1	0.857	0.858	0.597	0.642	0.770
2	1.000	1.000	1.000	0.955	1.000
3	0.876	0.930	0.821	0.779	0.857
4	0.938	0.930	0.834	0.914	0.913
5	0.807	0.787	0.502	0.472	0.653
6	0.975	0.977	0.986	0.879	0.985
7	0.923	0.945	0.737	0.654	0.846
8	0.965	0.948	0.932	0.922	0.954
All Observations	0.953	0.968	0.930	0.919	0.943

Table 6. Rank Order Correlation between FICA and WSALVAL (Census Data)

4. Conclusions

In summary, data shuffling offers the following advantages: (1) Disclosure risk is minimized for every sub-domain, (2) The marginal distribution of the shuffled data is exactly the same as that of the original data for the complete data set as well as for every sub-domain, and (3) The rank order correlation of the shuffled data is very similar to that of the original data for the complete data set as well as for every sub-domain. The SBLM approaches offers the following advantages: (1) Disclosure risk is minimized for every sub-domain for the perturbed data set when d = 0, but not in the other cases, (2) The mean vector and covariance matrix of the perturbed data is exactly the same as that of the original data for the complete data set as well as for every sub-domain. However, the marginal distribution of the perturbed data is different from that of the original data, and (3) The product moment correlation of the perturbed data is exactly the same as that of the original data for the complete data set as well as for every sub-domain. However, the rank order correlation of the perturbed data is very different from the original rank order correlation.

The selection of the specific approach would depend on the characteristics of the data. If the numerical data does not deviate significantly from normality and/or we are only interested in estimating linear relationships among variables, then the SBLM perturbation approach may be preferred since it offers the advantage that the results of traditional statistical analyses conducted on the masked data would yield *exactly* the same results as those using the original data. However, if the data is known to be non-normal and/or we are interested in estimating non-linear monotonic relationships, then shuffling would be preferred since it maintains the marginal distribution *exactly* and is also capable of maintaining monotonic non-linear relationships among variables. In practice, since data sets that exhibit multivariate normality are not very common, data shuffling would generally be the preferred approach.

References

1. Burridge, J. 2003. Information preserving statistical obfuscation. *Statistics and Computing*, 13 321-327.
2. Fuller, W.A. (1993). Masking procedures for microdata disclosure limitation. *Journal of Official Statistics*. 9, 383-406.
3. Muralidhar K., R. Parsa, R. Sarathy. 1999. A general additive data perturbation method for database security. *Management Science*, 45 1399-1415.
4. Muralidhar, K. and R. Sarathy. 2006. Data shuffling - A new masking approach for numerical data. *Management Science*, 52(5), 658-670.
5. Muralidhar, K. and R. Sarathy. 2007a. 'Easy to implement' is putting the cart before the horse – Effective techniques for masking numerical data. 2007 Federal Committee On Statistical Methodology Research Conference, Arlington VA, November 5-7.
6. Muralidhar, K. and R. Sarathy. 2007b. Generating Sufficiency Based Non-Synthetic Perturbed Data. Working paper.
7. Winkler, W. 2002. Single-ranking micro-aggregation and re-identification. *Research Report Series (Statistics 2002-08)*, US Census Bureau, http://www.census.gov/srd/papers/pdf/rrs2002-08.pdf.
8. Winkler, W. 2006. "Modeling and quality of masked microdata," *Research Report Series (Statistics 2006-01)*, US Census Bureau, http://www.census.gov/srd/papers/pdf/rrs2006-01.pdf.

Evaluating the disclosure risks of reporting quality measures to the public

Jerome P. Reiter*, Anna Oganian**, Alan F. Karr***

* Department of Statistical Science, Duke University, Durham, NC, USA, (jerry@stat.duke.edu)

** National Institute of Statistical Sciences, Research Triangle Park, NC, USA, (aoganian@niss.org)

*** National Institute of Statistical Sciences, Research Triangle Park, NC, USA, (karr@niss.org)

Abstract. To protect confidentiality, statistical agencies typically alter data before releasing them to the public. Ideally, although rarely done, the agency releasing data also provides a way for secondary data analysts to assess the quality of inferences obtained with the released data. Quality measures can help secondary data analysts to disregard inaccurate conclusions resulting from the disclosure limitation procedures, as well as have confidence in accurate conclusions. We propose an interactive computer system that analysts can query for measures of data quality. We focus on potential disclosure risks of providing these quality measures.

1 Introduction

Many national statistical agencies, survey organizations, and researchers disseminate microdata, i.e. data on individual units, to the public. Wide dissemination of data greatly benefits society, enabling broad subsets of the research community to access and analyze the collected data. Often, however, data disseminators cannot release microdata as collected, because doing so could reveal survey respondents' identities or values of sensitive attributes.

Data disseminators therefore limit what they release, typically by altering the collected data. Common strategies include recoding variables, such as releasing ages or geographical variables in aggregated categories; reporting exact values only above or below certain thresholds, for example reporting all incomes above $100,000 as "$100,000 or more"; swapping data values for selected records, e.g., switch the quasi-identifiers for at-risk records with those for other records to discourage users from matching, since matches may be based on incorrect data; and, adding noise to numerical data values to reduce the possibilities of exact matching on key variables

or to distort the values of sensitive variables (Willenborg and de Waal, 2001). Another approach is to replace sensitive values with multiple imputations, often called synthetic data (Little, 1993; Rubin, 1993; Reiter, 2003, 2004; Abowd and Woodcock, 2004).

Generally, increasing the amount of alteration decreases the risks of disclosures; but, it also decreases the accuracy of inferences obtained from the released data. Unfortunately, with most disclosure limitation strategies it is difficult for data users to determine how much their particular estimation has been compromised by the data alteration. This is especially true when data disseminators do not release detailed information about the disclosure limitation strategy.

Secondary data analysts would be greatly helped if statistical agencies provided some way for them to learn about the quality of inferences based on the released data. Benefits of providing data quality measures include (i) analysts can avoid publishing (in the broad sense) results with poor quality, (ii) analysts can feel more confident about results with good quality, and (iii) agencies can claim that they provided information about quality, so that they are not responsible if analysts arrive at inaccurate conclusions. Ideally, the measures are specific to particular inferential quantities rather than broad measures. For example, reporting comparisons of means, variances, and correlations in the observed and masked data does little to help analysts estimating complex models.

In this article, we discuss an approach that enables users of secondary data to assess the quality of inferences made on disclosure-proofed public use data. The idea is to create a *verification server*. This server, located at the statistical agency, would store the original and masked datasets. Analysts, who have only the masked data, would submit queries to the server for measures of data quality for certain estimands. The server would run the analysis on both the original and masked data, and report back to the analyst a measure of data quality that compares the inferences obtained from both sources. The server also serves as a data collection mechanism for the agency, capturing what quantities analysts care most about. Agencies might be able to utilize this information to improve the quality of future data releases.

Verification servers are not the proverbial free lunch. As we shall illustrate, providing measures of data quality also provides ill-intentioned users (henceforth called intruders) with more information for disclosure attacks. The usefulness of this additional information for intruders can be reduced by applying disclosure limitation strategies on the quality measures released by the server.

The remainder of this article is organized as follows. In Section 2, we give a formal representation of the verification server. In Section 3, we describe a data quality measure for the server. In Section 4, we illustrate potential disclosure attacks using the masked data and quality measures. In Section 5, we suggest some approaches for reducing the additional risks from releasing quality measures.

2 Formal description of verification server

The agency wants to release some version M of original microdata, D, to the public in a way the protects confidentiality of survey respondents' identities and sensitive attributes. As is typical of disclosure settings, we assume that the agency does not reveal precise details about the disclosure limitation strategy, except in broad terms like, "noise was added to some variables, while other variables were swapped."

The analyst seeks inferences about some quantity Q, such as a confidence interval for a regression coefficient. Let $Q(M)$ be the estimate of Q obtained from using M, and let $Q(D)$ be the estimate of Q obtained from using D. The analyst can compute only $Q(M)$. In general, whenever Q is based on units whose values were altered for disclosure protection, we expect that $Q(M) \neq Q(D)$. Of course, this is not always the case: $Q(M)$ could equal $Q(D)$ for some Q.

The agency wants to enable secondary data analysts to learn about the differences between $Q(M)$ and $Q(D)$. Given instructions from the analyst, the agency could manually compute $Q(M)$ and $Q(D)$ and describe the differences to the analyst, but this is a labor-intensive process. A preferable approach is to let analysts query a verification server for measures of the quality of their particular $Q(M)$. Let $FM(a, b)$ represent a numerical summary comparing a to b. We call FM a *fidelity measure*, as it represents the degree to which the quantity b is faithful to the quantity a. For queries for acceptable Q, the server reports back a value of the fidelity measure to the user. It never reports D or $Q(D)$.

The verification server has some advantages over model servers, also known as remote access systems, that give analysts $Q(D)$. As we shall illustrate, providing infinitely precise information about analyses of D, whether in the form of $Q(D)$ or $FM(Q(D), Q(M))$, can lead to high disclosure risks. Arguably, it is easier to coarsen $FM(Q(D), Q(M))$ than $Q(D)$. The fidelity measure is qualitative and so can be coarsened while still providing meaningful information about data quality. The $Q(D)$ is precise; altering $Q(D)$ essentially defeats the purpose of providing it. Additionally, model servers must limit the scope of analyses and details of output, since clever queries can reveal individual data values (Gomatam *et al.*, 2005). The verification server with coarsened $FM(Q(D), Q(M))$ arguably is not as susceptible to such tricks.

3 A proposed fidelity measure

Existing utility measures are of two types: (i) comparisons of broad differences between the original and released data, and (ii) comparisons of differences in specific models between the original and released data. Broad difference measures essentially quantify some statistical distance between the distributions of the data on the original and released files, for example a Kullback-Leibler or Hellinger distance. As the distance between the distributions grows, the overall quality of the released

data generally drops. Another approach is based on how well one can discriminate between the original and altered data. For example, Woo *et al.* (2007) stack the original and altered data sets in one file, and estimate probabilities of being "assigned" to the original data conditional on all variables in the data set. When the distributions of probabilities are similar in the original and altered data, the distributions of the variables are similar—this fact comes from the literature on propensity scores for matching in observational studies—so that the altered data have high utility.

Because global measures are only tangentially tied to specific estimands, we consider fidelity measures based on specific models. We use the confidence interval overlap measure of Karr *et al.* (2006). First, the server computes the 95% confidence intervals for the estimand from the masked data, $Q(M) = (L_r, U_r)$, and from the collected data, $Q(D) = (L_o, U_o)$. Then, the server computes the intersection of these two intervals, (L_i, U_i). The fidelity measure is

$$FM(Q(D), Q(M)) = \frac{U_i - L_i}{2(U_o - L_o)} + \frac{U_i - L_i}{2(U_r - L_r)}. \quad (1)$$

When the intervals are nearly identical, corresponding to high utility, the $FM \approx 1$. When the intervals do not overlap, corresponding to low utility, the $FM = 0$. The second term in (1) is included to differentiate between intervals with $\frac{U_i - L_i}{(U_o - L_o)} = 1$ but different lengths. For example, for two masked data intervals that fully contain the collected data interval, the measure (1) favors the shorter interval. Other fidelity measures are possible and could be released as part of the verification server's output.

4 Risks associated with measures

Values of $FM(Q(D), Q(M))$ provide intruders with additional information to attempt disclosure attacks. In this section, we describe examples of how intruders might utilize this information when M is constructed with common disclosure limitation strategies.

In what follows, let X_j represent the vector of quasi-identifiers for unit j, such as age, race, sex, marital status, and geography. We assume that these are known without error by the intruder for selected records in the database. Let Y_j be the vector of sensitive attributes collected in the survey for unit j, such as health, monetary, or other personal variables. We assume that these are not known by the intruder.

4.1 Data swapping

Consider a scenario where a small percentage of X_j are swapped, but Y_j is left alone. Whole vectors of X_j are swapped together. The agency does not reveal which or how many records were swapped. As we shall show, given infinitely precise fidelity measure values the intruder can undo many of the swapping protections.

4.1.1 Determining which records underwent swapping

The intruder can use a trial and error approach to determine which records underwent swapping. The intruder submits a query for the confidence interval for the slope in the regression of one variable in Y on some of the variables in X. When $FM(Q(D), Q(M)) = 1$, with high probability the values of X_j for the records submitted with the query are the original values. (It is possible that $FM(Q(D), Q(M)) = 1$ by random chance, but with sufficiently large sample size in the query this is a small probability.) When $FM(Q(D), Q(M)) \neq 1$, at least one record has undergone swapping. The intruder can isolate the swapped records by submitting many queries based on different subsets of records. This process can be repeated many times—an efficient algorithm can be devised to reduce the number of queries—to uncover all swapped records.

4.1.2 Determining original values for swapped records

Once the set of swapped records is determined, intruders can determine the original values for those records. To illustrate, suppose the intruder seeks the actual value of marital status for a target record with swapped value of marital status equal to "married." First, the intruder selects a set of unswapped records with sufficient numbers in each marital status category. Second, the intruder appends the target record to this unswapped set. Third, using the appended data in the query, the intruder asks for the fidelity measures for the proportion of people in each marital status category. Any marital status category for which $FM(Q(D), Q(M)) = 1$ is eliminated as a candidate for the target's true marital status. Only two categories have $FM(Q(D), Q(M)) \neq 1$. The target's true marital status is the remaining marital status not equal to the swapped value, e.g., the one other than "married" in our example. The intruder can repeat this process for all swapped values in X_j for all j, thus uncovering the true values in D.

As another attack strategy, the intruder could construct all possible datasets by permuting X_j. For each dataset, the intruder estimates a regression involving one component of Y on X, including all records in the regression. With high probability, the dataset for which $FM(Q(D), Q(M)) = 1$ is the actual D. This is a computationally expensive strategy for large datasets, particularly when the intruder does not know any details about the swapping strategy.

4.2 Top-coding

Suppose now that one component of Y_j for some records j is protected by top-coding; that is, large values of this variable are reported only as exceeding a certain threshold t. As we shall show, given infinitely precise fidelity measure values the intruder can undo many of the top-coding protections. In our examples, we assume that the intruder computes $Q(M)$ by setting the top-coded values equal to t.

4.2.1 Rank order top-coded values

The intruder can utilize the fidelity measure to order the top-coded records from smallest to largest values of the unobserved Y. First, the intruder obtains a subset of n records not subject to the top-coding. Second, for some record j subject to top-coding, the intruder appends the record to this subset. Third, the intruder asks for fidelity measures for the mean based on the $n+1$ records in this query. The intruder repeats the second and third steps of this process for each top-coded record, each time using the same n records not subject to top-coding. Finally, the intruder orders the values of $FM(Q(D), Q(M))$ from smallest to largest. This ordering matches the ranking of the unobserved Y, since $Q(M)$ successively worsens in quality as the true values deviate from t.

4.2.2 Determine values of top-coded records

The strategy used in Section 4.2.1 can be modified to learn exact values of top-coded records. As before, the intruder obtains $FM(Q(D), Q(M))$ for a subset of n records with $Y < t$ and one top-coded record j. Then, the intruder guesses a plausible value of the true Y_j for that record, using that guess to make a proposed true dataset, D^*, for those $n + 1$ records. The intruder computes $FM(Q(D^*), Q(M))$. The intruder repeats this process many times using different initial guesses of the true Y_j. The guess that results in $FM(Q(D^*), Q(M)) = FM(Q(D), Q(M))$ and exceeds t is the true value for that record.

4.3 Added noise

With clever transformations, the intruder can estimate the actual values of variables protected by additive noise. Specifically, the intruder submits a query based on n records in which all values but the one for record j are transformed to equal the same number. The intruder obtains $FM(Q(D), Q(M))$ for that query. Then, the intruder guesses a plausible value of the true Y_j, using that guess to make a proposed true dataset, D^*, for those n records. The intruder computes $FM(Q(D^*), Q(M))$. The intruder repeats this process many times using different initial guesses of the true Y_j. The guesses that result in $FM(Q(D^*), Q(M)) = FM(Q(D), Q(M))$ are candidates for the true value. Some of these values may be more plausible than the others, given other information released in the data.

4.4 Synthetic data

The attack strategies described for the other disclosure limitation methods can be applied on partially synthetic data, for which there is a one-to-one correspondence between D and the (multiple sets of) released M. As an example, to learn the original value Y_j for a record with Y synthesized, the analyst appends this record to a set of n records whose values of Y are not synthesized. The intruder guesses a plausible value of the true Y_j, using that guess to make a proposed true dataset, D^*, for those

$n+1$ records. The intruder computes $FM(Q(D^*), Q(M))$. The intruder repeats this process many times using different initial guesses of the true Y_j. The guesses that result in $FM(Q(D^*), Q(M)) = FM(Q(D), Q(M))$ are candidates for the true value. This attack is not possible when all values of the variable are synthesized.

5 Reducing risks for measures

The key factors driving the risks outlined in Section 4 include (i) the ability of intruders to submit queries based on subsets of records and transformations of variables and (ii) the availability of infinitely precise fidelity measures. Thus, to reduce these risks, it makes sense to limit what queries are answered with fidelity measures or to coarsen the reported fidelity measures.

When limiting queries, we should preserve as much as possible the ability of the legitimate user to obtain fidelity measures for complex models. Otherwise, the verification server has limited usefulness. When coarsening fidelity measures, we should provide the analyst with enough information to decide whether or not $Q(M)$ is "good enough" compared to $Q(D)$ for publication. The agency should not make that decision, as different analysts will have different quality requirements.

5.1 Limiting the query space

The verification server need not report fidelity measures for all possible queries. Certain attacks can be made much more difficult with query restrictions that may have minimal impact on legitimate data users. This section describes some query restrictions.

5.1.1 Require well-defined target populations

The verification server might require that records in any queries comprise well-defined sub-populations rather than arbitrary collections. For example, the server could require selection criteria to be functions with limited complexity, such as a maximum of three-way interactions among variables (e.g., all women under age 25 living in a city). The verification server could be programmed to force the analyst to enter the subset selection criteria rather than specify a collection of record identification numbers. Such restrictions go a long way towards defeating attacks based on subsetting records. They are not foolproof; a clever intruder might be able to cobble together desired records from legal subsetting requests.

5.1.2 Disallow certain transformations

Some transformations, such as those described in Section 4.3, are not useful for many legitimate applications but are useful for disclosure attacks. It may be possible to prevent some of these transformations. One approach is to allow a set of standard transformations, such as polynomials and low powers, and disallow all others. This is tricky, however, because the restrictions may rule out legitimate uses.

5.2 Coarsening the fidelity measure

Coarsening fidelity measures creates uncertainty in all of the attacks, because the true fidelity measure is not available for backsolving. This section describes some approaches to coarsen fidelity measures.

5.2.1 Interval reporting

Rather than report precise measures, the verification server can report measures in intervals; for example, between 90% and 100% overlap, between 80% and 90% overlap, etc. Such reporting may be sufficient to enable the user to evaluate quality, yet provide enough uncertainty in the fidelity measures to mask true values. For example, if the server never reports $FM(Q(D), Q(M)) = 1$ precisely, then the attacks on swapped data described in Section 4.1 can be prevented. However, there is no guarantee that interval reporting prevents all attacks. For example, when seeking to order top-coded records, the intruder can determine which record owns the largest value when the interval with the least overlap (e.g., 10% to 20% has smaller overlap than 30% to 40%) is unique.

Intruders can use trial and error attacks like those in Section 4.2.2 – 4.4 to obtain ranges of possible values for unknown Y. For example, to get a range for a particular top-coded value Y_j, the intruder appends that record to a collection of n other records that are not top-coded. The intruder submits queries about the mean of Y based on those $n+1$ records, obtaining the reported interval $FM(Q(D), Q(M))$. The intruder then proceeds as follows.

A1. Set f equal to the lower bound of the reported interval for $Q(M)$.

A2. Find the value Y_j^* such that, when used to make a plausible true dataset D^*, produces $FM(Q(D^*), Q(M))^* = f$, where $FM(Q(D^*), Q(M))^*$ is a single number measurement of the fidelity measure. Store this value of Y_j^*.

A3. Set $f = f + c$, where c is some small number like 0.5.

A4. Repeat Step 2 and 3 until f equals the upper bound of the reported interval for $Q(M)$.

The stored values Y_j^* are the plausible values for the original Y_j. When this distribution does not have sufficient variance, an inferential disclosure occurs.

To illustrate these attacks, we use data from the 1995 U.S. Current Population Survey. The variables include adjusted gross income (AGI) and amount of interest income (INTVAL). The intruder submits a query for the fidelity measure for the intercept in a regression of AGI on INTVAL. The regression is based on a set of 31 records, one of which has AGI top-coded. The intruder seeks to estimate the original value of the top-coded AGI. Figure 1 graphically displays the output from the attack protocol described above. The curved line connects the values of

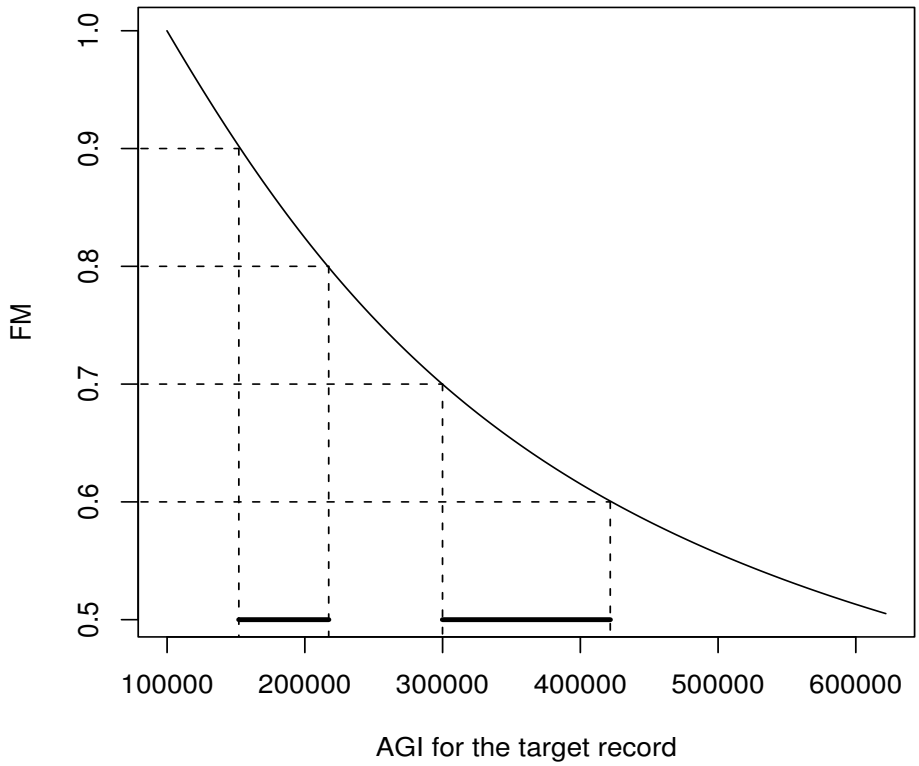


Figure 1: Illustration of intruder's bounding procedure when server reports interval fidelity measures. The solid horizontal lines are ranges of plausible values for original AGIs when the reported interval is (0.6, 0.7) or (0.8, 0.9).

$FM(Q(D^*), Q(M))^*$ for several guesses at the original AGI. If the top-coded record is such that the server reports an interval of (0.8, 0.9), the intruder knows that the original AGI is between \$152,082 and \$217,318. If the top-coded record is such that the server reports (0.6, 0.7), the original AGI is between \$299,950 and \$421,723.

To reduce confidentiality risks when providing interval measures, the server should report as wide an interval as necessary to generate sufficient variability in the prediction distribution, but not so wide as to destroy the usefulness of the interval measure. To do so, the agency can define acceptable bounds of uncertainty for at risk values, such as large incomes. Then, for any given query, the server determines the maximum and minimum values of the fidelity measures that yield these acceptable bounds, and it reports an interval that contains these bounds. This approach still is risky if, for example, the intruder can get close to the true Y_j by backsolving with the midpoint of the reported interval. Additionally, the intruder may be able to sharpen bounds by submitting multiple queries involving the same target value.

A related approach is to add random noise to the fidelity measures before re-

porting, again with the goal of obscuring the true fidelity measure. The amount and distribution of the noise should be hidden from the analyst to provide less information for backsolving attacks. The noise should be the same for all requests for the same $Q(M)$ to eliminate intruders' ability to sharpen estimates by averaging the results of multiple queries for the same analysis. But, the noise must differ by analysis to reduce the chance that the intruder can guess the value of the added noise for some queries, for example if the intruder knows $Q(D)$ and submits $Q(M)$ anyway. These two criteria can be met by tieing the random seed to some function of $Q(D)$ that provides unique seeds for almost all different queries.

The intruder could treat the reported, noisy fidelity measure as a true value, and attempt trial and error attacks. This suggests that the noise distribution should be based on acceptable bounds for risk, as described for the interval measures.

5.2.2 Computing on different datasets

As another approach, the verification server can report fidelity measures based on datasets that differ slightly from D and M. The server does not tell analysts how the datasets differ. As an example, the verification server can do the following.

B1. Delete k randomly sampled records from the data used to compute $Q(D)$. Let r_d and r_n be the row numbers of the deleted records (r_d) and the not deleted records (r_n). Let D_{r_n} and M_{r_n} be the data in D and M for the records in r_n.

B2. Sample k row numbers from r_d, with replacement. Let r_s be the sampled row numbers. Let D_{r_s} and M_{r_s} be the data in D and M for the records in r_s.

B3. Construct $D' = (D_{r_n}, D_{r_s})$ and $M' = (M_{r_n}, M_{r_s})$.

B4. Report $FM(Q(D'), Q(M'))$ to the analyst.

With large k, this creates a combinatorial explosion of possible true values for the records in D. A related approach for model servers was suggested by Steel and Reznek (2006).

This approach is not immune to disclosure risks in verification servers, although the intruder must work hard to guess original values sensibly. First, the intruder proposes a possible value of D, say D^*. Second, the intruder proposes a possible value of D' and M', say D'^* and M'^*, obtained by mimicking steps B1 – B3 on D^* and M. Third, the intruder computes $FM(Q(D'^*), Q(M'^*))$. The intruder repeats this process many times, collecting all D'^* for which $FM(Q(D'^*), Q(M'^*)) = FM(Q(D'), Q(M'))$. The values of the targeted Y_j among the D'^* meeting this criterion are plausible values of the original Y_j. If the distribution does not have sufficient variability, or if some obvious function like the mean/median of the plausible values is too close to the truth, there may be a disclosure risk.

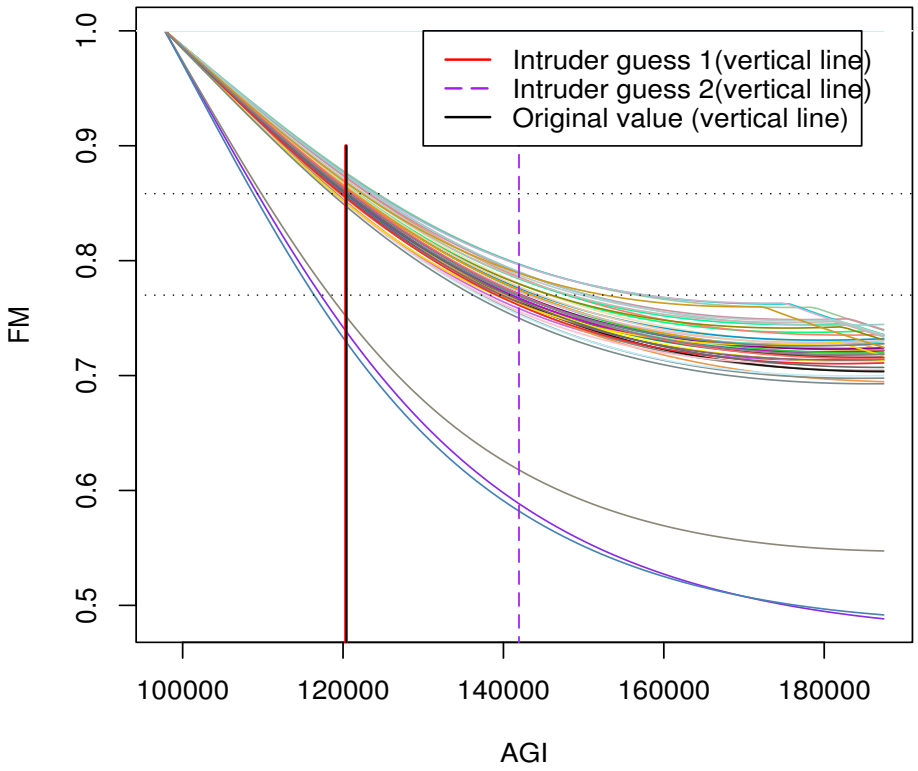


Figure 2: FM curves constructed by the intruder to defeat the *delete-add* strategy of the verification server.

Figure 2 illustrates this attack strategy for one top-coded value of AGI. The intruder's query is based on n records with $Y < t$ and one record subject to top-coding. To draw any one curve, we assume that the intruder knows k and follows steps B1 – B3 only with M to get M'^*. For all top-coded records in M'^*, we replace t with a guess of the original AGI, thus obtaining a D'^*. We then compute $FM(Q(D'^*), Q(M'^*))$. We repeat the last two steps for different guesses, and connect values of $FM(Q(D'^*), Q(M'^*))$ to make the curve. The figure shows 50 such curves, each derived from different records in M'^*. The intruder draws a horizontal line at the reported $FM(Q(D'), Q(M'))$. Its intersections with the curves gives the distribution of plausible values for the original datum. For one realization of D' we obtained $FM(Q(D'), Q(M')) = 0.87$. For this D', the intruder can average the plausible AGI values to get very close to the original value. However, for another realization of D' we obtained $FM(Q(D'), Q(M')) = 0.77$, for which the average of the plausible AGI values is not a close estimate of the original value. Although 0.87 and 0.77 are not that different in terms of data quality, Figure 2 indicates that protection is sensitive to which units are in D'.

6 Concluding remarks

Verification servers could have enormous benefit for statistical agencies and consumers of their data. However, releasing precise quality measures could threaten confidentiality. The examples in this article suggest that both restricting queries and coarsening fidelity measures show promise for providing sufficient protection.

References

Abowd, J. M. and Woodcock, S. D. (2004). Multiply-imputing confidential characteristics and file links in longitudinal linked data. In J. Domingo-Ferrer and V. Torra, eds., *Privacy in Statistical Databases*, 290–297. New York: Springer-Verlag.

Gomatam, S., Karr, A. F., Reiter, J. P., and Sanil, A. P. (2005). Data dissemination and disclosure limitation in a world without microdata: A risk-utility framework for remote access servers. *Statistical Science* **20**, 163–177.

Karr, A. F., Kohnen, C. N., Oganian, A., Reiter, J. P., and Sanil, A. P. (2006). A framework for evaluating the utility of data altered to protect confidentiality. *The American Statistician* **60**, 224–232.

Little, R. J. A. (1993). Statistical analysis of masked data. *Journal of Official Statistics* **9**, 407–426.

Reiter, J. P. (2003). Inference for partially synthetic, public use microdata sets. *Survey Methodology* **29**, 181–189.

Reiter, J. P. (2004). Simultaneous use of multiple imputation for missing data and disclosure limitation. *Survey Methodology* **30**, 235–242.

Rubin, D. B. (1993). Discussion: Statistical disclosure limitation. *Journal of Official Statistics* **9**, 462–468.

Steel, P. and Reznek, A. (2006). Issues in designing a confidentiality preserving model server. In P. D. Munoz and H. Brungger, eds., *Monographs of Official Statistics: Work Session on Statistical Data Confidentiality*, 29–36. Eurostat.

Willenborg, L. and de Waal, T. (2001). *Elements of Statistical Disclosure Control*. New York: Springer-Verlag.

Woo, M. J., Reiter, J. P., Oganian, A., and Karr, A. F. (2007). Global measures of data utility for microdata masked for disclosure limitation. *Journal of Privacy and Confidentiality* forthcoming.

On method-specific record linkage for risk assessment

Jordi Nin, Javier Herranz and Vicenç Torra

IIIA, Artificial Intelligence Research Institute
CSIC, Spanish National Research Council
Campus UAB s/n
08193 Bellaterra (Catalonia, Spain)
{jnin,jherranz,vtorra}@iiia.csic.es

Abstract. Nowadays, the need for privacy motivates the use of methods that permit us to protect a microdata file both minimizing the disclosure risk and preserving the statistical utility.

Nevertheless, research is usually focused on how data utility is preserved, and much less research effort is dedicated to the study of the tools that an intruder might use to compromise the privacy of the data or, in other words, to increase the disclosure risk.

Record linkage is a standard mechanism used to measure the disclosure risk of a microdata protection method. In this paper we present some improvements for the (standard) distance based record linkage. In particular, we test our improvements to evaluate the disclosure risk of rank swapping, which is higher than what was believed up to now. We will also present the results of the application of this approach to microaggregation.

1 Introduction

Nowadays, statistical agencies publish confidential microdata files in the Internet. This data can be accessible for a variety of users, as decision makers, politicians, researchers or general public. However, such publication has to fulfill laws and regulations to preserve the privacy of the respondents.

A good statistical practice is that the released data include a full description of the data as well as the anonymization criteria that has been applied. For instance, all available microdata files in the EUROSTAT web page [12] include a text description explaining all the anonymization criteria applied to the confidential data.

The main goal of data protection methods [1], is to minimize both the *disclosure risk* (DR) and the *information loss* (IL) of the protected released microdata. Disclosure risk measures the capacity of an intruder to obtain some sensitive information about the original dataset from the protected one, and information loss measures the reduction of the statistical utility of the protected microdata with respect to the original one.

Information loss is deeply studied in many works [2, 6, 14], and it is out of the scope of this paper. Although, we will use in our experiments the measures defined in [7] to compare several protection methods.

In this paper, we focus in the way of computing the disclosure risk. Many works [6, 20] use *record linkage* methods [18, 19] for this purpose. Such methods are widely used in the scenario where an intruder has a compete access to the protected data set, whereas he knows some records of the original data set obtained from other data sources (publicly available or not). The aim of the intruder is to use record linkage to link his original records with the corresponding protected records released by the statistical agency. Obviously, the more records are correctly linked, the more disclosure risk has the employed protection method. Some examples of standard record linkage methods are distance based ones and probabilistic ones.

As we have said before, a good practice for statistical agencies is to give a complete description of the anonymization process, therefore, the intruder has a valuable information about how protected data is obtained. For this reason, the common assumption that a real intruder will use a standard record linkage method is quite unrealistic.

Many protection methods like rank swapping [16] or univariate microaggregation [8], protect the data using only *local* information, so that information loss is kept low. For instance, rank swapping has a parameter which limits the swap interval, or univariate microaggregation build the clusters with the k nearest values when the original data is sorted.

In this paper, we present an ad-hoc record linkage method called *Location Record Linkage* (L-RL). Our method exploits such limitations (*i.e.* protection is made locally). Using this knowledge, the intruder can limit the records where the record linkage method is applied, decreasing in this way the probability of finding incorrect links. As a result there is an increase on the number of correct links, and, therefore, an increment in the disclosure risk of such protection methods.

The rest of the paper is organized as follows. In Section 2 we recall the three data protection methods (rank swapping, univariate and multivariate microaggregation) where we have tested the new ad-hoc record linkage technique. Then we explain in Section 3 the basic concepts related to data protection, disclosure risk, information loss and the standard definition of score. a description of the new record linkage method. In Section 4 we describe the Location Record Linkage (L-RL) technique, we define a new score which takes into account L-RL, and we test L-RL with the abovementioned protection methods. Finally, in Section 5, we draw some conclusions and present some future work.

2 A Review of Protection Methods

2.1 Rank Swapping

Rank swapping is a widely used microdata protection method, which was originally described [16] only for ordinal attributes. However, in the comparisons made in [7], rank swapping was ranked among the best microdata protection methods for numerical attributes.

Rank swapping with parameter p and with respect to an attribute $attr_j$ (i.e., the j-th column of the original dataset X) can be defined as follows: first, the records of X are sorted in increasing order of the values x_{ij} of the considered attribute $attr_j$. For simplicity, assume that the records are already sorted, that is $x_{ij} \leq x_{\ell j}$ for all $1 \leq i < \ell \leq n$. Then, each value x_{ij} is swapped with another value $x_{\ell j}$, randomly and uniformly chosen in the set of still unswapped values in the limited range $i < \ell \leq i + p$. Finally, the sorting step is undone. When rank swapping is applied to a dataset, the algorithm explained above is run for each attribute to be protected, in a sequential way.

2.2 Univariate Microaggregation

Another widely used microdata protection method is microaggregation. Given a data set of a attributes, microaggregation builds small clusters of at least k elements and replaces each original value by the centroid of the cluster to which the element belongs.

A few different approaches exist for microaggregation. The simplest one is when each attribute is protected independently. This corresponds to *univariate microaggregation*. At present, a few optimal univariate microaggregation algorithms have been developed. A good example is [13], where the authors implement an optimal univariate microaggregation using graph operations over a graph built from the confidential data.

2.3 MDAV Microaggregation

The MDAV (Maximum Distance to Average Vector) algorithm [15] is an heuristic algorithm for multivariate microaggregation. MDAV is an iterative algorithm; at each step, it computes first the average record of a set of records and then builds a cluster with the farthest k records of this average record. Then, in the same step, another cluster is built with the farthest k records from the centroid of the new built cluster. Then, all the records of these two clusters are removed, and the process is repeated until all original values are protected.

3 Disclosure Risk Scenario

The main objective of a protection method is to anonymize a data set. A data set can be viewed as a file containing a number of records, where each record contains a

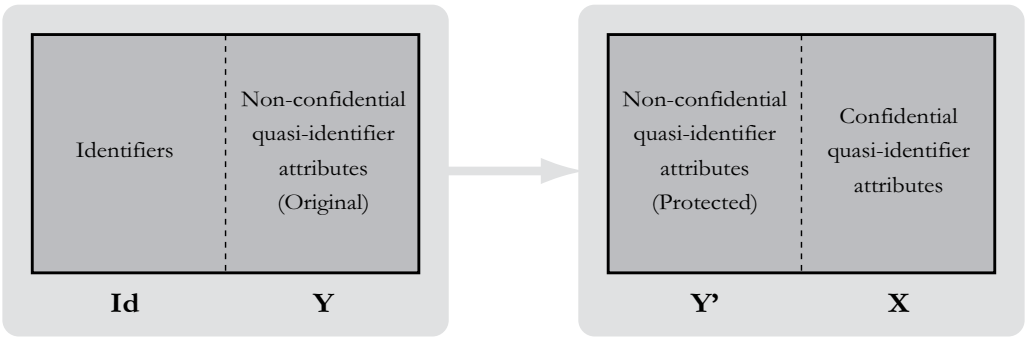


Figure 1: Re-identification scenario.

set of attributes describing an individual. The attributes in the original data set can be classified into two different categories, depending on their capability to identify individuals, as follows:

- **Identifiers.** They can be used to identify the individual unambiguously. A typical example of identifier is the passport number.
- **Quasi-identifiers.** They cannot identify a single individual when used alone. However, when they are combined with others quasi-identifiers attributes, they can uniquely identify an individual. Among the quasi-identifier attributes, we can distinguish between confidential and non-confidential, depending on whether they contain confidential information. An example of non-confidential quasi-identifier attribute is the postal code, while a confidential quasi-identifier is the salary.

When a data set is protected, identifiers are removed or encrypted to prevent an intruder to re-identify individuals easily. Typically, the remaining attributes are released, some of them protected. In this paper, we assume that non-confidential attributes are protected, while confidential attributes are not. This allows third parties to have precise information on confidential data without revealing to whom that confidential data belongs to.

In this scenario, as shown in Figure 1, an intruder might try to re-identify individuals by obtaining the non-confidential quasi-identifiers (Y) together with the identifiers (Id) from other data sources. Then, applying record linkage between the protected attributes (Y') and the same attributes obtained from other data sources (Y), the intruder might be able to re-identify a percentage of the protected individuals together with their confidential data (X). This is what protection methods try to prevent.

In general, the avoidance of all risk is not possible as this usually implies no information. Instead, we have to find a good trade off between information loss and disclosure risk. The score, presented in [6], was defined to measure this trade off in terms of information loss and disclosure risk measures. We define such measures below.

We will use these measures in our experiments as defined in [7].

- **Information Loss (IL).** Let X and X' be matrices representing the original and the protected data set, respectively. Let V and R be the covariance matrix and the correlation matrix of X, respectively; let $\overline{X}$ be the vector of variable averages for X and let S be the diagonal of V. Define V', R',$\overline{X}'$, and S' analogously from X'. The information loss is computed by averaging the mean variations of $X-X'$,$V-V'$,$S-S'$, and the mean absolute error of $R-R'$ and multiplying the resulting average by 100.

- **Disclosure Risk (DR).** The three different methods were presented in [18] to evaluate this risk: (i) *Distance Linkage Disclosure risk* (DLD), which is the average percentage of linked records using distance based record linkage, (ii) *Probabilistic Linkage Disclosure risk* (PLD), which is the average percentage of linked records using probabilistic based record linkage and (iii) *Interval Disclosure risk* (ID) which is the average percentage of original values falling into the intervals around their corresponding masked values. The three values are computed over the number of attributes that the intruder is assumed to know that, in our case, ranges from one to half of the attributes. The DR is a weighted mean that gives half weights to ID and the other half to linkage disclosure risk. That is:

$$DR = 0.5\,ID + 0.5\,[\frac{DLD + PLD}{2}]$$

- **Score:** The final score is defined as the arithmetic sum of IL and DR, therefore

$$score = 0.5\,IL + 0.5\,DR$$

4 Location Record Linkage

As we stated in Section 1, standard record linkage methods underestimate the real disclosure risk in the real world. Here, we consider a new protection method to be used when the intruder knows that only a subset of the protected records are eligible for being linked with the original one. We will call this method *location record linkage* (L-RL for short).

The rationale of our approach is intuitive: protection methods perturbate the original values in a controlled and predictive way to keep information loss as low as possible. For instance, for a given attribute, standard rank swapping swaps one original value with one of the p following values in the sorted table. Then, if the intruder knows all protected attributes (this is our case), he only needs to compare the original record x_i that he wants to link with $2p$ records in the protected data set (note that a protected value can be either the source or the destination in the swap process). The same problem happens with univariate microaggregation, where, if

original data is sorted, clusters are non-overlaped and the values of each cluster are contiguous.

Obviously, if more than one attribute are known, it is possible to repeat the process for each attribute. Formally, if the original record $x_i = (x_{i1}, \ldots, x_{ic})$ has c attributes $attr_1, \ldots, attr_c$, then, the matching protected record x'_ℓ will necessarily satisfy the condition

$$x'_\ell \in \cap_{1 \leq j \leq c} B(x_{ij}),$$

where $B(x_{ij}$ contains all the protected records whose j-th attribute is one of the $2p$ candidates to have been swapped with x_{ij}. That is, the search of the protected record is reduced to an intersection of the sets of possible protected records. Of course, the more attributes are considered, the less records will be in this intersection, and, therefore, the probability of finding the correct record linkage will increase. However, this is not the main concern, because for some combination of the protected attributes, the intersection gives a unique record: the intruder can be sure that this is the protected record which matches with the considered original record. This is so, because this linkage method does not introduce error probabilities. So, the method guarantees to the intruder that the link is correct.

4.1 Experiments

We have considered two different data sets in our experiments. The first one, called Census, has been extracted using the Data Extraction System (DES) from the U. S. Census Bureau [5]. This dataset contains 1080 records consisting of 13 attributes. The second one, called EIA, was extracted from the U.S. Energy Information Authority [11]. It contains 4092 records consisting of 10 attributes.

As we are interested in studying the effects of the L-RL, we have computed two different indicators:

Number of linkages. We study the number of correct links that L-RL is able to find using different sets of attributes. We will assume that the intruder knows all the protected records and he has partial knowledge of the attributes.

Score computation. We compute the standard score and a new variant of it which takes into account the L-RL method. Formally, we define these two score as

- **Score$_1$.** The standard score is computed as presented in Section 3. That is

$$score = 0.5\, IL + 0.125\, DLD + 0.125\, PLD + 0.25\, ID$$

- **Score$_2$.** Our variant of the score is defined by the following expression that includes the standard measures as well as the new L-RL (we use

	rs 2	rs 4	rs 6	rs 8	rs 10	rs 12	rs 14	rs 16	rs 18	rs 20
1	38.4	18.0	16.8	10.8	10.8	6.4	6.8	5.2	4.0	4.0
2	497.0	130.2	54.2	29.2	21.8	15.4	13.0	10.4	7.0	6.0
3	1034.2	761.2	420.6	197.2	99.0	60.2	45.2	32.4	28.8	24.2
4	1071.8	959.6	694.2	378.6	199.4	107.2	71.6	58.0	49.4	39.6
5	1076.8	1042.0	925.2	711.6	463.2	281.4	195.0	165.6	131.6	121.2
6	1079 .0	1063.2	1001.6	879.2	681.0	484.2	413.0	340.4	293.6	287.4
7	1079.2	1064.2	1018.6	913.8	733.0	547.4	475.4	432.4	408.6	339.2
8	1079.2	1077.6	1071.8	1042.6	972.0	861.6	701.6	528.6	472.4	386.4
9	1079.6	1077.6	1071.4	1065.6	1036.6	988.8	888.0	766.6	602.0	466.0
10	1079.6	1078.2	1072.1	1066.6	1039.2	996.6	930.8	824.8	677.4	544.0
11	1079.6	1079.1	1073.4	1069.2	1039.4	1001.0	939.4	846.8	706.2	574.0
12	1079.6	1079.1	1073.4	1069.6	1041.2	1002.0	942.0	853.4	726.4	593.8
13	1079.6	1079.1	1076.7	1070.3	1044.8	1004.4	944.2	871.0	745.6	615.4

Table 1: Number of correctly linked records when L-RL is applied to Census data set, protected with rank swapping. The first column shows the number of known attributes.

LLD to denote Location Linkage Disclosure risk) presented in Section 4. That is,

$$score = 0.5\,IL + 0.25\,(\frac{DLD + PLD + LLD}{3}) + 0.25\,ID$$

4.1.1 Rank Swapping

In Tables 1 and 2 we can observe detailed results about the number of correct links obtained by L-RL using different sets of attributes on data protected using rank swapping. It is easy to observe that the more attributes are known by the intruder, the more records are linked. Note that, for the five less protected data sets from Census, an intruder links more than 70% of the records when only half of the attributes are known. Another interesting result with the Census data set is that the intruder is always able to link more than 50% of the records if he knows all the attributes. Similar results are obtained for the EIA data set. For the three less protected datasets, the intruder is able to link more than 50% of records when all the attributes are known.

Tables 3 (Census data set) and 4 (EIA data set) present $score_1$ and $score_2$, as well as the original values of their components before their aggregation. We can observe that the largest disclosure risk measure, for all cases, is LLD. Therefore, it is clear that the L-RL method increases the risk with respect to standard ones for rank swapping.

4.1.2 Univariate Microaggregation

In Table 5 we can observe detailed results about the number of correct links obtained by L-RL using different sets of attributes for data protected using univariate microaggregation. The results show clearly that the intruder is able to link almost

	rs 2	rs 4	rs 6	rs 8	rs 10	rs 12	rs 14	rs 16	rs 18	rs 20
1	70.3	46.1	43.6	42.4	40.0	37.6	37.6	35.1	37.5	36.3
2	378.4	183.4	145.1	139.8	135.7	135.4	134.6	133.2	133.0	132.3
3	2174.8	338.8	284.1	246.8	236.1	229.1	226.1	224.7	224.3	223.5
4	2827.1	557.0	380.5	327.6	310.2	301.7	298.3	296.0	294.7	294.9
5	3402.9	1441.3	720.9	496.1	423.7	398.7	384.1	373.0	374.0	367.7
6	3582.9	1859.4	856.3	512.4	400.3	415.9	397.0	380.2	378.8	371.3
7	3699.8	2420.6	1325.3	709.6	431.4	423.9	410.1	393.1	424.5	391.6
8	3778.6	2699.0	1631.7	947.3	572.8	448.8	458.3	411.5	445.6	401.3
9	3810.1	2862.2	1808.5	1081.8	654.6	492.2	479.4	420.9	451.9	409.3
10	3831.5	2996.8	1986.3	1221.5	741.5	539.2	507.9	432.6	455.4	411.8

Table 2: Number of correctly linked records when L-RL is applied to EIA dataset, protected with standard rank swapping. The first column shows the number of known attributes.

all the records using only a few attributes. This happens for both data sets. It is clear that univariate microaggregation has a high disclosure risk, greater than the one of rank swapping.

Tables 7 and 8 present, in the same way than in the rank swapping example, $score_1$ and $score_2$, as well as, their components. We can observe that the largest disclosure risk measure is LLD in all cases. Therefore, it is clear that the L-RL method increases the risk with respect to standard ones for univariate microaggregation.

4.1.3 Multivariate Microaggregation

Table 9 presents the number of correct links obtained by L-RL using different sets of attributes for data protected using MDAV multivariate microaggregation. As we can observe in the table, the more groups of attributes are known, the less records are linked. This is due that to the fact that MDAV does not present the same locality problem than univariate microaggregation and rank swapping. In other words, not all the original records are assigned to the cluster represented by the nearest centroid. *I.e.*, some records might be assigned to the second or third nearest cluster. Therefore, L-RL is unsuitable for MDAV. This effect is also illustrated in Tables 10 and 11, where LLD is the lowest disclosure risk value, therefore $score_2$ is lower than $score_1$ and LLD should not be used for the evaluation of the disclosure risk for MDAV.

5 Conclusions

In this paper, we have presented a new type of record linkage designed to exploit the limitations of some protection methods. We have shown that this new method obtains a more accurate disclosure risk evaluation for rank swapping and univariate microaggregation.

	IL	LLD	DLD	PLD	ID	$Score_1$	$Score_2$
rs 2	3.89	77.73	73.52	71.28	93.98	43.54	43.98
rs 4	6.54	66.65	58.40	42.92	83.09	36.71	38.04
rs 6	10.57	54.65	43.76	22.49	72.12	31.60	33.39
rs 8	16.54	41.28	32.13	11.74	62.11	29.28	30.89
rs 10	20.18	29.21	23.64	6.03	53.28	27.12	28.32
rs 12	23.46	19.87	18.96	3.46	47.17	26.33	27.05
rs 14	28.93	16.14	15.63	2.06	43.39	27.52	28.13
rs 16	35.16	13.81	13.59	1.29	40.78	29.64	30.17
rs 18	32.52	12.21	11.50	0.83	38.90	27.53	28.03
rs 20	35.12	10.88	10.87	0.59	37.33	28.33	28.75

Table 3: Score calculation for rank swapping using the Census data set. IL stands for Information Loss, LLD stands for Location Linkage Disclosure, DLD stands for Distance Linkage Disclosure, PLD stands for Probability Linkage Disclosure, ID stands for Interval Disclosure, $Score_1$ is the score computed only using DLD and PLD and $Score_2$ is the score computed taking into account LLD results.

	IL	LLD	DLD	PLD	ID	$Score_1$	$Score_2$
rs 2	4.24	43.27	21.71	16.85	93.10	30.22	32.21
rs 4	9.67	12.54	10.61	4.79	82.09	27.28	27.69
rs 6	14.63	7.69	7.40	2.03	72.21	26.55	26.79
rs 8	18.71	6.12	5.98	1.12	63.90	26.22	26.43
rs 10	22.87	5.60	5.19	0.69	57.09	26.44	26.66
rs 12	26.60	5.39	4.87	0.51	51.64	26.88	27.11
rs 14	29.42	5.28	4.55	0.32	47.49	27.19	27.43
rs 16	32.38	5.19	4.54	0.23	44.19	27.83	28.07
rs 18	34.22	5.20	4.54	0.22	41.42	28.06	28.30
rs 20	36.27	5.15	4.36	0.18	38.97	28.45	28.69

Table 4: Score calculation for rank swapping using the EIA data set.

We have also presented some experiments using MDAV microaggregation that prove that in some sense MDAV is immune to the location problem described in this paper.

As future work, we plan to study the disclosure risk of MDAV and other protection methods using other ad-hoc, specific, record linkage methods.

Acknowledgements

Partial support by the Spanish MEC (projects ARES – CONSOLIDER INGENIO 2010 CSD2007-00004 – and eAEGIS – TSI2007-65406-C03-02) is acknowledged. Jordi Nin wants to thank the Spanish Council for Scientific Research (CSIC) for his I3P grant.

	Census				EIA			
k	2 Vars	3 Vars	4 Vars	5 Vars	2 Vars	3 Vars	4 Vars	5 Vars
10	1032	1079	1080	1080	3430	3923	3947	4035
20	892	1070	1077	1079	2609	3780	3872	3980
30	704	1054	1072	1078	1931	3599	3751	3900
40	531	1021	1065	1076	1388	3347	3621	3806
50	379	989	1054	1069	1012	3074	3427	3703

Table 5: Number of correctly linked records when L-RL is applied to Census data set, protected with univariate microaggregation.

Table 6: Score optimal univariate microaggregation using the Census data set

k	IL	LLD	DLD	PLD	ID	Score_1	Score_2
10	1.15	98.87	86.28	86.31	98.36	46.74	47.79
20	2.85	95.32	83.43	83.47	93.43	45.64	46.63
30	3.71	90.46	80.36	80.21	88.41	44.03	44.88
40	4.71	85.49	77.00	76.57	83.69	42.47	43.20
50	5.66	80.81	73.94	73.56	79.41	41.12	41.71

Table 7: Score calculation for optimal univariate microaggregation using the Census data set.

References

[1] Adam, N. R., Wortmann, J. C.,(1989), Security-control for statistical databases: a comparative study, ACM Computing Surveys, Volume: 21, 515-556.

[2] Bertino, E., Nai, I., Parasiliti, L., (2005), A framework for evaluating privacy preserving data mining algorithms, DMKD, Springer, 11:2 121-154.

[3] Brand, R., Domingo-Ferrer, J., and Mateo-Sanz, J. M., (2002) Reference datasets to test and compare sdc methods for protection of numerical microdata. Manuscript for [4].

[4] CASC: Computational Aspects of Statistical Confidentiality, European Project IST-2000-25069, http://neon.vb.cbs.nl/casc.

[5] Data Extraction System, U.S. Census Bureau, http://www.census.gov/

[6] Domingo-Ferrer, J., Torra, V., (2001), Disclosure control methods and information loss for microdata, Pages 91-110 of [10].

[7] Domingo-Ferrer, J., Torra, V., (2001), A quantitative comparison of disclosure control methods for microdata, Pages 111-133 of [10].

[8] Domingo-Ferrer, J., Mateo-Sanz, J. M. (2002) Practical data-oriented microaggregation for statistical disclosure control, KDE, 14 189-201.

	IL	LLD	DLD	PLD	ID	Score_1	Score_2
10	0.32	93.69	72.53	76.74	99.69	43.74	45.33
20	0.80	87.01	55.94	70.32	99.54	41.07	43.06
30	1.42	80.53	43.54	64.93	99.35	39.10	41.30
40	1.62	74.30	35.52	59.78	98.75	37.41	39.63
50	2.07	68.52	30.41	55.08	95.26	35.53	37.68

Table 8: Score calculation for optimal univariate microaggregation using the EIA data set.

	Census					EIA			
	2 GV	3 GV	4 GV	5 GV	6 GV	2 GV	3 GV	4 GV	5 GV
Mic2-05	120	133	94	76	69	1493	1616	1208	922
Mic2-15	42	178	197	201	200	593	1626	1178	1005
Mic2-25	23	91	130	177	205	349	1306	1551	1324
Mic3-05	73	108	159			414	311		
Mic3-15	28	74	199			170	441		
Mic3-25	12	20	99			82	294		
Mic4-05	45	138				449			
Mic4-15	4	44				141			
Mic4-25	2	12				69			

Table 9: Number of correctly linked records when L-RL is applied to MDAV microaggregation. GV stands for the number of groups of variables known by the intruder. Mic*i*-*k* corresponds to MDAV microaggregation using v variables and clusters of size k.

[9] Domingo-Ferrer, J., Torra, V. (2005) Ordinal, Continuous and Heterogeneous k-Anonymity Through Microaggregation, DMKD, 11 195-212.

[10] Doyle, P., Lane, J., Theeuwes, J., Zayatz, L., eds. (2001), Confidentiality, disclosure, and data access: theory and practical applications for statistical agencies, Elsevier Science.

[11] U.S. Energy Information Authority, http://www.eia.doe.gov/cneaf/electricity /page/eia826.html

[12] Statistical Office of the European Communities (EUROSTAT), http://epp.eurostat.ec.europa.eu

[13] Hansen, S., Mukherjee, S. (2003) A Polynomial Algorithm for Optimal Univariate Microaggregation. KDE, 15:4 1043-1044.

[14] Mateo-Sanz, J.M., Domingo-Ferrer, J., Sebé, F., (2005), Probabilistic information loss measures in confidentiality protection of continuous microdata, DMKD, Springer, 11:2, 181-193.

[15] Hundepool, A., Van de Wetering, A., Ramaswamy, R., Franconi, L., Capobianchi, A., DeWolf, P.-P., Domingo-Ferrer, J., Torra, V., Brand, R., Giessing,

	IL	LLD	DLD	PLD	ID	Score_1	Score_2
Mic2-05	19.30	9.11	69.06	49.22	74.77	43.13	38.96
Mic2-15	37.70	15.15	45.83	26.67	60.94	43.15	41.39
Mic2-25	47.16	11.59	28.56	16.81	51.93	42.23	41.31
Mic3-05	30.66	10.49	37.44	33.58	65.21	40.51	38.43
Mic3-15	42.76	9.29	22.75	19.38	54.79	40.34	39.36
Mic3-25	56.13	4.04	15.86	13.36	51.57	44.61	43.73
Mic4-05	34.67	8.47	31.90	24.35	61.37	39.71	38.07
Mic4-15	45.58	2.22	15.97	12.31	52.43	39.43	38.44
Mic4-25	54.60	0.65	11.20	7.08	45.09	40.86	40.15

Table 10: Score calculation for optimal MDAV multivariate microaggregation using the Census data set.

	IL	LLD	DLD	PLD	ID	Score_1	Score_2
Mic2-05	2.99	32.01	35.01	50.80	93.71	35.65	34.74
Mic2-15	5.49	26.89	20.02	31.49	86.50	30.81	30.90
Mic2-25	6.35	27.68	16.09	26.89	83.88	29.52	30.03
Mic3-05	7.64	8.86	21.47	34.53	85.52	32.20	30.60
Mic3-15	9.99	7.47	11.33	22.67	79.63	29.15	28.36
Mic3-25	11.12	4.59	9.60	18.32	77.63	28.46	27.68
Mic4-05	8.30	10.97	25.71	36.78	87.76	33.90	32.21
Mic4-15	19.16	3.45	12.66	21.31	81.57	34.22	33.09
Mic4-25	20.11	1.69	8.11	14.66	78.28	32.47	31.66

Table 11: Score calculation for optimal MDAV multivariate microaggregation using the EIA data set.

S. (2003) μ-ARGUS version 3.2 Software and User's Manual. Statistics Netherlands, Voorburg NL, feb 2003.

[16] Moore, R., (1996), Controlled data swapping techniques for masking public use microdata sets, U. S. Bureau of the Census (Unpublished manuscript).

[17] Nin, J., Herranz, J., Torra, V., Rethinking Rank Swapping to Decrease Disclosure Risk, Data and Knowledge Engineering, in press. http://dx.doi.org/10.1016/j.datak.2007.07.006

[18] Torra, V., Domingo-Ferrer, J., (2003), Record linkage methods for multidatabase data mining, Information Fusion in Data Mining, Springer, 101-132.

[19] Winkler, W. E., (2003), Data cleaning methods, Proc. SIGKDD 2003.

[20] Yancey, W. E., Winkler, W. E., Creecy, R. H., (2002), Disclosure risk assessment in perturbative microdata protection, Inference Control in Statistical Databases: From Theory to Practice, LNCS, Springer, 2316, 135-152.

Microaggregation Heuristics for p-Sensitive k-Anonymity.

Josep Domingo-Ferrer, Francesc Sebé and Agusti Solanas

Rovira i Virgili University, Dept. of Computer Engineering and Maths, UNESCO Chair in Data Privacy, Av. Països Catalans 26, E-43007 Tarragona, Catalonia. ({josep.domingo,francesc.sebe,agusti.solanas}@urv.cat)

Abstract. p-Sensitive k-anonymity is a sophistication of k-anonymity requiring that there be at least p different values for each confidential attribute within the records sharing a combination of key attributes. Like for k-anonymity, the computational approach originally proposed to achieve this property is based on generalization and suppression; this has several data utility problems, such as turning numerical key attributes into categorical, injecting new categories, injecting missing data, etc. We present and evaluate two heuristics for p-sensitive k-anonymity which, being based on microaggregation, overcome most of such drawbacks, while offering a smooth information loss increase as p and k grow.

1 Introduction

What is meant by database privacy largely depends on the context where this concept is being used. In official statistics, it normally refers to the privacy of the respondents to which the database records correspond (*respondent privacy*). In cooperative market analysis, it is understood as keeping private the databases owned by the various collaborating corporations (*data owner privacy*). In healthcare, both respondent and owner privacy are implicitly required: patients must keep their privacy and the medical records should not be transferred from a hospital to, say, an insurance company. In the context of dynamically queryable databases and, in particular, Internet search engines, the most rapidly growing concern is *user privacy*, that is, the privacy of the queries submitted by users (especially after scandals like the August 2006 disclosure of 658000 queries by the AOL search engine). Thus, what makes the difference is whose privacy is being sought.

Statistical disclosure control (SDC, see Dalenius, 1974; Willenborg and De Waal, 2001; Hundepool *et al.*, 2006) was born in the statistical community as a discipline to achieve respondent privacy. Privacy-preserving data mining (PPDM) appeared simultaneously in the database community (Agrawal and Srikant, 2000) and the cryptographic community (Lindell and Pinkas, 2000) with the aim of offering owner privacy: several database owners wish to compute queries across their databases in such a way that only the results of the queries are revealed to each other, not the contents of each other's databases. Finally, private information retrieval (PIR; Chor *et al.*, 1995) originated in the cryptographic community as an attempt to guarantee the privacy of user queries to databases.

Thus, the technologies to deal with the above three privacy dimensions (respondent, owner and user) have evolved in a fairly independent way within research communities with surprisingly little interaction. Fortunately, it turns out that some developments are useful for more than one privacy dimension, even if all three dimensions are independent (Domingo-Ferrer, 2007). Such is the case for k-anonymity and p-sensitive k-anonymity, which are useful properties both for respondent and owner privacy. Furthermore, in combination with private information retrieval, those two properties make all three privacy dimensions compatible. Thus, presenting efficient computational methods to meet those two properties is an especially relevant objective, which will be treated in this paper. Section 2 discusses k-anonymity for respondent and owner privacy, and recalls how to achieve it using microaggregation. Section 3 discusses p-sensitive k-anonymity and presents two heuristics to achieve this property via microaggregation. Section 4 contains an empirical performance evaluation of both heuristics. Conclusions are drawn in Section 5.

2 k-Anonymity for respondent and owner privacy

k-Anonymity is an interesting approach to face the conflict between information loss and disclosure risk, suggested by Samarati and Sweeney (1998). To recall the definition of k-anonymity, we need to enumerate the various (non-disjoint) types of attributes that can appear in a microdata set $\mathbf{X}$:

- *Identifiers*. These are attributes that *unambiguously* identify the respondent. Examples are passport number, social security number, full name, etc. Since our objective is to prevent confidential information from being linked to specific respondents, we will assume in what follows that, in a pre-processing step, identifiers in $\mathbf{X}$ have been removed/encrypted.

- *Key attributes*. Borrowing the definition from Dalenius (1986), key attributes are those in $\mathbf{X}$ that, in combination, can be linked with external information to re-identify (some of) the respondents to whom (some of) the records in $\mathbf{X}$ refer. Examples are job, address, age, gender, etc. Unlike identifiers, key attributes cannot be removed from $\mathbf{X}$, because any attribute is potentially a key attribute.

- *Confidential outcome attributes*. These are attributes which contain sensitive information on the respondent. Examples are salary, religion, political affiliation, health condition, etc.

We now have:

Definition. *A protected data set is said to satisfy k-anonymity for $k > 1$ if, for each combination of key attributes, at least k records exist in the data set sharing that combination.*

If, for a given k, k-anonymity is assumed to be enough protection for respondents, one can concentrate on minimizing information loss with the only constraint that k-anonymity should be satisfied. This is a clean way of solving the tension between data protection and data utility. The original computational approach in Samarati

and Sweeney (1998) to achieve k-anonymity relies on suppressions and generalizations, so that minimizing information loss translates to reducing the number and/or the magnitude of suppressions.

The drawbacks of partially suppressed and coarsened data for analysis were highlighted in Domingo-Ferrer and Torra (2005):

1. Satisfying k-anonymity with minimum data modification using generalization (recoding) and local suppression was shown to be NP-hard in Meyerson and Williams (2004) and Aggarwal *et al.* (2004);

2. Using global recoding for generalization causes too much information loss, and using local recoding complicates data analysis by causing old and new categories to co-exist in the recoded file;

3. There is no standard way of using local suppression (at the tuple level, at the attribute level, with blanking, with replacement by neutral values, etc.);

4. Analyzing partially suppressed data usually requires specific software (imputation software, censored data analysis, etc.);

5. Last but not least, when numerical attributes are generalized, they become non-numerical.

Joint multivariate microaggregation (in the way of Domingo-Ferrer and Mateo-Sanz, 2002) of all key attributes with minimum group size k was proposed in Domingo-Ferrer and Torra (2002) as an alternative to achieve k-anonymity; besides being simpler, this alternative has the advantage of yielding complete data without any coarsening (nor categorization in the case of numerical data). As a reminder, microaggregation seeks to split a data set into groups of records such that each group contains at least k records and groups are as homogeneous as possible; then records within a group are replaced with the average of all records in the group. Clearly, the higher the homogeneity of records in a group, the lower the information loss caused by replacement of those records by their average. In the case of the k-anonymity application, microaggregation is performed on the projection of records on key attributes, rather than on the entire records. If the microaggregated attributes are numerical, group homogeneity can be measured by the within-groups sum of squares SSE: the smaller SSE, the more homogeneous are the groups.

In Aggarwal and Yu (2004), masking through condensation (actually a special case of multivariate microaggregation) is proposed to achieve k-anonymity in the context of privacy-preserving data mining, and thus with the aim of owner privacy.

3 p-Sensitive k-anonymity via microaggregation

k-Anonymity is able to prevent identity disclosure, *i.e.* a record in the k-anonymized data set cannot be mapped back to the corresponding record in the original data set. However, in general, it may fail to protect against attribute disclosure. In Truta and Vinay (2006), an evolution of k-anonymity called p-sensitive k-anonymity was

presented. Its idea is that there be at least p different values for each confidential attribute within the records sharing a combination of key attributes. The following example illustrates a case where p-sensitive k-anonymity is useful because k-anonymity alone does not offer enough protection.

Example. *Imagine that an individual's health record is k-anonymized into a group of k patients with k-anonymized key attributes values* Age = "30", Height = "180 cm" *and* Weight = "80 kg". *Now, if all k patients share the confidential attribute value* Disease = "AIDS", *k-anonymization is useless, because an intruder who uses the key attributes* (Age, Height, Weight) *can link an external identified record*

(Name="John Smith", Age="31", Height="179", Weight="81")

with the above group of k patients and infer that John Smith suffers from AIDS (attribute disclosure).

Based on the above remarks, the following definition can be given

Definition. *A data set is said to satisfy p-sensitive k-anonymity for $k > 1$ and $p \leq k$ if it satisfies k-anonymity and, for each group of tuples with the same combination of key attribute values that exists in the data set, the number of distinct values for each confidential attribute is at least p within the same group.*

The computational approach proposed in Truta and Vinay (2006) and Truta *et al.* (2007) to achieve p-sensitive k-anonymity is an extension of the generalization/suppression procedure proposed in the original k-anonymity papers. Therefore it shares the same shortcomings pointed out in Domingo-Ferrer and Torra (2005) and listed above.

Like we did for k-anonymity in Domingo-Ferrer and Torra (2005), in Domingo-Ferrer (2006) we showed a way to achieve p-sensitive k-anonymity via microaggregation.

The goal is to obtain p-sensitive k-anonymous data sets without coarsened nor partially suppressed data. This makes their analysis and explotation easier, with the additional advantage that numerical continuous attributes are not categorized.

Note. *In addition to p-sensitive k-anonymity, a number of other sophistications of k-anonymity for protecting against attribute disclosure have recently been proposed, such as l-diversity (Machanavajjhala, 2006), (α, k)-anonymity (Wong* et al., *2006), t-closeness (Li* et al., *2007) and m-confidentiality (Wong* et al., *2007). All of them rely on generalizations, so the microaggregation approach proposed in this paper would be a novelty in all of them. For the sake of concreteness, we will focus here on p-sensitive k-microaggregation.*

We next present two different heuristics for microaggregation-based p-sensitive k-anonymity. The first one starts by achieving k-anonymity and then achieves p-sensitivity. The second one first achieves p-sensitivity and then k-anonymity.

3.1 k-Anonymity first

The heuristic in this section was described in Domingo-Ferrer (2006) without performance analysis and is as follows:

Algorithm 1 (k-Anonymity first)

1. *If $p > k$, signal an error ("p-sensitive k-anonymity infeasible") and exit the Algorithm.*
2. *If the number of distinct values for any confidential attribute in $\mathbf{X}$ is less than p over the entire dataset, signal an error ("p-sensitive k-anonymity infeasible") and exit the Algorithm.*
3. *k-Anonymize $\mathbf{X}$ using the MDAV microaggregation algorithm described in Domingo-Ferrer and Torra (2005). Let $\mathbf{X}'$ be the microaggregated, k-anonymized dataset.*
4. *Let $\hat{k} := k$.*
5. *While p-sensitive k-anonymity does not hold for $\mathbf{X}'$ do:*
   (a) *Let $\hat{k} := \hat{k} + 1$.*
   (b) *$\hat{k}$-Anonymize $\mathbf{X}$ using microaggregation. Let $\mathbf{X}'$ be the $\hat{k}$-anonymized dataset.*

The above algorithm is based on the following facts:

- A $k+1$-anonymous dataset is also k-anonymous;
- By increasing the minimum group size, the number of distinct values for confidential attributes hopefully increases (in the extreme case, if there is a single group as large as the entire dataset, all distinct values for all attributes are in the group).

3.2 p-Sensitivity first

The heuristic below first achieves p-sensitivity and then completes groups in order for them to include k or more records.

Algorithm 2 (p-Sensitivity first)

1. *Let $x_1, x_2, \ldots, x_n$ be the records in the original data set $\mathbf{X}$; let L be the set of confidential attributes. Let Q be the set of key attributes and let $x_j(Q)$ be the projection of record x_j on its key attributes.*
2. *Let P be an initially empty partition.*
3. *While there are at least k records in $\mathbf{X}$ and such records contain at least p different values for each attribute in L do:*
   (a) *Compute the average record $\overline{x}(Q)$ of the projections $x_1(Q), \ldots, x_n(Q)$.*
   (b) *Consider record $x_r \in \mathbf{X}$ so that Euclidean distance between $x_r(Q)$ and $\overline{x}(Q)$ is maximum.*

(c) *Create a new group C that initially contains record x_r.*

(d) *While confidential attributes of the records in C do not satisfy p-sensitivity:*

i. *Take $x_s \in \mathbf{X}$ so that $x_s(Q)$ is the nearest record to $x_r(Q)$ such that $x_s(L)$ contributes to the compliance of p-sensitivity by $C(L)$ (records in C projected on the confidential attributes in L).*

ii. *Add x_s to C and remove it from $\mathbf{X}$.*

(e) *While C does not contain at least k records:*

i. *Take $x_s \in \mathbf{X}$ so that the distance between $x_s(Q)$ and $x_r(Q)$ is minimum.*

ii. *Add x_s to C and remove it from $\mathbf{X}$.*

(f) *Add C to P.*

4. *For each record x remaining in $\mathbf{X}$:*

- *Add x to the group $C \in P$ satisfying that the distance from $x(Q)$ to $Centroid(C)(Q)$ (the centroid of the projections on Q of records in C) is minimum.*

5. *For $i = 1$ to n:*

- *Let x'_i be x_i with $x_i(Q)$ replaced by $Centroid(C)(Q)$, where C is the group in P to which x_i has been assigned.*

6. *The microaggregated, p-sensitive, k-anonymous data set X' is formed by records $x'_1, \cdots x'_n$.*

4 Empirical results

The test data set was generated from the "Census" data file, which was used in the European CASC project (Brand *et al.*, 2002) and in several papers in the microaggregation literature (Domingo-Ferrer *et al.*, 2001; Dandekar *et al.*, 2002; Yancey *et al.*, 2002; Laszlo and Mukherjee, 2005; Domingo-Ferrer and Torra, 2005; Domingo-Ferrer *et al.*, 2007). This data set contains 1080 records with 13 numerical attributes.

The following procedure was used to generate the test data set:

1. The first six attributes of "Census" were taken as key attributes. These attributes were standardized by subtracting their mean and dividing by their standard deviation.

2. Three categorical confidential attributes with 15 categories each were generated from the next three continuous attributes from "Census", respectively. To generate a categorical attribute with 15 categories from a numerical attribute, the range comprised between the minimum value and the maximum value of the attribute is divided into 15 intervals of the same length. Continuous values falling into the first interval are recoded into the first category, those falling into the second interval are recoded into the second category, and so on.

Algorithms 1 and 2 were run for different values of k and p. Results are presented in Tables 1 and 2, respectively.

Table 1: $100 \times SSE/SST$ ratio of the p-sensitive k-anonymous microaggregations yielded by Algorithm 1

	p				
k	**1**	**3**	**5**	**7**	**10**
3	3.69	44.96			
5	6.20	44.96	58.85		
7	7.93	44.96	58.85	71.67	
10	9.71	44.96	58.85	71.67	100

Table 2: $100 \times SSE/SST$ ratio of the p-sensitive k-anonymous microaggregations yielded by Algorithm 2

	p				
k	**1**	**3**	**5**	**7**	**10**
3	3.69	23.13			
5	6.20	23.28	47.15		
7	7.93	22.31	47.15	57.63	
10	9.71	23.13	47.15	57.63	100

Some comments on Tables 1 and 2 follow:

- For $p = 1$, Algorithm 1 is equivalent to the MDAV microaggregation algorithm (Domingo-Ferrer and Torra, 2005). For $p = 1$, Algorithm 2 is equivalent to Centroid-based fixed-size microaggregation (CBFS, Laszlo and Mukherjee, 2005), an algorithm very similar to MDAV. This similarity explains that for $p = 1$ and all values of k reported in Tables 1 and 2, the same SSE/SST ratio is obtained.
- However, for higher values of p, Algorithm 1 behaves clearly worse than Algorithm 2, with higher SSE/SST ratios. The explanation is that the average size of the computed groups is greater for Algorithm 1 than for Algorithm 2: indeed, not caring about p-sensitivity from the start results in a penalty in terms of group size and, consequently, of SSE/SST ratio.
- The reason that Table 1 does not seem to reflect a dependency on k for $p = 3, 5, 7, 10$ is that the actual minimum group size $\hat{k}$ computed by Algorithm 1 turns out to be always greater than or equal to $k = 10$.
- In Table 2, such a lack of dependency on k seems to occur for $p = 5, 7, 10$ also for Algorithm 2: again the explanation is that, for those values of p, the minimum group size if greater than or equal to 10.

- For $p = 10$, both algorithms yield an SSE/SST ratio equal 100. This means that they yield a partition with a single group. This is not really surprising, because the categorical attributes in our data set have only 15 different categories, some of which are rare (*e.g.* those intervals corresponding to the tails of the attribute range).

5 Concluding discussion

p-Sensitive k-anonymity is a sophistication of k-anonymity, whose idea is to avoid that all records sharing a combination of key attributes in a k-anonymous data set also share the values for one or more confidential attributes. The computational approach originally proposed to achieve this new property is based on generalization and suppression and has a number of data utility problems enumerated above.

We have proposed and evaluated two heuristics for p-sensitive k-anonymity which, being based on microaggregation, preserve the numerical nature of key attributes, do not introduce missing data and gracefully degrade data utility as p grows. The first heuristic starts by ensuring k-anonymity and then seeks to achieve p-sensitivity. The second heuristic proceeds the other way round: first p-sensitivity is satisfied and then k-anonymity. This second strategy seems to be clearly better in terms of within-groups homogeneity and, consequently, of data utility.

Disclaimer and acknowledgments

The authors are solely responsible for the views expressed in this paper, which do not necessarily reflect the position of UNESCO nor commit that organization. This work was partly supported by the Spanish Ministry of Education through projects TSI2007-65406-C03-01 "E-AEGIS" and CONSOLIDER CSD2007-00004 "ARES", and by the Government of Catalonia under grant 2005 SGR 00446.

References

Aggarwal, C. C. and Yu, P. S. (2004) "A condensation approach to privacy preserving data mining". In E. Bertino, S. Christodoulakis, D. Plexousakis, V. Christophides, M. Koubarakis, K. Böhm, and E. Ferrari, editors, *Advances in Database Technology - EDBT 2004*, volume 2992 of *Lecture Notes in Computer Science*, pages 183–199, Berlin Heidelberg.

Aggarwal, G., Feder, T., Kenthapadi, K., Motwani, R., Panigrahy, R., Thomas, D.,and Zhu, A. (2004) "k-Anonymity: Algorithms and hardness". Technical report, Stanford University.

Agrawal, R., and Srikant, R. (2000) "Privacy preserving data mining". In *Proceedings of the ACM SIGMOD*, pages 439–450. ACM.

Brand, R., Domingo-Ferrer, J., and Mateo-Sanz, J. M. (2002) "Reference data sets to test and compare SDC methods for protection of numerical microdata". European Project IST-2000-25069 CASC, http://neon.vb.cbs.nl/casc.

Chor, B., Goldreich, O., Kushilevitz, E., and Sudan, M. (1995) "Private information retrieval". In *IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 41–50.

Dalenius, T. (1974) "The invasion of privacy problem and statistics production. an overview". *Statistik Tidskrift*, 12: 213–225.

Dalenius, T. (1986) "Finding a needle in a haystack - or identifying anonymous census records". *Journal of Official Statistics*, 2(3):329–336.

Dandekar, R., Domingo-Ferrer, J., and Sebé, F. (2002) "LHS-based hybrid microdata vs rank swapping and microaggregation for numeric microdata protection". In J. Domingo-Ferrer, editor, *Inference Control in Statistical Databases*, volume 2316 of *Lecture Notes in Computer Science*, pages 153–162, Berlin Heidelberg, Springer.

Domingo-Ferrer, J. (2006) "Microaggregation for database and location privacy". In O. Etzion, T. Kuflik, and A. Motro, editors, *Next Generation Information Technologies and Systems-NGITS 2006*, volume 4032 of *Lecture Notes in Computer Science*, pages 106–116, Berlin Heidelberg.

Domingo-Ferrer, J. (2007) "A three-dimensional conceptual framework for database privacy". In *Secure Data Management-4th VLDB Workshop SDM'2007*, volume 4721 of *Lecture Notes in Computer Science*, pages 193–202, Berlin Heidelberg, 2007.

Domingo-Ferrer, J., and Mateo-Sanz, J. M. (2002) "Practical data-oriented microaggregation for statistical disclosure control". *IEEE Transactions on Knowledge and Data Engineering*, 14(1):189–201, 2002.

Domingo-Ferrer, J., Mateo-Sanz, J. M., and Torra, V. (2001) "Comparing SDC methods for microdata on the basis of information loss and disclosure risk". In *Pre-proceedings of ETK-NTTS'2001 (vol. 2)*, pages 807–826, Luxemburg: Eurostat.

Domingo-Ferrer, J., Sebé, F., and Solanas, A. (2007) "A polynomial-time approximation to optimal multivariate microaggregation", *Computers & Mathematics with Applications*, (to appear).

Domingo-Ferrer, J., and Torra, V. (2005) "Ordinal, continuous and heterogeneous k-anonymity through microaggregation". *Data Mining and Knowledge Discovery*, 11(2):195–212.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Lenz, R., Longhurst, J., Schulte-Nordholt, E., Seri, G., and DeWolf, P.-P. (2006) *Handbook on Statistical Disclosure Control (version 1.0)*. Eurostat (CENEX SDC Project Deliverable).

Laszlo, M., and Mukherjee, S. (2005) "Minimum spanning tree partitioning algorithm for microaggregation". *IEEE Transactions on Knowledge and Data Engineering*, 17(7):902–911.

Li, N., Li, T., and Venkatasubramanian, S. (2007) "T-closeness: privacy beyond k-anonymity and l-diversity". In *Proceedings of the IEEE ICDE 2007*.

Lindell, Y., and Pinkas, B. (2000) "Privacy preserving data mining". In *Advances in Cryptology - CRYPTO'00*, volume 1880 of *Lecture Notes in Computer Science*, pages 36–53, Berlin Heidelberg.

Machanavajjhala, A., Gehrke, J., Kiefer, D., and Venkitasubramanian, M. (2006) "L-diversity: privacy beyond k-anonymity". In *Proceedings of the IEEE ICDE 2006*, 2006.

Meyerson, A., and Williams, R. (2004) "On the complexity of optimal k-anonymity". In *Proc. of the ACM Symposium on Principles of Database Systems-PODS'2004*, pages 223–228, Paris, France, ACM.

Samarati, P., and Sweeney, L. (1998) "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression". Technical report, SRI International.

Truta, T. M., Campan, A., and Meyer, P. (2007) "Generating microdata with p-sensitive k-anonymity". In *Secure Data Management-4th VLDB Workshop SDM'2007*, volume 4721 of *Lecture Notes in Computer Science*, pages 124–141, Berlin Heidelberg.

Truta, T. M. and Vinay, B. (2006) "Privacy protection: p-sensitive k-anonymity property". In *2nd International Workshop on Privacy Data Management PDM 2006*, page 94, Berlin Heidelberg. IEEE Computer Society.

Willenborg, L., and DeWaal, T. (2001) *Elements of Statistical Disclosure Control*. Springer-Verlag, New York.

Wong, R. C. W., Fu, A. W. C., Wang, K. and Pei, J. (2007) "Minimality attack in privacy preserving data publishing". In *Proceedings of the VLDB 2007*, Vienna.

Wong, R. C. W., Li, J., Fu, A. W. C., and Wang, K. (2006) "(α,k)-Anonymity: an enhanced k-anonymity model for privacy-preserving data publishing". In *Proceedings of the ACM KDD*, pages 754–759, New York.

Yancey, W. E., Winkler, W. E., and Creecy, R. H. (2002) "Disclosure risk assessment in perturbative microdata protection". In J. Domingo-Ferrer, editor, *Inference Control in Statistical Databases*, volume 2316 of *Lecture Notes in Computer Science*, pages 135–152, Berlin Heidelberg, Springer.

The use of protected micro data in tabulation: A case of SDC-methods, microaggregation and PRAM

Janika Konnu*

* Statistics Finland, P.O. Box 5V, FI-00022 Statistics Finland, Finland, janika.konnu@stat.fi

Abstract: The development of Statistical Disclosure Control (SDC) methods has been fast and focused on micro data. However, the main problem in statistical agencies is how to produce safe tabular data. It is commonly known that the most suitable and efficient SDC method can be used only if the attributes of the micro data and its variables are properly taken into account. We chose to take deeper look at two of the methods, namely microaggregation and the Post RAndomization Method (PRAM). These methods were used to protect personal data. In our study the main objective was to analyse micro data protection from the data user's point of view. We compiled several tables using both original and protected data in the process. One of the interests of the study was to see whether there are significant differences in some basic tables of frequencies - and when there are, which parameter values generate acceptable differences.

1 Introduction

Statistical agencies have to take good care of the data they have collected. In the case of register-based data, it is easy to have access to all kinds of data the agency needs, but when it comes to surveys, the respondents' willingness determines whether they give the information or not. If respondents can trust that their data will be used appropriately by statistical agencies and there will be no risk of disclosure, they are more willing to provide accurate information. If respondents suspect their information is at risk of disclosure, it is only natural for them to refuse to answer or to provide inaccurate information.

In section 2 we introduce the Statistical Disclosure Control (SDC) methods and the data we used in the study. The results are described in section 3 and the conclusion is presented in section 4.

2 Methods and data

We used two methods for micro data protection. The first one was microaggregation, which is normally applied to continuous variables, and the second was the Post RAndomization Method (PRAM). In the study we applied these SDC methods with the μ-Argus software. Statistics Netherlands originally developed the software, but

upgrading it was one of the main tasks in two European Union projects: Computational Aspects of Statistical Confidentiality (CASC) and a CENtre of EXcellence for Statistical Disclosure Control (CENEX-SDC). As a result of work done there, the software now includes more methods and can be downloaded as freeware from the projects' web pages.

When data are protected with μ-Argus, it is important that metadata are specified very carefully. The software uses these specifications to estimate disclosure risk. The actual protection is applied on the basis of these estimates. The software has the property of generating safe data when final suppressions are allowed at the end of protection. In our research we wanted to analyse methods and consequently no suppressions were allowed.

2.1 Microaggregation

The SDC method called microaggregation is based on counting averages and releasing those instead of the original values in a record. This method has been proposed over a decade ago and it is used in many European countries and by Eurostat. However, its usability has been under discussion. Microaggregation is a method originally developed for continuos data (Group Crises, 2004), but as we demonstrate below, it can be modified to be used for categorical data.

Microaggregation is one of the SDC methods available in the μ-Argus software. In the software fixed size groups are formed using a MDAV (Maximum Distance to Average Vector) algorithm. This means that the average values for all the variables in the data are counted and records are grouped using the difference from these averages. When a MDAV algorithm is used, all the similar records included in the data form groups. In this way it is possible to try to minimise the information loss that releasing averages instead of actual values can entail.

There have been attempts to modify the microaggregation method for categorical data. Through trial and error we noticed that the software available already allowed for it, when we changed the codes of the categories so that they look like continuous values.

2.2 The Post RAndomization Method

The Post RAndomization Method (PRAM) is an SDC method based on misclassification, and it can be applied only to categorical data (de Wolf & van Gelder, 2004). In PRAM the values of the variables are changed on the basis of a chosen probability distribution. It has been reported that data protected by PRAM have to be analysed taking into account the PRAM matrix used in protection. If the matrix is forgotten, the results obtained may deviate extensively from ones that would be obtained by using the original data.

We were interested in analysing the change that actually occurs when PRAM is applied. We wanted to see if it is possible to use PRAM in cases where the researcher gets strongly protected data for the purpose of planning the analysis. The final results are then derived from the original data by the staff of the NSI. At Statistics Finland there is at least one example of such a procedure. The Finnish Longitudinal Employer-Employee Data contain such sensitive information on both companies and their employees that researchers can only have access to strongly protected parts of it.

2.3 Statistical methods

The purpose of the study was to get an idea on how microaggregation or PRAM change the properties of the data as they are used to protect it. Our interest was on the tables we can form with the protected data. How much do the cell values change? And if the values change, how much the interpretation of the table changes?

2.4 Data

In our study of SDC methods we wanted to test the proposed methods on some typical data that researchers want to have for their research. Detailed data on enterprises can only be studied on the premises of Statistics Finland. In this case the restrictions and SDC methods have been thought out quite carefully. At Statistics Finland most of the problems arise when a researcher wants to have very detailed personal data and use it outside of our premises. There is a need for general guidelines on how to protect these kinds of data, and how to determine when data are considered so sensitive that researchers can have access to them only in our research laboratory.

The data we used in our study contain information on teachers in Finland. As the main focus of the study was on SDC methods, only a part of this large data set was used. We decided that data containing high school teachers N=7798 was suitable for our purposes. We chose to protect identifying variables, even though if the data were to be used for actual research, it would have been easier to protect variables containing information on teachers' proficiency.

3 Results

Our study is only the beginning of extensive research, in which we try to analyse the use of statistical disclosure control methods proposed in literature. The data we used in this part of our study are generally difficult to handle, but they give a very good idea of how useful these methods are in practice.

Microaggregation was applied to categorical data with some changes described in section 2. PRAM was applied in two different ways and it was assumed that

protected data would be used without the PRAM matrix. First we tested PRAM without any bandwidth restrictions. This lead to considerable changes in frequencies, even with small changing probabilities. We tried to overcome this problem by choosing a bandwidth of 2.

3.1 Results of microaggregation

It is suggested in literature that microaggregation is not strong enough to protect data when it is applied one variable at the time and when it is the only protection the data has. Because of this, in our study we applied microaggregation to three variables at once. The variables we protected were the age of the teacher, his/her position and the school level he/she teaches at. Here we see the changes in the frequencies of the teachers' school level. Later in this chapter you can see how the frequencies of this same variable, school level, change when the data were protected by PRAM.

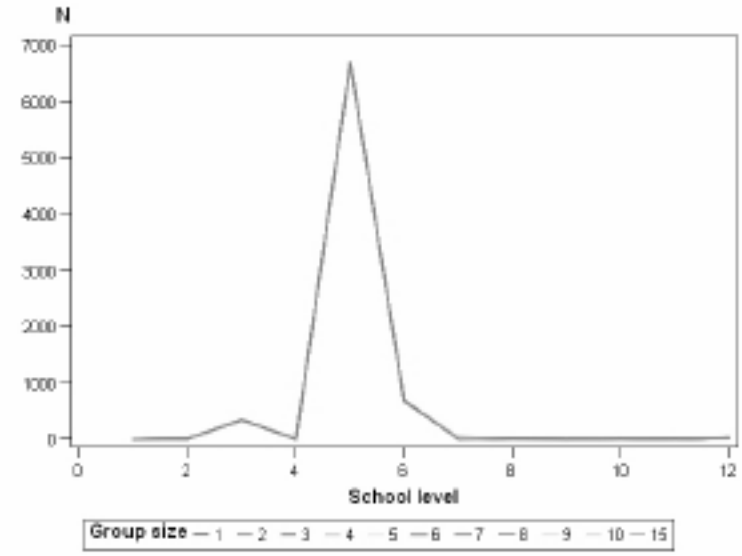


Fig 3.1 Changes in frequencies when data are protected with microaggregation

Using microaggregation in the case of a categorical variable seems to have no significant effect on frequencies, as seen in Figure 3.1. When we consider this and are only concerned with information loss, it is a very good result. However, if data are protected by microaggregation only, this leads to a problem with disclosure risk. It is obvious that there are only a few changes in the values of a record. This procedure brings hardly any uncertainty when it comes to identification of a record. The changes in frequencies are so small that they are hard to see from Figure 3.1. To get a better idea, the actual values can be found in Table 3.1.

	1	2	3	4	5	6	7	8	9	10	15
1	1	0	0	0	0	0	0	0	0	0	0
2	10	8	9	8	10	6	7	8	9	10	15
3	343	346	345	344	345	348	350	344	351	340	345
4	7	6	6	4	5	0	0	8	0	10	0
5	6672	6672	6673	6678	6673	6676	6671	6670	6673	6668	6688
6	686	684	684	684	680	684	693	672	693	690	690
7	22	24	24	20	30	30	21	24	18	30	0
8	7	8	9	12	5	6	7	8	9	0	15
9	7	4	6	4	10	6	7	8	0	10	0
10	16	18	15	16	15	12	14	16	18	10	15
11	3	4	3	4	0	6	0	8	9	10	15
12	24	24	24	24	25	24	28	24	18	20	15

Table 3.1 Changes in frequencies when data are protected with microaggregation

3.2 Results of PRAM

As mentioned before, our data were not accommodating in any way, and it was expected that our SDC methods would fail in some ways. In this case, one must definitely question whether PRAM should be used for these data at all. There are more than 6,500 records in one of the categories in our data, and then there are categories that have nearly no observations. This lead to a situation where our distribution tends to smooth out when protection is applied, as demonstrated in Figure 3.2. The actual values can be found in Table 3.2 and illustrate in detail how those small frequencies tend to increase.

	0.00	0.03	0.05	0.07	0.10	0.15	0.20	0.25	0.30	0.35	0.40
1	1	22	30	59	74	94	142	164	200	246	301
2	10	25	45	61	79	92	139	194	221	241	265
3	343	357	361	359	395	382	430	437	459	455	505
4	7	28	37	57	61	104	140	165	206	272	319
5	6672	6487	6347	6202	5978	5701	5350	5049	4693	4418	4011
6	686	679	678	693	684	697	684	663	648	671	674
7	22	39	61	72	89	150	145	192	228	232	272
8	7	36	33	49	77	108	158	172	250	247	299
9	7	24	44	51	71	101	152	177	229	231	275
10	16	36	64	73	89	119	152	188	233	272	315
11	3	27	43	60	99	110	142	196	201	266	262
12	24	38	55	62	102	140	164	201	230	247	300

Table 3.2 Changes in frequencies when data are protected with PRAM and no bandwidth is used

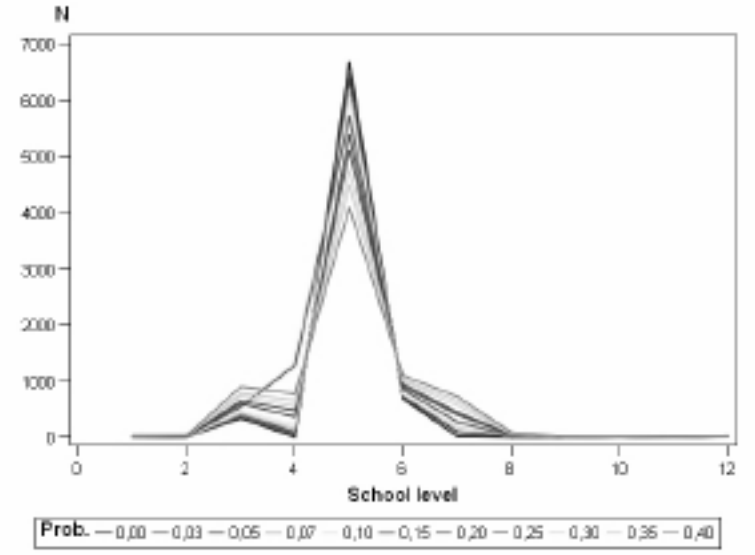


Fig 3.2 Changes in frequencies when data is protected with PRAM and no bandwidth was used

If we look at a changing probability of 0.10 or more, it is clear that the results from these data will not coincide with the ones from the original data unless a probability matrix is taken into account. However, if we are interested in using PRAM for some kind of demonstration data, changes that occur with a 0.10 probability could be acceptable.

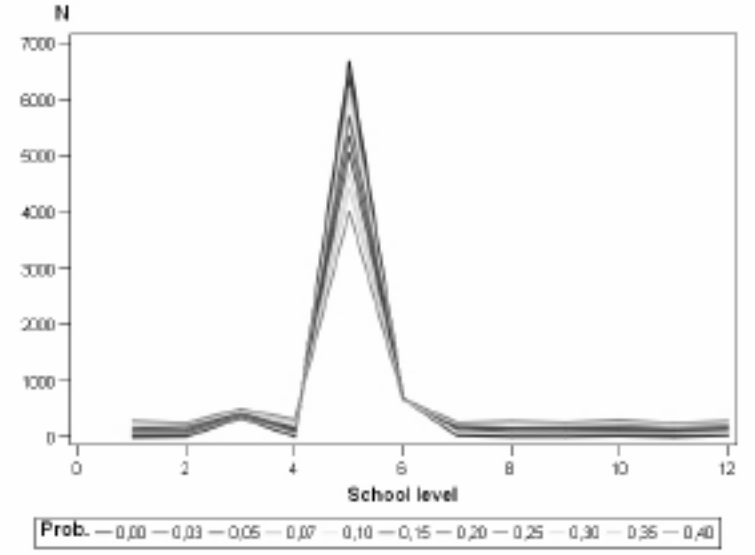


Fig 3.3 Changes in frequencies when data are protected with PRAM and the bandwidth is 2

When we obtained these results, we were slightly disappointed to see this smoothing effect on distribution. We decided to use a bandwidth of 2 and were expecting it to help. Unfortunately, protection of our data where the frequencies differ by so much did not improve with this either. Restricting the change has an effect when it comes to categories that are not next to category 5, into which most of the cases fall. The categories next to 5 showed the same remarkable increase as they did in the case without any bandwidth, as can be seen from Figure 3.3 and Table 3.3.

As can be seen in Figures 3.2 and 3.3, how well an SDC method performs depends not only on the data, but also on the attributes. If the handler of the data gets carried away and forgets to check the information content of the data, the results from the protected data may differ greatly from the ones from original data.

We concentrated on the usefulness of data, so it is advisable to briefly consider also whether the proposed values of the attributes are appropriate to protect all individuals against disclosure. It is also beneficial to keep in mind that since PRAM is based on a probability distribution, the user gets different protected data every time the procedure is applied. This means that our results are an example of how this method performs, not an absolute truth of its usefulness.

	0.00	0.03	0.05	0.07	0.10	0.15	0.20	0.25	0.30	0.35	0.40
1	1	2	3	5	9	18	25	26	31	28	29
2	10	15	12	14	20	23	19	26	41	35	36
3	343	376	417	431	469	547	605	651	735	789	904
4	7	50	88	150	218	1287	368	483	559	660	778
5	6672	6475	6377	6182	6047	5717	5382	5114	4748	4485	4083
6	686	711	736	763	767	849	899	946	995	1031	1100
7	22	97	99	175	194	276	408	449	575	654	728
8	7	13	15	27	25	30	40	54	61	68	92
9	7	6	8	8	7	8	8	3	10	10	9
10	16	17	16	17	17	18	16	20	19	17	11
11	3	4	3	4	4	7	4	6	5	4	13
12	24	22	24	22	21	18	24	20	19	17	15

Table 3.3 Changes in frequencies when data are protected with PRAM and the bandwidth is 2

4 Conclusion and future study

This part of research was only the beginning of more extensive research. At this moment it has no actual measures of information loss or identification risks. This is a severe weakness but we have made some assumptions for measuring both. In our case, we were more interested in general usability than in exact individual figures.

It is clear that when data are protected using PRAM, researchers have to use the PRAM matrix in order to obtain correct results. However, there are not too many researchers who are willing to undertake extra work because their data have been protected. There are even some researchers who simply are not able to do such work. In our opinion, PRAM is quite a promising method when we want to add some uncertainty into identification. It can be seen that PRAM can be used as it is if the data handler is working with some kind of demonstrative data. Researchers can be allowed access to this type of data in case we can not allow them access to the actual data even on the NSI's premises. This method can have potential in the future, if researchers are more familiar with statistics and mathematics than they are at present.

We used microaggregation for categorical data even if it was meant to be used for continuous data only. This choice resulted in some problems, but imitated reality in which you only have limited options to choose from. In our opinion, the usefulness of microaggregation lies in numerical data. We can see some possibilities in using microaggregation for categorical data, but in that case some other protection must also be applied to the data.

References

A CENtre of EXellence for Statistical Disclosure Control. http://neon.vb.cbs.nl/cenex/

Computational Aspects of Statistical Confidentiality. http://neon.vb.cbs.nl/casc/

de Wolf, Peter-Paul, van Gelder, Ilan (2004). An empirical evaluation of PRAM. Discussion paper 04012. Statistics Netherlands, Voorburg/Heerlen.

Gouweleeuw, J., Kooiman, P., Willenborg, L., and de Wolf, P. (1998). Post Randomisation for Statistical Disclosure Control: Theory and Implementation. *Journal of Official Statistics*. Vol. 14, No.4, 463-478.

Group Crises (2004). *Microaggregation for Privacy Protection in Statistical Databases*. Research Reports. In July 2005. <http://vneumann.etse.urv.es/publications/reports/>.

Konnu, Janika (2006). Mikroaineistojen tilastolliset tietosuojamenetelmät henkilötilastoissa. (Statistical Disclosure Control Methods for Personal Microdata; in Finnish only). Master's theses in Statistics. University of Jyväskylä: Department of Mathematics and Statistics.

Anonymisation of Linked Employer Employee Datasets using the example of the German Structure of Earnings Survey.

Hans–Peter Hafner*, Rainer Lenz**

* Research Data Centre of the Statistical Offices of the Länder, Statistical Office of Hesse, Rheinstr. 35/37, 65185 Wiesbaden, Germany (hhafner@statistik-hessen.de)

** Department I: architecture, civil engineering and geoinformatics, University of Applied Sciences Mainz, Holzstr. 36, 55116 Mainz, Germany (rainer.lenz@fh-mainz.de)

Abstract. The anonymisation of linked employer employee datasets constitutes a special problem for data producers. Concerning the employees there is generally less risk of reidentification, but their information can be used to identify the employer. We present a strategy that permits to measure dependencies between employer and employee data, to evaluate whether these dependencies have an impact on the reidentification risk of the employer and, if necessary, to anonymise the data of the employees in such a manner that the reidentification of the employer is very complicated. Finally, we embed this strategy in the generation process for a scientific use file of the German Structure of Earnings Survey 2001.

1 Introduction

Linked employer employee datasets (LEED) enable labour market researchers to split observed effects in one fraction caused by the employer and one fraction dependent on the employee. Since the middle of the 1990ies the number of analyses in this field has escalated. Abowd and Kramarz (1999) provide an overview on projects executed during the 90ies and on datasets from 17 countries that were available at that time.

LEED for Germany that are currently available to the scientific community are the Linked Employer Employee Data of the Institute for Employment Research (LIAB) and the Structure of Earnings Survey of the Federal Statistical Office and the statistical offices of the Länder (federal states). A description of the LIAB and selected studies conducted with it can be found in Alda et al. (2005); an overview on the Structure of Earnings Survey and related studies is provided by

Hafner and Lenz (2007).
As interesting as LEED are to the science community, it is very complicated for the data producers to generate anonymised scientific use files from such sources so that researchers can work with them in their institutes.

In this paper we propose a procedure, which besides classical information reducing methods, applies only selective one-dimensional microaggregation to especially sensitive variables. We show that thereby the reidentification risk associated with the data is reduced.
Chapter 2 summarises the thoughts that have to be considered for the anonymisation of the employer data. Chapter 3 deals with the information about the employees. Here we present a method that guarantees that the variables of the employees do not increase the disclosure risk of the employer. Following an overview on the methodology and the attributes of the Structure of Earnings Survey (Chapter 4) we apply our procedure to the German dataset of this survey of the year 2001.

2 Anonymisation of the Employer Data

In order to reidentify an enterprise, a data intruder needs additional knowledge, for example from commercial databases. This additional knowledge must have attributes in common with the target file (key variables). For Germany, the greatest electronically available resource is the so called Markus database. Details can be found at http://www.creditreform.de/.
The risk of reidentification can now be determined by means of matching experiments between the target file and the additional knowledge. The aim of the data intruder is to decide whether or not the pair $(a, b) \in A \times B$ of records belongs to the same employer. In a non-technical way, the concept of matching may be introduced as a way of bringing together pieces of information in pairs from two records taken from different data sources. For this purpose, a reasonable concept of similarity is necessary. Roughly spoken, the greatest possible similarity between two records turns into identity if the considered records correspond with regard to all key variables. In the case of small deviations of the key variables, two objects are felt to be strongly related, so that the matching result essentially depends on the concept of similarity. For technical details see Lenz (2006).
In recent years, regarding German business statistics, often the value 0.5 was accepted as the upper bound for the reidentification risk, provided that this value is reached only for a few parts of the file, for example for large companies in low frequented branches of economic activity, see Ronning et al. (2005). When evaluating the risk one has to consider that the calculation presumes that a data intruder has some knowledge of participation about an enterprise. Therefore the risk in sample surveys is reduced by a factor corresponding to the sample fraction of the stratum.

In practice, the risk is minimised by combining especially vulnerable classes of categorial variables with others and by applying data perturbing procedures like microaggregation to numerical variables.

3 Anonymisation of the Employee Data

In most cases the risk of reidentification for employees is negligible since there is no systematic additional knowledge. Furthermore we restrict our thoughts to sample surveys. Hence a data intruder has no participation knowledge about a person, s.t. the information about the employees is sufficiently anonymised when taken alone. However, it might be possible to draw conclusions from it about the enterprises so that the anonymisation made can be reversed in parts by a data intruder. To formalise these thoughts we need some notations.
Let A be an employer file whose attributes are the random variables $S_1, \ldots, S_n$, and B an employee file with attributes $T_1, \ldots, T_m$. $C = (A, B)$ is a *linked employer employee file* if it is possible to assign the employees covered in B to the employers covered in A.
An attribute Y of an employee file that is independent of an attribute X of the employer file would not yield further insights to a data intruder. But in practice absolute independence is rather the exception. Hence we need measures to calculate the degree of dependence between two variables. In doing so we have to differentiate by the scale of the attributes:

- Both attributes are metric. Then Pearson's correlation coefficient measures the degree of dependency.
- Both attributes are ordinal. Then one can use Spearman's rank correlation coefficient. The same holds if one attribute is ordinal and the other metric.
- Both attributes are nominal. Then Cramer's V is an adequate measure, which can also be applied if one attribute is ordinal and the other nominal.
- One attribute is nominal, the other metric. In this situation one can use the measure of association η. In contrast to the other measures, η is not symmetric. That means the metric attribute is the dependent variable while the nominal attribute is the independent one. Let $X = (x_1, \ldots, x_l)$ be the nominal attribute, n_i the number of observations in category i, $\overline{y}$ the mean of Y and $\overline{y}_i$ the mean of Y in category i. Then we define

$$\eta = \sqrt{\frac{\sum_{i=1}^{l}\sum_{j=1}^{n_i}(y_{ij}-\overline{y})^2 - \sum_{i=1}^{l}\sum_{j=1}^{n_i}(y_{ij}-\overline{y}_i)^2}{\sum_{i=1}^{l}\sum_{j=1}^{n_i}(y_{ij}-\overline{y})^2}} \qquad (1)$$

In empirical research it is common to assume a strong association between the two variables whenever the measure exceeds a value of 0.3. As regards the application

to datasets of official statistics, the determination of the bound will depend on the need for data protection. There must be more protection if the data are very sensitive or if the benefit of a reidentification seems very high. This benefit is influenced by the age of the data and by the availability of the information through other sources.
To test which combinations of variables of the linked employer employee file $C = (A, B)$ are especially vulnerable, we compute the measures of association for all key variables of A and all variables of B.
For the rest of this chapter, let S be an attribute of the employer file A and T an attribute of the employee file B so that the value of the measure of association between S and T is above the predefined bound. To simplify matters, we suppose that S and T are both categorical, which can always be achieved by grouping values. Let s be the number of categories of S, t the number of categories of T. Furthermore, A^* should be the anonymised version of A, and S^* the attribute that originated from S in A^*.

At this point it seems necessary to meet some assumptions about the behaviour of a data intruder. These are of pure theoretic nature since the existence of a data intruder is a hypothetical construct.

Below we describe in three steps how a data intruder might proceed in order to find out more about an employer by using the association between S and T.
Step 1: Since the anonymisation of A took place without using data modifying methods, the intruder knows (because of the description of the anonymisation) for every value x^* of S^* the set $X = \{x_1, \ldots, x_k\}$ containing the corresponding original value x.

Step 2: The intruder manages to get the marginal distributions of T for every category of S. Maybe he finds them in a publication of the statistical office or he asks for calculation via remote data access.

Step 3: The intruder compares every employer's distribution of T with the marginal distributions of T for $x_1, \ldots, x_k$ and he chooses that x_i which presents the smallest differences in respect to the distribution of T.
The adequate statistical procedure for step 3 is to perform a test of goodness of fit. In the case of discrete variables the χ^2 test is the most common tool. Using this tool, the observed sample is analysed as to whether it can be a random sample of a specific distribution by comparing the observed and the expected frequencies. Let $e_i, i = 1, \ldots, t$, be the expected frequency for category i of T and f_i the observed frequency. Then the well known test statistic χ^2 is obtained by

$$\chi^2 = \sum_{i=1}^{t} (f_i - e_i)^2 / e_i \qquad (2)$$

The null hypothesis indicates that the observed sample originates from the assumed distribution. It is rejected if the value of χ^2 is greater than the quantile for the chosen level of significance.
The distribution function of (5) fits asymptotically the distribution function of the χ^2 function with $t-1$ degrees of freedom. The rule of thumb mostly mentioned for the application of the χ^2 test is that for at least 80 percent of the categories the expected frequencies should be 5 or more and that the expected frequencies of the other categories should be at least 1. Koehler and Larntz (1980) show that the appoximation is suitable even for smaller expected frequencies provided that the square of the number of observations at least equals the number of categories multiplied by 10.
To improve the goodness of the prediction, one can combine categories whose fraction is very low in all marginal distributions that have to be tested. A modification of the χ^2 test that can be applied to small samples and small expected frequencies has been developed by Haldane. He does not compare the test statistic with the asymptotic χ^2 distribution, but with the exact distribution that holds under the null hypothesis, for instance see Bortz et al. (1990).
Let us now return to the data intruder. We assume that he has conducted his tests and calculated the statistics. Before he makes his decision, he has to determine a level of significance. We suppose that he can live with a probability of error of 20 percent; that is he chooses $\alpha = 0.2$. If the null hypothesis is accepted for exactly one value the decision is clear. The case that the null hypothesis is accepted for more than one value should not occour in practice; but it will happen very often that none of the alternatives is accepted since the χ^2 test almost always rejects the null hypothesis if the sample size is large. In this case one can decide on the basis of the corrected contingency coefficient C_{corr}.

The χ^2 test and the contingency coefficient provide global measures for the match of distributions. If no definite decision can be made on this basis it is additionally recommended that the components of contingency are compared. The component of contingency c_i for the category $i \in \{1, \ldots, t\}$ is defined as

$$c_i = \begin{cases} +\sqrt{\frac{(f_i-e_i)^2/e_i}{n+((f_i-e_i)^2/e_i)}} & \text{if } f_i - e_i \geq 0 \\ \\ -\sqrt{\frac{(f_i-e_i)^2/e_i}{n+((f_i-e_i)^2/e_i)}} & \text{if } f_i - e_i < 0. \end{cases}$$

A component of contingency with value 0 corresponds to a perfect match of the observed sample with the testing distribution with respect to this category; a negative (positive) component indicates a lower (higher) fraction of the category in the sample. The data intruder will now look at the components of contingency of the categories of T which are typical for the alternative x_i that should be tested. We call a category d of T *characterising* for x_i if the following conditions are satisfied:

1. The fraction of d in the marginal distribution of T given x_i exceeds a value f.

2. The fraction in category x_i exceeds a bound g in the distribution of d over the categories of S.

3. There is at least one alternative $x_j \neq x_i$ such that $|P(s|x_i) - P(s|x_j)| > h$ where h is a specified bound.

The last condition excludes that the fractions of the category are nearly equal for all alternatives. Such a category would not contribute to the decision-making.
Regarding the selection of f, g and h, one has to look at the distributions of the attributes involved; therefore, no universally valid statements are possible. After the selection of the characterising categories, the data intruder looks at the corresponding components of contingency. He may either sum up all these components separately for every alternative or sum up only the positive components in each case. An argument for summing up only the positive components is that not in every case all characterising categories are represented equally well. A value appearing above the average of some characterising categories is often a better indicator as we will see in our application in chapter 5. If the global contingency coefficient and the analysis of the single components yield the same result, the decision comes down to this alternative. In all other cases there is no sufficient confidence.
Finally, we have to evaluate the thoughts outlined above from the point of view of the data producers. Let $r_1, \ldots, r_s$ be the risks of reidentification for the categories $1, \ldots, s$ of S in the original employer file A and let r^* be the risk of reidentification for the anonymised category x^*. Furthermore, let p be the fraction of employers whose original value x_i can be derived from x^* with the help of the attribute T and the method described above. For these employers the risk of reidentification is as high as if no anonymisation with respect to S had taken place. Hence, the risk of reidentification for an employer of category x_i adds up to

$$pr_i + (1 - p)r^*. \tag{3}$$

If (7) exceeds 0.5, the categories of T that contributed most to the disclosure of x_i (that means, the categories with the highest fractions of positive components of contingency) have to be combined with other categories. This subsumption can be carried out for the complete dataset or, alternatively, only for those employers with value x^* for S.

4 The Structure of Earnings Survey

Based on an EC regulation of 1999, the survey is held in all EU countries every four years, so that the data produced are comparable all over Europe. As most

countries conducted the latest survey for 2002, the next one will be performed for 2006.
The group of reporting units comprises local units of the industry and selected parts of the service sector. The survey covers all employees who are subject to social insurance contributions and receive a remuneration in the month of report (October of the year of survey).The SES is a two-stage sample survey. In the first sampling stage, a stratified random sample is drawn from the local units. At the second stage, the employees to be included from the selected local units are determined through the personal identification number shown on the staff lists. For 2001, a total of a good 22,000 local units supplied data on over 845,000 employees.

There are separate questionnaires for data on the local unit and one each (or several for larger local units) for white-collar and blue-collar employees. Further information on the methodology and variables of the 2001 SES is contained in Frank-Bosch (2003) and in the metadata provided on the web site of the research data centres of the statistical offices of the Federation and the Länder (http://dok.fdz-metadaten.de/6/62/621/621110/erheb/200100/).

5 Anonymisation of the German Structure of Earnings Survey 2001

Since spring 2005 the research data centres of the Federal Statistical Office and the statistical offices of the Länder have conducted a project with the aim to generate a scientific use file of the German structure of earnings survey taken in 2001. The project has been concluded in autumn 2006 with the publication of the file. Scientists participated in an advisory capacity in the conception of the anonymised dataset to ensure that the result will be of interest to a broad circle of users.
At first, the key question was which regional units should be displayed. Two alternatives with five and eight regions consisting of adjacente federal states were tested. Depending on the model used, some 30 to 40 economic sectors were displayed. Furthermore, the number of employees of a company was microaggregated if a company had at least a thousand employees or if it was among the three largest companies of the economic sector in the region. Each group for microaggregation consisted of at least three companies.
The key variables which the employer dataset had in common with commercial databases were the region, the economic sector, the number of employees of the enterprise and the influence of the public sector. The last-mentioned attribute can be compared with *partner - agency, state, administration* in the Markus database. Using these attributes, matching experiments as described in chapter 2 were conducted to calculate the risks of reidentification for the several alterna-

tives. It transpired that the risks for some economic sectors were too high when eight regions were displayed. Thus we opted for five regions and we lowered the threshold from which the number of employees was microaggregated to 500. It turned out that the attribute *participation of the public sector* was not critical with respect to reidentification; hence we could display the original value.

Now we will describe the anonymisation of the employee data more precisely following the scheme we developed in chapter 3. As an example we take the two combined economic sectors of the drapery / clothing trade and the leather industry. Of these sectors, 429 local units and 14,826 employees are contained in the survey.
At first we must examine which attributes of the employees have a strong association with the economic sector. Table 1 shows that the measure of association indicates a strong dependence on the economic sector only for the occupation class. Since all other values are far below 0.3 we conclude that there are no further variables with a strong relation to the economic sector. Hence we can carry

Table 1: Measures of association between the economic sector and the attributes of the employees

Sex	0.044
Wage Tax Class	0.045
Allowance for Children	0.043
Position in Job	0.119
Education	0.071
Type of Contract of Employment	0.022
Occupation Class (2-digit)	0.659
Paid Working Hours Total	0.003
Gross Earnings in Accounting Period	0.041
Extra Pay for Shift Work	0.077
Extra Pay for Night Work	0.113
Income Tax	0.035
Pension and Unemployment Insurance	0.048
Health and Care Insurance	0.055
Gross Annual Earnings	0.035
Supplementary Grants in the Reporting Year	0.051
Net Annual Earnings	0.033
Holiday Entitlement	0.003
Net Earnings in Accounting Period	0.043

out two χ^2 tests between the economic sector and the occupation class. First we test the observed distribution of the occupation classes against the distribution of the drapery / clothing trade, and then against the one of the leather industry. For that purpose we combine some occupation classes which contain only few

employees so that we obtain 18 classes. According to Koehler and Larntz (1980) , the χ^2 distribution is a good approximation if the number of observations is at least $\sqrt{10 * 18} = 13.42$. Choosing $\alpha = 0.2$, one of the two null hypotheses is accepted in only 15 cases. In 12 of theses cases the prediction is correct, in the other cases the number of observations is smaller than 14. As expected, the χ^2 value taken alone is not a good predictor. For this reason, the next step consists of the calculation of the contingency coefficients. On the basis of these coefficients, the prediction is correct in 394 of 429 cases. Moreover, there is only a small difference in the fraction of correct predictions between the drapery / clothing trade and the leather industry. While for the drapery / clothing trade 92.6 percent of the assignments are correct, this applies to only 90.4 percent of the local units of the leather industry.

Before we start to calculate the components of contingency, we have to decide which occupation classes are characterising for the economic sectors under review. Tables 2 and 3 show that textile fabricators and textile producers doubtlessly are characterising occupation classes for the drapery / clothing trade, and leather producers, leather and coat fabricators for the leather industry. Furthermore, 73.8 % of all workers with spinning occupations and 76.8 % of the textile refiners are employed in the drapery / clothing industry, so that these occupations can also be regarded as characterising for this economic sector. For all other occupations listed in the tables below, the fraction of corresponding workers amounts to less than 5 % in the two economic sectors. In accordance with condition 2. for characterising categories, we leave these occupation classes out of consideration.

Table 2: Fractions of the most frequent occupation classes: Drapery / Clothing Trade

Textile Fabricators	19.3 %
Office Workers	14.6 %
Textile Producers	9.3 %
Product Inspectors, Shipping Finalisers	6.3 %
Technicians	5.5 %
Product Traders	5.5 %
Storekeepers, Warehousemen, Transport Workers	4.4 %
Spinning Occupations	4.4 %
Textile Refiners	3.0 %

If we add up only the positive coefficients we can make a prediction for 326 of the 429 local units. Out of 238 predictions for units of the drapery / clothing trade 225 are correct (94.5 %), out of 88 predictions for units of the leather industry 86 are correct (97.3 %). If we sum up all coefficients, we reach a correct prediction for only 56.6 % of the units of the drapery / clothing trade, while the fraction of correct predictions in the leather industry is 88.5 %. These figures suggest that the results are getting worse by using negative coefficients. Thus it

Table 3: Fractions of the most frequent occupation classes: Leather Industry

Leather Producers, Leather and Coat Fabricators	47.4 %
Office Workers	16.9 %
Product Traders	4.8 %
Technicians	3.9 %
Entrepeneurs, Organisers, Accountants	3.1 %
Plastics Fabricators	3.1 %
Storekeepers, Warehousemen, Transport Workers	3.0 %

might be better to use only the positive coefficients and to be content with fewer assignments. In exchange the risk of a misclassification is small.
If we combine the results of the analysis of the contingency coefficient and of the separate components and assign an economic sector to a local unit only if both predictions correspond, then there is a very small risk for the data intruder. He can make predictions for 82 local units of the leather industry and all are correct; for the drapery / clothing trade 181 of 190 (95.3 %) possible predictions are correct. Thus he can choose whether he wants to assign more units with a higher risk or fewer units with a lower risk.
We suppose the data intruder to be risk averse and assigns only units for which both analyses yield the same result. Then he can correctly assign 82 of the 104 units (78.9 %) of the leather industry.

As mentioned in section 2, matching experiments are applied in order to estimate the reidentification risk for employers of specific industries. Regarding the leather industry, the resulting risk is 0.61. If one joins the two sectors leather industry and drapery / clothing trade, the risk is reduced to 0.12. Hence, with (7) the overall risk for local units of the leather industry is estimated by $0.789 * 0.61 + 0.211 * 0.12 = 0.506$. Since this value exceeds 0.5 we have to combine the characteristic occupation classes of the sectors of the drapery / clothing trade and leather industry.

The application shows that our methods can be applied to linked employer employee datasets in order to increase the data protection. However, further experience is needed to improve and to standardise the suggested methods so that they will be easier applicable and less time-consuming.

References

Abowd, J. M. and Kramarz, F. (1999) "The Analysis of Labor Markets Using Matched Employer-Employee Data", In: Ashenfelter, O., Card, D. (eds.): *Handbook of Labor Economics*, Amsterdam, vol. **3**, 2629–2733.

Alda, H., Bender, S. and Gartner, H. (2005) "The linked employer-employee dataset of the IAB (LIAB)", *IAB Discussion Paper* **06/2005**, Institute for

Employment Research, Nuremberg, Germany

Bortz, J., Lienert, G. A. and Boehnke, K. (1990) "Verteilungsfreie Methoden in der Biostatistik", Berlin, Heidelberg: *Springer*

Frank-Bosch, B. (2003) "Verdienststrukturen in Deutschland: Methode und Ergebnisse der Gehalts- und Lohnstrukturerhebung 2001", *Wirtschaft und Statistik* **12/2003**, 1137–1151

Hafner, H.-P. and Lenz, R. (2007) "Die Gehalts- und Lohnstrukturerhebung. Methodik, Datenzugang und Forschungspotential", *discussion paper* **18**, Research data centres of the statistical offices of the federal and the states

Koehler, K. and Larntz, K. (1980) "An empirical investigation of goodness-of-fit statistics for sparse multinomials", *JASA* **370 (75)**, 336–344

Lenz, R. (2006) "Measuring the disclosure protection of micro aggregated business microdata. An analysis taking as an example the German Structure of Costs Survey", *Journal of Official Statistics* **22 (4)**, Sweden, 681–710

Ronning, G., Sturm, R., Höhne, J., Lenz, R., Rosemann, M., Scheffler, M. and Vorgrimler, D. (2005) "Handbuch zur Anonymisierung wirtschaftsstatistischer Mikrodaten", *Statistik und Wissenschaft*, volume **4**, Statistisches Bundesamt

Comparing Fully and Partially Synthetic Data Sets for Statistical Disclosure Control in the German IAB Establishment Panel

Jörg Drechsler, Stefan Bender* and Susanne Rässler**

* Institute for Employment Research (IAB), Regensburger Straße 104, 90478 Nürnberg, Germany, joerg.drechsler@iab.de, stefan.bender@iab.de
** Otto-Friedrich-University Bamberg, Department of Statistics and Econometrics, Feldkirchenstraße 21, 96045 Bamberg, Germany, susanne.raessler@sowi.uni-bamberg.de

Abstract: In this paper we discuss the advantages and disadvantages of two approaches that provide disclosure control by generating synthetic data sets: The first, proposed by Rubin (1993), generates fully synthetic data sets while the second suggested by Little (1993) imputes values only for selected variables that bear a high risk of disclosure. Changing only some variables in general will lead to higher analytical validity. However, the disclosure risk will also increase for partially synthetic data sets since true values remain in the data. Thus, agencies willing to release synthetic data sets will have to decide, which of the two methods balances best the trade-off between data utility and disclosure risk for their data. We offer some guidelines to help making this decision.
We apply the two methods to a set of variables from the 1997 wave of the German IAB Establishment Panel and evaluate their quality by comparing regression results from the original data with results we achieve for the same analyses run on the data set after the imputation procedures. The results are as expected: In both cases the analytical validity of the synthetic data is high with partially synthetic data sets outperforming fully synthetic data sets in terms of data utility. But this advantage comes at the price of a higher disclosure risk for the partially synthetic data.

1 Introduction

A new approach for statistical disclosure control was suggested by Rubin in 1993: Generating fully synthetic data sets to guarantee confidentiality. His idea was to treat all the observations from the sampling frame that are not part of the sample as missing data and to impute them according to the multiple imputation framework. Afterwards, several simple random samples from these fully imputed data sets are released to the public. Because all imputed values are random draws from the posterior predictive distribution of the missing values given the observed values, disclosure of sensitive information is nearly impossible. However, the quality of this method strongly depends on the accuracy of the model used to impute the "missing" values.

To overcome this problem, a related approach suggested by Little (1993) replaces observed values with imputed values only for variables that bear a high risk of disclosure or for variables that contain especially sensitive information leaving the rest of the data unchanged. This approach, discussed as generating partially synthetic

data sets in the literature, has been adopted for some data sets in the US (see for example Abowd and Woodcock, 2001, 2004 or Kennickell, 1997).

In this paper we apply both methods to an establishment survey of the German Institute for Employment Research (IAB) and discuss advantages and disadvantages for both methods in terms of data utility and disclosure risk.

2 Application of the Two Synthetic Data Approaches to the IAB Establishment Panel

2.1 The IAB Establishment Panel

The IAB Establishment Panel is based on the employment statistics aggregated via the establishment number as of 30 June of each year. Consequently the panel only includes establishments with at least one employee covered by social security. For the imputation of the IAB Establishment Panel, we use additional information from the German Social Security Data (GSSD). The basis of the GSSD is the integrated notification procedure for the health, pension and unemployment insurances. We use the establishment identification number to match the selected establishment characteristics aggregated from the employment register with the IAB Establishment Panel.

2.2 Generating Fully Synthetic Data Sets

We only impute values for a set of variables from the 1997 wave of the IAB Establishment Panel. As it is not feasible to impute values for the millions of establishments contained in the German Social Security Data for 1997, we sample from this frame, using the same sampling design as for the IAB Establishment Panel: Stratification by establishment size, region and industry. After the imputation procedure, all original observations from the Establishment Panel are omitted and only the imputed values are kept for analysis.

2.3 Generating Partially Synthetic Data Sets for the IAB Establishment Panel

For this study, we replace only two variables (the number of employees and the industry, coded in 16 categories) with synthetic values, since these are the only two variables that might lead to disclosure in the analyses we use to evaluate the data utility of the synthetic data sets. Especially large firms can be identified without difficulty using only these two variables.

We define a multinomial logit model for the imputation of the industry code and a linear model stratified by four establishment sizes defined by quartiles for the number of employees.

3 Comparison Between the Original and the Imputed Data Sets

3.1 Data Utility

To create the fully synthetic data sets we draw ten new samples from the German Social Security Data (GSSD) and impute every sample ten times using chained equations as implemented in the software IVEware by Raghunathan, Solenberger and Hoewyk. For the imputation procedure we use 26 variables from the GSSD and reduce the number of panel variables to be imputed to 48 to avoid multicollinearity problems. For the partially synthetic data sets, we use the same number of variables in the imputation model, but no samples are drawn from the GSSD, since the original sample is used. We generate the same number of synthetic data sets, but the modelling is performed using own coding in R.

For an evaluation of the data utility of the synthetic data, we compare analytic results achieved with the original data with results from the synthetic data. The comparison is based on an analysis by Thomas Zwick: 'Continuing Vocational Training Forms and Establishment Productivity in Germany' published in the German Economic Review, Vol. 6(2), pp. 155-184 in 2005.

Zwick analyses the productivity effects of different continuing vocational training forms in Germany. He argues that vocational training is one of the most important measures to gain and keep productivity in a firm. Zwick runs a probit regression using the 1997 wave of the Establishment Panel. The regression shows that establishments increase training if they expect to loose workers. For his analysis, Zwick runs the regression only on units with no missing values for the regression variables, loosing all the information on establishments that did not respond to all questions used. This might lead to biased estimates if the assumption of a missing pattern that is completely at random (see for example Rubin, 1987) does not hold. For that reason, we compare the regression results from the synthetic data sets that by definition have no missing values, with the results, Zwick would have achieved if he would have run his regression on a data set with all the missing values multiply imputed. Comparing results from Zwick's regression run on the original data and on the synthetic data sets are presented in Table 1.

All estimates are very close to the estimates from the real data and except for the variable "high number of maternity leaves expected", for which the significance level decreases to 5% for the fully synthetic data, remain significant on the same level when using the synthetic data. Obviously Zwick would have come to the same conclusions in his analysis, no matter if he would have used the fully synthetic data or the partially synthetic data instead of the real data.

However, if we compare the results from the partially synthetic and the fully synthetic data sets more closely, we see that the estimates from the partially synthetic data sets are closer to the original estimates for most coefficients, although the

industry dummies are used as covariates in the regression. Note that the univariate distribution of the industry will always be identical to the true distribution for the fully synthetic data sets, because the industry code is part of the sampling design which is identical for the original and for the fully synthetic data.

Exogenous variables	**Coeff. from org. data**	**Fully synthetic data**	**Partially synt. data**	$\beta_{fully} - \beta_{org}$	$\beta_{partially} - \beta_{org}$
Redundancies expected	0.250***	0.251***	0.260***	0.001	0.010
Many employees are expected to be on maternity leave	0.266**	0.244*	0.318**	-0.021	0.052
High qualification need exp.	0.648***	0.625***	0.642***	-0.023	-0.006
Apprenticeship training reaction on skill shortages	0.113*	0.147*	0.118*	0.034	0.005
Training reaction on skill shortages	0.527***	0.523***	0.547***	-0.004	0.019
Establishment size 20-199	0.686***	0.645***	0.702***	-0.041	0.017
Establishment size 200-499	1.355***	1.203***	1.329***	-0.152	-0.027
Establishment size 500-999	1.347***	1.340***	1.359***	-0.007	0.012
Establishment size 1000 +	1.964***	1.778***	1.815***	-0.187	-0.149
Share of qualified employees	0.778***	0.820***	0.785***	0.043	0.008
State-of-the-art technical equipment	0.169***	0.168***	0.170***	-0.001	0.001
Collective wage agreement	0.254***	0.313***	0.268***	0.059	0.014
Apprenticeship training	0.484***	0.406***	0.503***	-0.078	0.020
Number of observations	7,332	7,332	7,332		

Table 1: Comparison between the regression coefficients from the real data and the coefficients from the synthetic data

15 industry dummies and East Germany dummy

Notes: *** Significant at the 0.1% level,** Significant at the 1% level, * Significant at the 5% level; the standard errors are heteroscedasticity-corrected.

Source: IAB Establishment Panel 1997 without newly founded establishments and establishments not represented in the employment statistics of the German Federal Employment Agency; regression according to Zwick (2005)

Another way to determine the data utility is to look at the overlap between the confidence intervals for the estimates from the original data and the confidence intervals for the estimates from the synthetic data as suggested by Karr et al. (2006). For every estimate the average overlap is calculated by:

$$J_k = \frac{1}{2}\left(\frac{U_{over,k} - L_{over,k}}{U_{org,k} - L_{org,k}} + \frac{U_{over,k} - L_{over,k}}{U_{syn,k} - L_{syn,k}}\right),$$

where $U_{over,k}$ and $L_{over,k}$ denote the upper and the lower bound of the overlap of the confidence intervals from the original and from the synthetic data for the estimate k, $U_{org,k}$ and $L_{org,k}$ denote the upper and the lower bound of the confidence interval for the estimate k from the original data, and $U_{syn,k}$ and $L_{syn,k}$ denote the upper and the lower bound of the confidence interval for the estimate k from the synthetic data. This utility measure is more accurate in the sense that it also considers the significance level of the estimate, because estimates with low significance might still have a high confidence interval overlap and by this a high data utility even if their point estimates differ considerably from each other, because the confidence intervals

will increase with lower significance. For more details on this method see Karr et al. (2006). Results for our regression example are presented in Table 2.

The confidence interval overlap is high for both approaches, often more than 90%, but again the partially synthetic approach yields better results than the fully synthetic approach. The overlap is higher for all estimates except for the variable that indicates whether the establishment expects many employees to be on maternity leave. Especially, if we look at the average CI overlap over all estimates, the improvements for the partially synthetic data sets become clearly evident with an increase of the average overlap from 80.8% to 92.6%.

Exogenous variables	**CI overlap for the fully synthetic data**	**CI overlap for the partially synthetic data**
Redundancies expected	0.950	0.954
Many employees are expected to be on maternity leave	0.945	0.861
High qualification need exp.	0.923	0.980
Apprenticeship training reaction on skill shortages	0.846	0.973
Training reaction on skill shortages	0.897	0.908
Establishment size 20-199	0.760	0.901
Establishment size 200-499	0.421	0.923
Establishment size 500-999	0.955	0.973
Establishment size 1000 +	0.735	0.792
Share of qualified employees	0.846	0.972
State-of-the-art technical equipment	0.953	0.996
Collective wage agreement	0.675	0.916
Apprenticeship training	0.594	0.883
Average overlap	0.808	0.926

Table 2: Comparison of the average confidence interval overlap between the original data set and the synthetic data sets

The advantages of the partially synthetic approach become even more obvious, if we look at a regression of the number of employees transformed on a logarithmic scale on the 15 industry dummies. This model might not be the most interesting model from an economic perspective (the R^2 is low, 0.134 for the original data) but it provides useful information for our study, since it contains exactly the two variables that are synthesized for the partially synthetic approach. Table 3 shows the estimates for both approaches compared to the real estimates and the average confidence interval overlap.

Again, the partially synthetic approach provides better results, although the estimates for the fully synthetic data sets are based on exact marginal distribution for the industry. The deviation from the original estimates is lower for eleven of the 16 estimates. The significance level changes slightly for six estimates when using the fully synthetic data sets, but only for two estimates when using the partially synthetic data sets. The confidence interval overlap is higher for 13 estimates if only some variables are synthesized and the average overlap over all estimates further underlines the higher data utility for partially synthetic data sets.

Exogenous variables	Coefficients from org. data	Fully synthetic data	Partially synthetic data	CI overlap fully synthetic data	CI overlap part. synthetic data
Industry dummy 1	-1.606***	-1.794***	-1.531***	0.653	0.834
Industry dummy 2	0.774***	0.757***	0.723***	0.849	0.919
Industry dummy 3	0.098	-0.006	0.148	0.731	0.878
Industry dummy 4	-0.029	-0.204*	0.016	0.470	0.864
Industry dummy 5	-0.96***	-1.162***	-0.923***	0.477	0.908
Industry dummy 6	-1.276***	-1.495***	-1.234***	0.470	0.880
Industry dummy 7	-1.696***	-1.884***	-1.600***	0.507	0.718
Industry dummy 8	-0.505***	-0.286*	-0.605***	0.515	0.786
Industry dummy 9	0.334*	0.362**	0.320*	0.871	0.975
Industry dummy 10	-0.547*	-0.62**	-0.713***	0.914	0.799
Industry dummy 11	-1.431***	-1.531***	-1.342***	0.781	0.781
Industry dummy 12	-0.318**	-0.346***	-0.258*	0.929	0.851
Industry dummy 13	-0.442***	-0.623***	-0.395***	0.537	0.883
Industry dummy 14	-1.641***	-1.844***	-1.529***	0.589	0.731
Industry dummy 15	-0.703***	-0.719**	-0.820***	0.966	0.841
Intercept	4.831***	4.85***	4.779***	0.926	0.774
Average overlap				0.699	0.839

Table 3: Comparison of the estimates and confidence interval overlaps for a regression of the number of employees on industry dummies (the 16th dummy is the reference category)

Notes: *** Significant at the 0.1% level,** Significant at the 1% level, * Significant at the 5% level

Of course, partially synthetic data sets should always provide results that are at least as good as the ones from the fully synthetic data set for analyses that are based solely on variables left unchanged in the partially synthetic data. So, in terms of data utility, partially synthetic data sets will outperform fully synthetic data sets in most cases. Furthermore, there might be instances where defining imputation models for all variables is simply impossible, because there are so many logical constraints, bounds, and skip patterns in the data that a useful model cannot be obtained. And if it is possible to come up with a model, the imputed values might be biased and this bias is then introduced in all the other variables that are imputed on a later stage, based on the imputations for this variable.

However, the data utility benefits of the partially synthetic data sets come at the price of an increased disclosure risk that should be discussed in the following Section.

3.2 Disclosure risk

In general, the disclosure risk for the fully synthetic data is very low, since all values are synthetic values. It is not zero however, because, if the imputation model is too good and basically produces the same estimated values in all the synthetic data sets, it doesn't matter that the data are all synthetic. It might look like the data from a potential survey respondent an intruder was looking for. And once the intruder thinks, he identified a single respondent and the estimates are reasonable close to the true values for that unit, it is no longer important that the data is all made up. The

potential respondent will feel that his privacy is at risk. Still, this is very unlikely to occur since the imputation models would have to be perfect and the intruder faces the problem that he never knows if the imputed values are anywhere near the true values.

The disclosure risk is higher for partially synthetic data sets especially if the intruder knows that some unit participated in the survey, since true values remain in the data set and imputed values are generated only for the survey participants and not for the whole population. So for partially synthetic data sets assessing the risk of disclosure is an equally important evaluation step as assessing the data utility. It is essential that the agency identifies and synthesizes all variables that bear a risk of disclosure. A conservative approach would be, to also impute all variables that contain the most sensitive information. Once the synthetic data is generated, careful checks are necessary to evaluate the disclosure risk for these data sets. Only if the data sets proof to be useful both in terms of data utility and in terms of disclosure risk, a release should be considered.For this study, the disclosure risk evaluation is still in progress. First results show however that the disclosure risk is still very low for the partially synthetic data sets considered here.

4 Discussion and Conclusion

Releasing microdata to the public that guarantees confidentiality for survey respondents on the one hand, but also provides a high level of data utility for a variety of analyses on the other hand is a difficult task. In this paper we discussed two closely related approaches based on multiple imputation: The generation of fully and partially synthetic data sets. While fully synthetic data sets will never contain any originally observed values, original values are replaced only for key identifiers and/or sensitive values in partially synthetic data sets. Since imputed values can be generated for the whole population with fully synthetic data sets, but only for the survey respondents with partially synthetic data sets, knowing that a certain unit participated in a survey will be a benefit for the intruder only for the partially synthetic data sets.

Nevertheless, partially synthetic data sets have the important advantage that in general the data utility will be higher, since only for some variables the true values have to be replaced with imputed values, so by definition the correlation structure between all the unchanged variables will be exactly the same as in the original data set. The quality of the synthetic data sets will highly depend on the quality of the underlying model and for some variables it will be very hard to define good models. But if these variables don't contain any sensitive information or information that might help identify single respondents, why bother to find these models? Why bother to perturb these variables first place? Furthermore, the risk of biased imputations will increase with the number of variables that are imputed. For, if one of the variables is imputed based on a 'bad' model, the biased imputed values for that variable could be

the basis for the imputation of another variable and this variable again could be used for the imputation of another one and so on. So a small bias could increase to a really problematic bias over the imputation process.

The findings in this paper underline these thoughts. The partially synthetic data sets provide higher data quality in terms of lower deviation from the true estimates and higher confidence interval overlap between estimates from the original data and estimates from the synthetic data almost for all estimates. Still, this increase of data utility comes at the price of an increase in the risk of disclosure. Although the disclosure risk for fully synthetic data sets might not be zero, the disclosure risk will definitely be higher if true values remain in the data set and the released data is based only on survey participants. Thus, it is important to make sure that all variables that might lead to disclosure are imputed in a way that confidentiality is guaranteed. This means that a variety of disclosure risk checks are necessary before the data can be released, but this is a problem common to all perturbation methods that are based only on the information from the survey respondents.

References

2. Abowd, J.M., Woodcock, S.D. (2001). Disclosure limitation in longitudinal linked data. *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*. North-Holland, Amsterdam, p.215-277

3. Abowd, J.M., Woodcock, S.D. (2004). Multiply-Imputing Confidential Characteristics and File Links in Longitudinal Linked Data. *Privacy in Statistical Databases.* Springer Verlag, New York, p.290-297

13. Karr, A.F., Kohen, C.N., Oganian, A., Reiter, J.P. and Sanil, A. P. (2006). A framework for evaluating the utility of data altered to protect confidentiality, *The American Statistician, Vol. 60*, p.224 - 232

14. Kennickell, A.B. (1997). Multiple imputation and disclosure protection: The case of the 1995 Survey of Consumer Finances. *Record Linkage Techniques*. National Academy Press, Washington D.C., p.248-267

16. Little, R.J.A. (1993). Statistical Analysis of Masked Data, *Journal of Official Statistics, Vol. 9,* p.407-426

25. Rubin, D.B. (1987). Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons, New York

26. Rubin, D.B. (1993). Discussion: Statistical Disclosure Limitation. Journal *of Official Statistics, Vol. 9*, p.462-468

31. Zwick, T. (2005). Continuing Vocational Training Forms and Establishment Productivity in Germany. *German Economic Review, Vol. 6(2)*, p.155-184

Disclosure scenario and risk assessment: structure of earnings survey

Daniela Ichim, Luisa Franconi

Istituto Nazionale di Statistica, via C. Balbo 16, Rome, Italy.
(ichim@istat.it, franconi@istat.it)

Abstract. The anonymisation of Structure of Earnings Survey microdata is addressed. A practical implementation of the statistical disclosure control paradigm is illustrated. Two realistic disclosure scenarios for enterprise and employee re-identification are presented. Both scenarios are based on a careful analysis of the microdata file information content. Global recoding and a constrained regression model are implemented to protect the units at risk of re-identification. Using this approach, only few variables and only the records at risk are modified. This leads to an important information preservation.

1 Introduction

The Structure of Earnings Survey (SES) provides detailed information on the level and structure of remuneration of employees, their individual characteristics and the enterprise or local unit to which they belong to. The SES outcome represents an uniquely rich data source on gross earnings in Europe which is increasingly important for evidence-based policy making, in particular for monitoring economic growth and social cohesion. Furthermore, the SES data are indispensable for employers and employees as regards the demand and supply of labour.

National Statistical Institutes (NSI) disseminate more and more microdata files for research purposes. Such information dissemination is always constrained by the preservation of confidentiality of respondents. Statistical disclosure limitation methods are commonly applied to any survey microdata file. Special characteristics of microdata should be taken into account by the protection methods. Given the particular characteristics and needs of the typical user, the microdata files for research purposes should leave the possibility to perform very reliable analysis. An optimal trade-off between risk of re-identification and information loss should always be looked for and found by the NSIs.

The theoretical strategy proposed in [1] was put into practice in order to anonymise the Italian SES microdata file. Both respondent and user requirements were considered.

The paper is organized as follows. Section 2 briefly presents the Italian structure of earnings survey (reference year 2002). The adopted disclosure scenarios are described in section 3. Some preliminary work on variables is presented in section 4. Risk estimation issues and protection of enterprises and employees are discussed in section 5 and 6, respectively. An information loss assessment is provided in section 7.

2 Italian Structure of Earnings Survey

SES is one of the business surveys specified in the European Commission Regulation 831/2002. In Italy, SES data was collected by means of a sampling survey. A two stage sampling scheme was followed. Especially for this wave, Italy had obtained a derogation allowing a stratified sampling of enterprises instead of local units, as would be required by the European Commission Regulation 1916/2000. The enterprises sampling frame was the most up-to-date version of (Archivio Statistico delle Imprese Attive - statistical business register of active enterprises). Employees were sampled through a proportional to size sampling scheme. The observed variables are those indicated in the Regulation 1916/2000. Generally, the optional variables were not observed in the Italian survey. At enterprise level, structural economic variables were registered: economic activity (*Nace*), number of employees (*SizeEnt*), geographical location (*Nuts*), form of economic and financial control and existence of collective pay agreements. On employees, besides *Gender* and *Age*, variables related to education, profession and contractual position were surveyed. Annual and monthly earnings corresponding to the reference month (October 2002) were registered together with their various components. The working time was accounted for through variables like number of paid hours. More details on the Italian survey may be found in [4].

The sampling weights were computed in order to indicate each unit representativeness, see [2]. Generally, the weights should satisfy some restrictions, for example the preservation of some known population totals. The enterprise sampling weights were derived from the inverses of the inclusion probabilities by means of a multivariate calibration procedure, see [3, 4]. To compute the employees final weights, the procedure indicated in [2] was followed.

3 Disclosure scenario

Coherently to the SES hierarchical structure, two disclosure scenarios were adopted, taking into account the particular observed phenomenon: one for enterprise re-identification and one for employee re-identification. No nosy colleague or external register scenarios were deemed realistic.

3.1 Enterprise scenario

Given the existence of publicly available business registers, it may be supposed that an intruder could use such information to re-identify an enterprise. Moreover, it is known that, for business surveys, large enterprises are always included in the sample. Public business registers report general information on name, address, number of employees, principal economic activity, geographical location, turnover, exports, etc. Among these variables, *Nace*, *Nuts* and *SizeEnt* were observed in SES. Consequently, they were the key variables of the adopted disclosure scenario. Of course, only enterprises belonging to combinations with very small frequencies should be considered at risk of re-identification.

The survey confidential information related to enterprises is represented by the economic/ fiscal/ social policies that could be deduced/inferred from the variables observed on employees. But the content to be protected against confidentiality breaches is not due to the variables directly observed on enterprises. In conclusion, a disclosure scenario based on confidential variables was not deemed realistic for enterprise re-identification.

3.2 Employee scenario

The observed variables related to employees may be classified in two categories. There are "social" and "fiscal" variables. The latter cannot be subject to an external disclosure scenario since they are not publicly available, at least in Italy. The "social" variables observed in the Italian 2002 SES (age, gender, etc.) may be available in external registers, but they exhibit very high frequencies. Irrespective of other variables, the "social" ones cannot be considered as identifying variables.

In absence of any other information, variables on earnings, number of paid hours and absence days cannot be subject to any spontaneous identification. Combined only with *Gender* and *Age*, these variables cannot either be used for spontaneous employee identification. The reason is that the observed social variables are not at all identifying (high frequencies). In other words, it was believed that an intruder could not identify an employee only by means of gender, age, number of paid hours and earnings variables, for example. Moreover, the variable Management position or supervisory position (*ManPos*) is hardly identifying because of the second option in its definition.

Instead, enterprise information, representing an employee activity, could be used to identify an employee. Assuming such knowledge, an employee could be identified by means of the "social" variables. In Italy, there does not exist any reliable external register containing information on occupation and/or education. Consequently, these variables could be hardly used by an intruder, except for spontaneous identification based on very detailed personal *a-priori* knowledge. But in such cases, the disclosure information content would probably be substantially diminished. The same reasoning applies to variables related to the type of contract and length of stay in service. In conclusion, it was considered that only *Gender* and *Age* could be combined with enterprise information for employee identification.

About the disclosure content, it was assumed that an intruder could be interested only in extremely high earnings. "Small-medium" earnings were not judged at risk since in Italy many occupational categories are subject to some kind of national contract, at least as a common basis. Hence such earnings should not be "appealing" for an intruder. Given the Italian economy structure, it was believed that small-medium size enterprises (and their employees, too) were hardly identifiable because of their high frequencies. As the microdata file is to be released for research purposes, such interest of an intruder in small-medium enterprises cannot be fully claimed. Hence, it was supposed that only high earnings corresponding to large (and generally well-known) enterprises could be considered interesting by an intruder.

In conclusion, the adopted disclosure scenario assumes that the employee identification is possible by means of: a) information on enterprise (*Nace* x *Nuts* x

SizeEnt), b) social variables (*Gender* x *Age*) and c) extremely high earnings in large enterprises.

Since the anonymised microdata file would be disseminate for research purposes, modification of only key and confidential variables was deemed sufficient. The other variables could be released unchanged.

4 Recoding

For the Italian 2002 SES microdata file, several variables were recoded taking into account both confidentiality issues and user requirements.

1. Number of employees was recoded in 4 categories: $E10-49$, $E50-249$, $E250-999$, $E1000+$. The new variable was called $Size$.

2. Coherently with Istat dissemination policy, $Nace$ divisions 10-14 were aggregated together, as well as the divisions 15 and 16.

3. Age was recoded in 14-19, 20-29, 30-39, 40-49, 50-59, 60+.

4. *Length of service in the enterprise* was recoded in intervals of 4 years. The original variable was kept, too. The recoded variable, Len, was used in the perturbation stage.

5. Total gross annual earnings in the reference year ($AnnualEarnings$) was recoded in categories of 10000 euro. The original variable was kept, too. The recoded variable, $AnnEarn$, was used to determine the employees at risk of re-identification.

5 Enterprises anonymisation

5.1 Enterprises at risk of re-identification

The key variables in the enterprises disclosure scenario were $Nace, Nuts$ and $Size$. These key variables are all categorical. With respect to the adopted disclosure scenario, the enterprises at risk were those belonging to combinations of key variables with frequencies below an *a-priori* given threshold. The drawback of this approach is that it does not consider the survey characteristics. As the Italian 2002 SES was a *sampling* survey, both sample frequencies and population frequencies were considered. When a population combination of key variables contains many enterprises, it would be more difficult to identify a sampled enterprise, even if it is a sample unique. Consequently, a sampled enterprise was considered at risk when both population and sample frequencies were simultaneously inferior to 3.

Considering the key variables recoded as in the previous section, the sample of enterprise contains 641 non-empty combinations. Instead, the population of enterprises contains 910 such combinations. The table 1 presents the frequencies of sample and population rare cases. With respect to the adopted disclosure scenario, 62 enterprises were considered at risk of re-identification.

Rare case	Frequency
Sample uniques	70
Population uniques	78
Sample and population uniques	**28**
Sample doubles	50
Population doubles	54
Sample uniques and population doubles	**19**
Sample and population doubles	**15**

Table 1: Frequency of combinations *Nace* x *Nuts* x *Size* with 1 or 2 units.

Original		**After recoding**	
Size	**Frequency**	**Size**	**Frequency**
$E10-49$	4852	$E10-49$	4794
$E1000$	247	$E1000$	213
$E250-999$	1257	$E250-999$	1112
$E50-249$	2461	$E50-249$	2322
		$E10-49\_E50-249$	53
		$E10-49\_E50-249\_E250-999\_E1000$	13
		$E250-999\_E1000$	152
		$E50-249\_E250-999$	152
		$E50-249\_E250-999\_E1000$	6

Table 2: Frequencies of *Size* before and after the application of global recoding.

5.2 Protection

Protection of enterprises at risk of re-identification was achieved by means of a dedicated global recoding procedure, see [5]. Since *Nace* and *Nuts* are the most important from the user point of view, only *Size* was recoded.

As the enterprises at risk of re-identification were defined by means of both sample and population frequencies, the global recoding was applied controlling the population frequencies. Indeed, for each sample combination at risk, the corresponding population combination was identified. Two sample combinations were aggregated if the overall frequency of the population combinations was higher than the threshold. Obviously, this procedure can be applied only if the population frequencies are available.

As stated by the survey experts, it is preferable to aggregate a *Size* category with the category corresponding to immediately larger enterprises. When such recoding was not sufficient (the population frequency could still be smaller than the threshold), aggregation with the category corresponding to immediately smaller enterprises was investigated. If needed, aggregation of all *Size* categories for a given *Nace* x *Nuts* combination was performed. The sample frequencies before and after global recoding are presented in table 2.

6 Employees anonymisation

6.1 Employees at risk of re-identification

With respect to the adopted disclosure scenario, an employee could be identified only if information on enterprise is used. Considering the information content, only extreme earnings in large (and well-known) enterprises could be interesting for an intruder. For the anonymisation of the Italian 2002 SES microdata, enterprises with more than 250 employees were considered as large enterprises. In this sample there are 1504 enterprises with more than 250 employees (over a total of 8817 in the sample), while the population contains 3093 such enterprises (over a total of 193256). There are 40687 (over 81975) sampled employees belonging to a large enterprise. Due to the performed *SizeEnt* global recoding, further uncertainty is introduced. For example, an intruder wouldn't know whether an enterprise belonging to the category $E50-249\_E250-999$ belongs to $E50-249$ or $E250-999$.

Concerning the variable used for identification, several considerations hold. Firstly, as previously discussed, only a spontaneous identification scenario was adopted. This means that it was supposed that a possible intruder has no access to values of some variables possibly known only by a nosy colleague (for example, *paid hours* or *absence days*). Moreover, re-identification of an employee using his/her *number of absence days* or *number of working days* would be quite difficult. Therefore, an employee identification would be possible only by means of some previous "guesses" on ranges of earnings variables. Considering that the microdata file would be released for research purposes, the *Annual earnings* was deemed more adequate for spontaneous identification purposes. This was due to the "management" bonuses generally included in the annual earnings.

An intruder wouldn't be interested in differences of few thousands of euro between two values. Moreover, he wouldn't be even able to consider as different such two close values. Consequently, to determine the employees at risk of identification, *AnnualEarnings* was recoded into categories, resulting in a new variable *AnnEarn*, see section 4.

AnnualEarnings values exceeding a certain threshold T were considered as extremely high earnings, hence subject to spontaneous identification. T should be the same for all combinations of categorical key variables. An intruder would probably try to identify those employees with earnings greater than a certain $T_{intruder}$, an *a-priori* value imagined/thought by him. This $T_{intruder}$ is the limit above which the intruder would consider all earnings as high and interesting. As $T_{intruder}$ probably depends on both personal experience and knowledge of the studied economical phenomenon, its value cannot be determined by the data protector. Based on the observed data only, $T_{intruder}$ was estimated by T. Considering the skew probability distribution of $AnnualEarnings$, for the Italian 2002 SES, T was computed as the 99% quantile of the distribution. Obviously, it was assumed that employees with extremely low earnings were not at risk of re-identification.

Then, for each $Nace, Nuts, Size, Gender, Age, AnnEarn$ combination, the sampled employees with earnings greater than T were counted. If there was a single employee with such characteristics, it was considered at risk of re-identification.

Note that, in this way, the number of employees at risk of re-identification is not *a-priori* defined. In the Italian 2002 SES file, using this procedure, 317 employees were considered at risk.

6.2 Protection

Since the microdata would be disseminated for research purposes, perturbation of only records at risk of re-identification was deemed sufficient. Protection was applied taking into account also some probable usages of the microdata file. As stated by survey experts, regression models are frequently used to estimate the possible differences between different categories. For example, a researcher could be interested in estimating the difference on *AnnualEarnings* between two regions (estimating differences between regional politics). The employees protection was achieved by a perturbation method based on a regression model.

For the Italian 2002 SES, only parametric linear models were considered. The response variable was *AnnualEarnings*.

The explanatory variables choice is the most crucial step for the protection procedure. Firstly, the variables having a significant impact on the *AnnualEarnings* behaviour should be considered. For example, if it is believed that *Nace* significantly explains the *Annual Earnings* variation, *Nace* should be an explanatory variable. Secondly, the assumed model should simulate an user analysis. It was supposed that *AnnualEarnings* can be modelled as a linear combination of *Size, Gender, Age, ManPos, Occupation,FullTimePartTime, Len, MonthlyEarnings* and *PaidHoursMonth*, respectively. The model was used for each combination of *Nace* and *Nuts*.

Taking into account the already published totals, a constrained minimization problem was solved, see [6]. The main constraint was: for each combination of *Nace* and *Nuts*, the relative difference between the original and perturbed value should not be higher than 0.5%, as required by the survey experts. Moreover, each perturbed value was restricted to belong to the interval (0.5*OriginalValue, 2*OriginalValue). Additionally, each perturbed value was required to be higher than the threshold T. These last two constraints actually controlled the perturbation introduced in each record. The *AnnualEarnings* values of the employees at risk of re-identification, were replaced by the corresponding fitted values. Then, the *MonthlyEarnings* values of the records at risk were proportionally modified.

For the Italian 2002 SES microdata file, because of particular values of the factors involved in the regression model, the above methodology couldn't be applied for one combination of *Nace* and *Nuts*. For this particular combination, the *AnnualEarnings* values of the units at risk of re-identification were micro-aggregated, see [7].

By perturbing only the records at risk of re-identification, surely the *AnnualEarnings* and *MonthlyEarnings* weighted means would not be exactly preserved. But, the trend behaviour of these variables would be actually preserved. This behaviour was slightly lowered because only high earnings may be at risk of re-identification.

7 Information loss assessment

Only 0.39% of employees records were modified. The extreme earnings were generally decreased. Over the 317 units at risk of re-identification, the *AnnualEarnings* values were increased for 88 units. The summary statistics of the records at risk perturbation are indicated in table 3.

Min	Q1	Median	Mean	Q3	Max
-50	-1.70	4.75	8.69	19.07	100

Table 3: Percentages of the absolute relative perturbation of *AnnualEarnings*.

Since the *MonthlyEarnings* was proportionally modified with respect to *AnnualEarnings*, their consistency was automatically preserved. *Average gross hourly earnings in the representative month* was still computed as a ratio with respect to the perturbed *MonthlyEarnings*. Consequently, their consistency was maintained. Time related variables (number of worked hours, absence days, etc.) were not at all modified. The order relationships between *MonthlyEarnings*, *Special payment for shift work* and *Earnings related to overtime* were maintained, as well as their correlation coefficients, 0.03 and 0.14, respectively. A similar conclusion may be stated for the relationships between *AnnualEarnings*, *Total annual bonuses* and *Annual bonuses based on productivity*.

Generally, any microdata release is anticipated by the publication of a set of tables containing information on the survey variables. It is not always possible to exactly preserve the already published totals without a significant information loss with respect to other statistics. In such cases, at least an assessment of the difference one might obtain between the published totals and the ones computed using the microdata file is necessary. By definition, the applied protection method controls weighted totals for each combination of $Nace$ and $Nuts$.

For the Italian 2002 SES microdata, several tables were already published, involving mainly the following variables: $Nace$, $Nuts$, $Size$, $Gender$, Age, $FullTimePartTime$, $ManPos$, $Occupation$. The microdata file contains 26295 combinations of these variables. Only 263 (1%) of these combinations were modified by the applied perturbation. The absolute relative changes in the weighted means of *AnnualEarnings* were all inferior to 0.3. The mean of these relative perturbations was 0.06. Considering that the already published tables do not have the same maximum level of detail with respect to which these evaluations were performed, it was supposed that, at higher hierarchical levels, these relative perturbations on the means would tend to compensate one another. These were the main reasons for not applying any further adjustment.

8 Conclusions

A detailed analysis of possible disclosure scenarios and an accurate definition of related identifying variables are the key points of this anonymisation procedure.

Considering that the microdata file would be released for research purposes, the identification of units at risk is based on individual risk measures. This approach of a tailored definition of units at risk allows for dedicated protection methods that can save more information content.

Consideration of different scenarios is a key issue. For the Italian 2002 SES, for the enterprises an spontaneous identification scenario was adopted. Frequencies in the population of enterprises were considered for rare cases identification. Instead, for employees it was adopted a spontaneous identification scenario based on both an estimated threshold earning value and a rarity concept.

For the Italian 2002 SES microdata file, enterprise protection was achieved by aggregating categories of enterprise size. Consequently, two of the most important survey variables (*Nace* and *Nuts*) remained completely unchanged. A perturbation was applied only to records of employees considered at risk of re-identification, resulting in a significant reduction of the information loss. The perturbation method was derived from an analysis of user requirements. A set of constraints derived from a data utility criteria were used, too. The sampling weights were unchanged. An evaluation of the information loss was also performed.

The sampling weights, as well as the number of respondents, could also contribute to the enterprise identification. Other issues related to the information content analysis will be subject to further research.

References

[1] Hundepool, A. *et. al.* (2006) "Handbook on Statistical disclosure control", available at www.cenex.org.

[2] "Structure of Earnings Survey 2002 - Eurostat's arrangements for implementing the Council Regulation 530/1999 and the Commission Regulation 1916/2000", Eurostat, 6 April 2004.

[3] Deville, J.C., Sarndal, C.E. (1992) "Calibration estimators in survey sampling", *Journal of the American Statistical Association*, **87**, 376-382.

[4] ISTAT (2005) "Dipendenti, ore lavorate e retribuzioni nelle imprese dell'industria e dei servizi Anno 2002", Statistiche in breve, available at www.istat.it.

[5] Willenborg, L. & De Waal, T. (2001) "Elements of statistical disclosure control", Lecture Notes in Statistics, New York: Springer.

[6] Draper N.R., Smith H. (1998) *Applied regression analysis*, Wiley-Interscience.

[7] Defays, D., Anwar, M.N. (1998) "Masking Microdata Using Micro-Aggregation", *Journal of Official Statistics*, **14 (4)**, 449–461.

Microdata risk assessment in an NSI context

Jane Longhurst* and Paul Vickers*

*Office for National Statistics, Segensworth Road, Fareham, UK, Jane.Longhurst@ons.gov.uk/ Paul.Vickers@ons.gov.uk

1. Introduction

National Statistics Institutes (NSIs) collect and publish a wide range of economic and social data. Making confidentiality commitments and keeping these commitments is an important factor in maintaining trust between data providers and the NSI. Official statistics are generally released in the form of tables and microdata (individual level records) and the demand from policy makers and researchers for more detailed data and innovative ways to supply the data is continuing to increase. This increased demand increases the potential disclosure risks and this places greater pressure on NSIs to develop sophisticated methods to identify the level of risk posed by any release and to minimise this risk where it is deemed too great.

Traditionally NSIs have adopted microdata risk assessment procedures based on checklist criteria, ad-hoc rules and simple data-based summary measures. More recently there has been a recognised demand for quantitative disclosure risk measures in order to gain more objective criteria for release. This paper will focus on methods that estimate disclosure risk measures for microdata based on the population that make use of statistical models to estimate risk from the sample information in particular a probabilistic disclosure risk approach based on the Poisson Distribution and log-linear models (Elamir and Skinner (2006)). The scope is to investigate methodological aspects and practical issues related to the implementation of the method in particular for anonymised social survey microdata released under license.

In order to provide context for this work Section 2 gives an overview of the microdata release process at the Office for National Statistics (ONS) in the UK. Sections 3 and 4 describe in detail the research that has been carried out to investigate the feasibility of using the method based on the Poisson Distribution and log-linear models and to develop it further in an NSI context. The conclusion is drawn that the method could be used to assess the risk of a microdata file at the individual level but that further research is required to assess the risk at the household level.

2. Background to microdata release at the ONS

ONS has a long history of releasing microdata from its surveys. The release of all microdata from the ONS is approved by the Microdata Release Panel (MRP). Approval of release will depend on a number of factors described by Jackson and Longhurst (2005) and summarised in the following sections.

2.1 Legal and Policy Issues

The sixth United Nations Fundamental Principle of Official Statistics states:
Individual data collected by statistical agencies for statistical compilation, whether or not they refer to natural or legal persons, are to be strictly confidential and used exclusively for statistical purposes.

Access for research purposes is an exception to this default position. As such it needs to be a properly planned, managed, authorised and publicly acceptable activity. For research access to be lawful requires all these to be addressed, because compliance with the law is not just a matter of following legislation, but requires compliance with the laws for public administrative and generic information and common law at both the national and European Union (EU) level.

2.2 Risk Assessment

It is important that any disclosure risk assessment is carried within the legal and policy frameworks. For microdata, disclosure risk occurs when there is a possibility that an individual can be re-identified using information contained in the file, and on the basis of that, confidential information is obtained. Microdata is released only after taking out directly identifying variables, such as names, addresses, and identity numbers. However, other variables in the microdata can be used as indirect identifying variables such as gender, age, occupation, place of residence, country of birth, family structure. This combination of indirectly identifying variables is defined as a key and provides the basis for identification of a respondent and hence the disclosure risk.

In terms of disclosure risk assessment an intruder is someone who deliberately or inadvertently determines confidential information about a respondent from a dataset or attempts to do so. To assess the disclosure risk, one firsts need to make realistic assumptions about what an intruder might know about respondents and what information will be available to them to match against the microdata and potentially make an identification and disclosure. These assumptions are known as disclosure risk scenarios. ONS considers various disclosure risk scenarios in carrying out its disclosure risk assessment. The scenarios cover topics such as political attacks, private and public database cross match, journalists, local search and inquisitive neighbours, Elliot and Dale (1998).

These disclosure risk scenarios can be used to define the key variables within a microdata set. ONS has developed a checklist that can be used to focus on these variables. Responses to the questions on the checklist from data suppliers allow a disclosure risk assessment to be made. This assessment is generally made using subjective judgements and precedents. Sections 3 and 4 describe work to provide a quantitative risk assessment to support these subjective judgements.

2.3 Risk Management

The outcome of the risk assessment determines whether further measures need to be carried out or put in place to allow the data to be released. The MRP recognises all microdata releases are not risk free and aims to release microdata in a way that minimises this risk. It uses a combination of disclosure control methods (mostly recoding) and licence agreements or safe settings to control how researchers and policy makers use the data and present the results of any analysis.

3. Quantitative risk assessment

3.1 Introduction

The focus of this paper is the use of quantitative disclosure risk measures and how they can be used for risk assessment at the ONS for social survey microdata released under licence. Social surveys are typically samples of households where the

characteristics of the population are not fully known. When the population is unknown there is a need to rely on models or heuristics in order to quantify the disclosure risk.

Previous work (Shlomo and Barton (2006)) has been undertaken to compare the performance of three different methods; the Special Uniques Detection Algorithm (SUDA) (Elliot et al (2005)), the probabilistic disclosure risk method based on the Poisson Distribution and log-linear models (Elamir and Skinner (2006)) and the method based on the Negative Binomial Distribution (Polettini and Seri (2003)) which is embedded in the Computational Aspects of Statistical Confidentiality (CASC) project software Mu-Argus (CASC (2004)). An evaluation of these three methods has been carried out for a range of different sample sizes but limited key sizes, mostly 6 variables. For the examples considered the results show that the Poisson model with log-linear modelling performs the best but is more complex than the other methods and requires more computing time and intervention in a model search algorithm.

The scope of this paper is to investigate the practical issues related to the implementation of the probabilistic disclosure risk method based on the Poisson Distribution and log-linear models. In particular addressing the performance and feasibility of this method for larger key sizes. The method can be used to estimate risk for individual records and globally for a whole file. The focus here is on file level measures of risk that can be used within a microdata release procedure. Consideration will also be given to the feasibility of estimating disclosure risk for hierarchical microdata sets, e.g. individuals within households.

To introduce the basic measure of identification risk, suppose a key has K cells and each cell $k = 1,\ldots K$ is the cross product of the categories of the identifying variables. Let F_k and f_k be the population and sample counts in cell k respectively. The aim of quantitative disclosure risk assessment methods for microdata is to estimate an individual per-record disclosure risk measure that is formulated as $\frac{1}{F_k}$, that is the probability that a record in the microdata and a record in the population having the same values of identifying key variables will be correctly matched. Since the uniques in the population, $F_k = 1$, are the dominant factor in the disclosure risk measures, this measure is focused on the case when $f_k = 1$, i.e. for sample unique cells. This leads to the following record-level risk measure: $r_k = E\left[1/F_k \middle| f_k = 1\right]$

3.2 The Poisson Model and Log-Linear Modelling

As described in Bethlehem et al (1990) consider models where the F_k are realisations of independent Poisson random variables with means λ_k $(k = 1,\ldots\ldots K)$, $F_k \sim P(\lambda_k)$. Assume that the sample is drawn by Bernoulli sampling with common inclusion probability π so that $f_k \sim P(\mu_k)$ where $\mu_k = \pi\lambda_k$. The record level measure can be expressed as $r_k = E\left[1/F_k \middle| f_k = 1\right] = \frac{1}{\lambda_k(1-\pi)}(1 - e^{-\lambda_k(1-\pi)}) = h(\lambda_k)$

This measure depends on unknown λ_k. In order to 'borrow strength' between cells suppose the μ_k are related via the log linear model $\log \mu_k = x_k' \beta$ where x_k is a design vector denoting the main effects and interactions of the model for the key variables. Using standard procedures, such as iterative proportional fitting, this model is fitted to the sample data to obtain the maximum likelihood estimates for the vector β and the fitted values $\hat{\mu}_k = \exp(x_k' \hat{\beta})$ are calculated. The estimate for $\hat{\lambda}_k$ is then substituted in the formula for r_k which can be aggregated across sample uniques to obtain the following file level measure estimate:

$$\hat{\tau} = \sum_{SU} \hat{r}_k = \sum_{SU} \hat{E}\left[1/F_k \middle| f_k = 1\right] = \sum_k I(f_k = 1) h(\hat{\lambda}_k)$$

the expected number of correct matches for sample uniques, where $SU = \{k : f_k = 1\}$. Such an approach has been described in Skinner and Holmes (1998) and Elamir and Skinner (2006).

A key issue with this method is that inference may be sensitive to the adequacy of the specification of the log linear model. Skinner and Shlomo (2006) develop criteria for assessing whether the vector x_k may be expected to lead to accurate estimated risk measures. Standard approaches such as Pearson or likelihood-ratio tests or Akaike's Information Criterion are discounted since they are not appropriate for the large and sparse tables considered in this application. In this analysis the minimum error test statistic $\frac{\hat{B}}{\sqrt{v}}$ is used, as defined in Skinner and Shlomo (2006). It has an approximate standard normal distribution under the hypothesis that the expected value of $\hat{B}$ is zero. A positive value under 1.96 accepts the fit of the log linear model for obtaining a good disclosure risk measure.

3.3 Method

The data used for this analysis was taken from the 2001 UK Census. Following the method used by Shlomo and Barton (2006) five different samples were drawn from the Census data for England and Wales covering 52 million people within 22 million private households (communal establishments were excluded). The different samples simulate typical sample sizes for different ONS social surveys and the samples were drawn using a similar design as standard social surveys. Clustered samples are not simulated since it is assumed that this aspect of sample design would not affect the risk measurements. Some ONS social surveys sample disproportionately across geographies, this is not simulated here. Table 1 describes the five different samples.

Sample	Sampling Fraction	Number of households in the sample	Number of persons in the sample
A	0.000323	7,000	16,651
B	0.001062	23,000	54,560
C	0.002308	50,000	119,618
D	0.006924	150,000	357,888
E	0.010155	220,000	524,399

Table 1: Samples used in the analysis

The Poisson model with log-linear modelling is used to estimate $\hat{\tau}$ for the different sample sizes and using different key sizes. Since the samples have been drawn from Census data, the true risk measure, τ, can be calculated and used to evaluate the performance of the method.

Following Skinner and Shlomo (2006) a forward search algorithm is used, starting from simpler models and adding interaction terms until the specification is judged to be adequate. As in Shlomo and Barton (2006) the analysis is based on the private database cross match scenario, Elliot and Dale (1998). As a first stage the analysis is restricted to 6 variables with categories typically used in ONS social survey microdata releases: region (11), age (96), sex (2), number of residents (7), marital status (6), number of cars (5). An 8 variable key is also considered, this covers the 6 variable key plus number of earners (5) and number of dependent children (5). The assumption is made that there are no discrepancies in the values of the key variables between the microdata and the intruder's data.

3.4 Results

Table 2 displays the results (final model, estimated and true risk and the minimum error test statistic) for the five samples for the 6 variable key where K = 443,520, replicating the results in Shlomo and Barton (2006). A detailed breakdown of the model search for Sample C is included in the Appendix as an example. For all samples the estimated risk is close to the true risk, in all cases within 10%. The breakdown of the model search shows that as outlined in Skinner and Shlomo (2006) large negative values for the test statistic (overfitting) lead to an underestimate of the true risk and large positive values (underfitting) lead to an overestimate. When the test statistic for a model is small this can lead to either over or under estimation. Each model takes a few minutes to run, some intervention is required through the model search algorithm to select the interactions for inclusion and so models with more interactions take longer to run. As the sample size increases, the global risk estimate increases and the model becomes more complex.

Sample	**Final Model**	**True Risk (τ)**	**Estimated Risk ($\hat{\tau}$)**	**Minimum error test statistic ($\frac{\hat{B}}{\sqrt{v}}$)**
A	All 2-way interactions	83.0	81.4	0.14
B	All 2-way interactions + {age, residents, mstatus}	220.3	204.5	1.13
C	All 2-way interactions + {age, mstatus, cars} +{age, residents, mstatus}	446.6	410.6	1.65
D	All 2-way interactions + {age, mstatus, cars} +{age, residents, mstatus} +{region, age, residents}	1193.8	1182.4	1.74
E	All 2-way interactions + {age, residents, mstatus} +{age, residents, cars} +{age, mstatus, cars} +{region,	1701.8	1569.7	1.03

	age, residents} +{age, sex, mstatus} +{region, residents, mstatus}			

Table 2: Results for 6 variable key

The modelling exercise was repeated for the 8 variable key where K = 11,088,000 but problems were encountered related to computing times. Even running just the all 2-way interaction model takes approximately 40 hours[1]. The different samples take approximately the same time, run time is dependent on key size and number of variables, although as shown in Table 2 sample size tends to affect model complexity which is related to the time taken to complete the model search. It can be calculated that the next stage in the model search, i.e. testing each 3-way interaction term for inclusion in the model would take about 2000hrs or over 33 days! This would not be practical. The next section of the paper describes an approach to improving the time taken to fit the log-linear models.

3.4.1 Partitioning

The approach considered here to reduce the time taken for the model search and model fitting procedures involves partitioning the data into smaller datasets. Models are fitted to each of these subsamples to obtain risk estimates which are then aggregated over all the subsamples to obtain to final risk estimate for the whole file. Since this approach will reduce the size of the sample and key size (if partitioning is based on a key variable) it should result in simpler models which will be easier and faster to fit. Two issues are considered; which variables should be used to partition the sample and how many subsamples should the sample be divided into.

Partitioning by a particular key variable assumes an underlying interaction with that variable in the model in each of the sub-tables. The Cramer's V statistic was used to measure the strength of association between pairs of the key variables in Sample C. The strongest association is between age and marital status (0.434). Sample C was partitioned into different subsamples based on different key variables and the estimated risk measures for the 6-variable key were compared, Table 3.

Partition variable	**Subsamples**	**Average size of partition**	**Key size (K)**	**Estimated Risk ($\hat{\tau}$)**
Age	4 (24 years in each)	29905	110,880	457.1
Region	4 (2 or 3 regions in each)	29905	120,960 or 80,640	503.6
Marital status	4 (1 or 2 marital status categories in each)	29905	147,840 or 73,920	404.3
Age	2 (48 years in each)	59809	221,760	480.2
Sex	2 (male/female)	59809	221,760	501.7

Table 3: Results for partitioning Sample C, 6-variable key, $\tau = 446.6$

The results show that different models are selected for different subsamples, as expected these tend to be simpler models (since the sample size of them is smaller). For some subsamples the true risk is overestimated for others it is under estimated. As

[1] Run times are obviously very dependent on the PC used. All results were obtained on a standard ONS PC; 2.8 GIG processor, 512 MEG memory, Windows XP, SAS Version 8.2.

expected partitioning by age produces the best risk estimate when splitting into 4 or 2 subsamples. The results are robust, the risk estimates for each subsample are good and the overall risk estimates for partitioning by age are more accurate than the estimate with no partitioning (see Table 2).

Now consider the best number of subsamples to implement when partitioning by age. In general as sample sizes get smaller models become simpler. Ideally one would want to find the case when the all 2-way interaction model or the independence model fits. These cases are quicker and easier to run since they require no stepwise procedure and therefore no user intervention in the modelling process. This approach is tested using the 8 variable key across all samples. The all 2-way interaction model is fitted to all subsamples. Different sized subsamples were selected depending on the overall sample size. Table 4 summarises the partitioning results across all samples, displaying the partition that produced the best risk estimate.

Sample	No. of sub samples	Average size of subsample	Key size (K)	True Risk (τ)	Estimated Risk ($\hat{\tau}$)	Range for $\frac{\hat{B}_2}{\sqrt{v}}$	Time taken
A	4 age bands	4163	2,772,000	406.0	414.0	[-0.22,17.5]	16 hrs
B	32 age bands	1705	346,500	1284.6	1231.9	[-0.7,7.1]	5 hrs 14 mins
C	32 age bands	3738	346,500	2693.4	2918.9	[-0.9,8.2]	5 hrs
D	32 age bands	11,184	346,500	7703.5	9052.3	[-1.0,22.1]	5 hrs 25 mins
E	32 age bands	16,388	346,500	11111.6	13224.5	[-0.2,20.4]	5 hrs 50 mins

Table 4: Partitioning results, 8-variable key, all 2-way interactions models

The results show that for all samples (other than Sample A) the best results are obtained using 32 partitions. In general as the size of the partition decreases the fit of the all 2-way interaction model improves and hence the accuracy of the risk estimate increases. The results for Sample A and B have shown that if sample sizes become too small (average around 2000 individuals) the all 2-way interaction model starts to overfit the data and leads to an underestimation of the final risk. Overall as the sample sizes increase the error in the risk estimate increases. Ideally one would implement more than 32 partitions for the larger samples (D and E) in order to improve the fit of the all 2-way interaction model and increase the accuracy of the risk estimate. The results suggest that an ideal subsample size for this 8-variable key is 2000-4000 individuals.

4. Measuring risk for hierarchical files

4.1 Introduction

Many social surveys are hierarchical in nature, allowing groups of individuals to be recognised within the file, the most common case being households. If it is possible to

link individuals within the released microdata file then it is important to take into account this dependence when measuring disclosure risk. Within the software Mu-Argus (CENEX (2006)) household risk is defined as the probability that at least one individual in the household is identified and is computed from the individual risk of the household members. An alternative scenario considered here would be that the intruder matches directly at the household level, where a single record represents a household characterised by the identifying variables of all household members. This approach was adopted for the disclosure risk assessment for the Household Sample of Anonymised Records (SARs) (CCSR (2005)) from the 2001 UK Census. Different approaches relate to the assumed availability of hierarchical external databases. Initial observations from the CAPRI (Confidentiality and Privacy Group) Data Monitoring Service (CCSR (2004)) indicate that some hierarchical household information is available in many datasets (particularly when more than one adult lives in the household).

4.2 Method

When measuring risk at the household level the key variables will need to be modified to incorporate information on all the individuals within the household as well as some household level variables, Elliot (2005). Here analysis is restricted to a key with basic household information for all members of the household, e.g. age, sex and preliminary results are outlined for 2 person household only.

Before fitting the model the microdata file needs to be modified. The file is split by household size and one record is created for each household that contains information on all members of the household. The key is then constructed using variables on each member of the household and household level variables. Care needs to be taken in constructing this key in terms of ordering the members within the household. It is possible that two households could be identical, but not recognised as such if they are sorted in different orders.

4.3 Results

Table 5 details the results of the stepwise modelling procedure for all 2 person households in Sample A (of which there are 2414) using a 4-variable key constructed using age and sex of both household members, K = 36,864. The household members are ordered within the key by age and then sex. The minimum error test statistic for the all 2-way interaction model is negative, providing evidence of overfitting and as expected the true risk is underestimated. The forward stepwise procedure is used to add 2-way interaction terms to the independence model. As 2-way terms are added to the model the test statistic decreases (indicating a better fit) but the risk estimate moves further away from the truth. No model can be found with a test statistic that is positive and less than 1.96. A similar pattern of results is observed for sample B. For sample C, D and E the all 2-way interaction model produces the best estimate of risk but this is not reflected in the test statistic. The results show that the modelling procedure is not as effective in this case as it was for modelling at the individual level.

Constructing keys at the household level and in particular the ordering of household members will introduce dependencies into the variables in the key which could affect the validity of fitting a log-linear model to the data. Consider two person households and a simple key of age and sex for both household members where the household

members are ordered by age and then sex. The age of the second household member will always be less than or equal to the age of the first household member.
The format of the household level key introduces structural zeros into the key. In order to reduce these problems the proposal is made to order the key by sex and then age rather than age and then sex. This will not totally overcome the problem interdependencies between the two age variables will still exist but there should be less structural zeros forced into the key by the ordering procedure.

Another factor that may be affecting the results seen here is the number of variables in the key. A 4-variable key has been investigated whereas previous analysis investigated 6 and 8 variable keys. Further analysis should be carried out with larger keys to investigate whether including more variables in the key improves the performance of the models. In particular household type should be considered as a key variable and potentially used as a partitioning variable.

Model	**Estimated Risk ($\hat{\tau}$)**	**Minimum error test statistic ($\frac{\hat{B}}{\sqrt{v}}$)**
Independence	9.39	21.5
All 2-way interaction	1.57	-1.2
Independence + {sex, sex2}	11.75	8.1
Independence + {sex, sex2} + {age, sex2}	11.5	4.6
Independence + {sex, sex2} + {age, sex2} + {age, sex}	12.2	3.4
Independence + {sex, sex2} + {age, sex2} + {age, sex} + {age2, sex2}	1.47	-1.7

Table 5: Results for Sample A, 2 person households, 4-variable key, $\tau = 4.8$

5. Conclusions

There is a strong, widespread and increasing demand for National Statistics Institutes (NSIs) to release microdata files. These data sets are a vital resource for key research and thus it is important to make the microdata as detailed as possible while protecting the confidentiality of the information provided by the respondents. This paper has investigated issues concerned with the practical implementation of a method for quantitatively assessing disclosure risk for microdata based on the Poisson Distribution and log-linear models.

The results have shown that it is feasible to assess the risk of a microdata file at the individual level for a 6 and 8-variable key and that the results are robust. Quantitative file level measures of risk can be used within the microdata release approval process alongside current (more subjective) measures such as the checklist. Splitting the risk assessment by subpopulations reduces computational demands for the 8-variable key. The results show that final risk estimates are more accurate when partitions are determined by a key variable that is most correlated with the other key variables, here age. In general as sample sizes get smaller the best fitting log-linear models become simpler. The recommendation is made that the all 2-way interaction model is taken for each subsample. This will necessarily impact on the quality of the final risk estimate

but will avoid lengthy model search procedures. The results suggest that an ideal subsample size for the 8-variable key is 2000-4000 individuals.

Assessing disclosure risk for larger keys is possible with partitioning but the time taken to carry out the modelling is anticipated to take days rather than hours on a standard ONS PC. Future work should consider the likely availability of external databases with more than 8 or 10 identifying variables in order to gauge whether risk assessments are required for these larger keys. In addition the availability of hierarchical external databases needs to inform the hierarchical disclosure risk scenarios.

The preliminary results outlined here have indicated that as currently implemented the modelling procedure is not as effective for estimating risk at the household level as it is at the individual level. Further analysis is required to overcome the interdependencies in the key variables introduced by the hierarchical structure. Further analysis is also required to investigate larger keys and alternative scenarios.

6. Acknowledgements
A special thank you to Natalie Shlomo and Jeremy Barton for their help in preparing the samples and initial SAS programmes used in the empirical work and to Natalie Shlomo and Chris Skinner for their general support.

References

Bethlehem, J., Keller, W., and Pannekoek, J. (1990) *'Disclosure Control of Micro-data'*. Journal of the American Statistical Association 85 No. 49 (1990) 38-45
CASC website (2004), *Computational Aspects of Statistical Confidentiality*, www.neon.vb.cbs.nl/casc/default.htm
CCSR. (2004) *'ONS Disclosure control report – Special licence Household SAR'*. www.ccsr.ac.uk/sars/guide/2001/disclosure.html
CCSR. (2005) '*A scoping study for the establishment of a data monitoring service'*. www.ccsr.ac.uk/research/datamonitor.htm
CENEX website (2006), *Centre of Excellence for Statistical Disclosure Control,* www.neon.vb.cbs.nl/cenex/
Elliot, M. J., and Dale, A. (1998) *'Disclosure Risk for Microdata'*, Report to the European Union ESP/204 62/DG III.
Elliot, M. J. and Dale, A. (1999) *'Scenarios of Attack: The data intruder's perspective on statistical disclosure risk'*. Invited paper for special edition of Netherlands Official Statistics.
Elamir, E. and Skinner, C. (2006) *'Record-Level Measures of Disclosure Risk for Survey Micro-data'*. Journal of Official Statistics 22 525-539(2006)
Elliot, M. (2005) *'Assessment of Disclosure Risk for Hierarchical Microdata Files',* ONS Report, Confidentiality and Privacy Group, Cathie March Centre, University of Manchester, 2005.
Elliot, M., Manning, A., Mayes, K., Gurd, J. and Bane, M. (2005) *'SUDA: A Program for Detecting Special Uniques'*. Monographs of Official Statistics, UNECE and Eurostat Work Session on Statistical Data Confidentiality Geneva (November 2005).
Jackson, P. and Longhurst, J. (2005*) 'Providing access to data and making microdata safe, experiences of the ONS',* Monographs of Official Statistics, UNECE and Eurostat Work Session on Statistical Data Confidentiality Geneva..

Polettini, S and Seri, G. (2003) *'Guidelines for the protection of social micro-data using individual risk methodology - Application within mu-argus version 3.2'*, CASC Project Deliverable No. 1.2-D3.
Shlomo, N. and Barton, J. (2006) *'Comparison of Methods for Estimating Disclosure Risk Measures for Microdata at the UK Office for National Statistics'*, Privacy in Statistical Databases, CENEX-SDC Project International Conference, PSD 2005 proceedings.
Skinner, C. J. and Holmes, D. (1998) *'Estimating the Re-identification Risk Per Record in Micro-data.'* Journal of Official Statistics 14 No. 4, 361-372.
Skinner, C. J. and Shlomo, N. (2006) *'Assessing Identification Risk in Survey Micro-data Using Log-Linear Models'*. www.eprints.soton.ac.uk/41842

Appendix

Model	**Estimated Risk** ($\hat{\tau}$)	**Minimum error test statistic** ($\frac{\hat{B}}{\sqrt{v}}$)
All 2-way interaction	578.4	8.0
All 3-way interaction	271.8	-2.4
All 2-way interaction + {age, mstatus, cars}	516.3	4.1
All 2-way interaction + {age, mstatus, cars} + {age, residents, mstatus}	410.6	1.65

Table A1: Detail of model search for 6-variable key, sample C, τ = 446.6

Using Targeted Perturbation of Microdata to Protect Against Intelligent Linkage

Mark Elliot[1]

[1] Centre for Census and Survey Research, University of Manchester, UK, M13 9PL

Keywords. Record Linkage, Disclosure Risk, Microdata

1 Introduction

This paper describes linkage experiments using the 2001 individual Sample of Anonymised Records (the SARs see http://www.ccsr.ac.uk/sars/) from the UK census to the microdata output from the UK Labour Force Survey (LFS) for spring 2001.

The objective of the study was to assess the impact of the Statistical Disclosure Control (SDC) methods used on the 2001 SARs on the ability to link an external dataset and SARs. The Labour Force Survey was selected as the external file because it was of sufficient size to produce a large enough overlap with the SARs and was collected around the census date.

The project follows the tradition of other such studies with official data; e.g. Muller, et al (1992), Elliot and Dale (1998). However, the study here elaborates on that earlier work by examining the impact of a targeted disclosure control technique on the ability of an intruder to attack a dataset by focusing on the high risk records.

2 Method

Data

Three datasets were used in this study:

1. The spring 2001 quarter of the standard Labour Force Survey (LFS).
2. The standard release version of the 2001 individual level SAR (post-SDC SAR).
3. The pre-SDC version of the 2001 individual level SAR (pre-SDC SAR).

The 2001 SARS were subject to extensive statistical disclosure risk assessment and targeted control methods. The risk assessment was carried through collaboration

between Manchester University and ONS staff using the software system SUDA 1.4; Elliot and Manning (2004) The disclosure control was a mixture of global recoding and local suppression and reimputation based on a variant of the PRAM (post randomisation) method. This involves using a replacement probability matrix for each value for each variable. The result was a file which was regarded as having been sufficiently protected against deliberate disclosure attempts; see Bycroft and Merrett (2005). A version of the SARS without the SDC applied is available in safe settings within ONS. This is in effect the pre-SDC file used here.

2.1 Procedure

The procedure consisted of seven steps:

1) The matching variables were selected and then the codings of these variables on the different datasets were harmonised.

2) The matching was conducted.

3) The SUDA Software was run over the SARs to obtain DIS-SUDA scores for the matches.[1]

4) All the unique matches (one-to-one) were sent to ONS.

5) The matches were verified by ONS Census and LFS divisions.

6) The matches were returned with an indicator placed on the match file indicating whether the match was true or false.

7) The proportion of correct matches was generated under several different assumptions.

2.2 Variable Selection and Mapping

[1] SUDA (Special Uniques Delectation Algorithm) is a method and a software system for detecting risky records within microdata. It does by identifying all minimal sample uniques (sample uniques with no unique supersets) within each record up a user specified size. By using a heuristic to combine all the information for each record with a well established file level probability measure, a per record risk measure is obtained. This enables the records to be sorted in order of risk. The use of this software here enables the simulation of a more sophisticated intruder who is able to take account of the structure of the target file in order to improve his/her confidence in the matches s/he finds.

The first step in the matching was to harmonise the data. This involved selecting the variables that were to be used as the match key and then recoding them on one or both datasets so that the codings were consistent.

Frequent practice when considering disclosure risk assessment is to use standard scenarios for key selection; see Elliot (2006). However, the variables available on the intersection between the LFS and SARs did not correspond to any of the scenarios in Elliot (2006). Since we were mimicking an attack from an *external database* that Scenario was used as a base. However, three of the standard variables within this scenario: "Number of cars in the household", "distance of travel to work" and "workplace" could not be included. To make the scale of the attack represent what was possible under this scenario, ethnic group and country of birth were added to the key, effectively creating a blend of two scenarios. To simplify the process only the data for England and Wales and persons in households were included. Non-resident students were excluded.

This created a key with the following nine key variables:

Region
Age
Sex
Marital Status
Number of residents in household
Tenure
Primary economic status
Ethnic group
Country of birth

2.2.1 Harmonising the key variables

As with most matching studies there then followed a complicated process of variable harmonisation; this required manipulation of all three datasets and resulted in the following
Age (95 categories for the pre SDC SARS-LFS match, 44 categories for the post-SDC SARs-LFS match)
Sex (2 Categories)
Marital Status (5 Categories)
Region of residence (11 Categories)
Number of Residents (7 categories)
Primary economic status (9 categories)
Country of birth (14 Categories)
Ethnic group (15 categories)

Tenure (5 categories)

2.3 Matching

Once the variable harmonisation had been conducted, the matching process was relatively straightforward. In principle, we could have used fuzzy matching methods – to allow for data divergence – however, the number of direct one to one matches was very large on both files and therefore this was deemed to cause an unnecessary administrative burden at the match verification stage. Therefore, a simple combine and sort algorithm was used for the matching.

In all there were 6085 one to one matches between the pre-SDC SAR and the LFS and 3130 one to one matches between the released SAR and the LFS. The SAR id and LFS identifier fields for the matched records were combined into a file and these were sent to ONS for verification.

2.3.1 Match verification

A problem occurred at the verification stage. For a significant number of matches there was no address linkable to the LFS identifying variables. This affected 1602 matches (26.32%) against the pre-SDC SARS file and 895 matches (28.95%) against the post-SDC SARS file.

There was also some interaction between some of the matching variables and whether a name and address was found. Ethnic group in particular was related to whether a name and address could be found with 22.6% of white people and 34.2% of non-white people on the post-SDC file not being found ($V=0.126$, $p<0.0005$). On the pre-SDC match file the figures were 23.6% and 33.5% respectively ($V=0.101$, $p<0.0005$).

There was some apparent interaction with the SUDA score indicating that there was a higher probability of a name and address not being found if a record had a high DIS-SUDA score(>0.3). However, this was only slight and not statistically significant ($V=0.005$, $p=0.773$).

Although there is some cause for concern about the number of cases for which it was not possible to verify whether a match was true and the interaction between that and some of the match key variables, it was decided that the non significant association with the SUDA score meant that it was acceptable to proceed with the analysis discounting the matches for which there was no name and address found.

3 Results

The headline results are given in tables 1 and 2. Overall there were 3130 matches between the post-SDC SARS file and the LFS and 6085 between the pre-SDC SARS file and the LFS. After discounting the matches for which no name and address was found these figures dropped to 2235 and 4483 respectively. The difference between these two totals bears discussion in itself. In effect the disclosure control has reduced the number of matches (true and false) by approximately half. This result is probably mostly due to the global recoding particularly of the age variable; as we might expect perturbation to produce new false matches as well as disguising true existing ones.

In terms of the absolute number of matches the SDC process can be said to have had a significant impact with 123 correct matches pre-SDC dropping to 51 post-SDC, nearly a 60% reduction.

		Frequency	Percent	Valid Percent
Valid	False match between SAR and LFS	2184	69.78	97.72
	Correct Match between SAR and LFS	51	1.63	2.28
	Total	2235	71.41	100.00
Missing	No name and address from LFS file	895	28.59	
Total		3130	100.00	

Table 1: Match Indicator for matches between post-SDC SAR and the LFS

		Frequency	Percent	Valid Percent
Valid	False match between SAR and LFS	4360	71.65	97.26
	Correct Match between SAR and LFS	123	2.02	2.74
	Total	4483	73.67	100.00
Missing	No name and address from LFS file	1602	26.33	
Total		6085	100.00	

Table 2: Match Indicator for matches between pre-SDC SAR and the LFS

In terms of the match rates we can observe that the raw match rates on both pairs of files are low: 2.74% in the matching with the pre-SDC SARs and 2.28% on the post-SDC SARs. This indicates that the SDC has had some effect in the gross match rates. However, even the match rate for the pre-SDC SAR is low and these are the sorts of

figures that should be of little concern to ONS. An intruder faced with that sort of success probability would be in effect swamped by false matches.

However, this is not the whole picture. In the second stage of the experiment SUDA was run over the SARs file, this represents how an intelligent intruder would use knowledge about the structure of the attack file to focus on the higher probability matches. Table 3 shows the frequency of true and false matches for various 7 bands of DIS-SUDA scores for the pre-SDC match file. There appears to be increasing risk as the DIS-SUDA score increases. The message here is more clearly seen in table 4, where we consider an intruder who uses various confidence thresholds. Using the pre-SDC match file the sophisticated intruder would be able to achieve matching rates of up to 22% using medium threshold matches. These are rates where an NSI might start to be concerned.

DIS-SUDA Band	False matches	Correct matches	% correct	Total
0->0.1	3900	81	2.03	3981
0.1->0.2	218	8	3.54	226
0.2->0.3	111	11	9.02	122
0.3->0.4	77	9	10.47	86
0.4->0.5	33	8	19.51	41
0.5->0.6	10	3	23.08	13
0.6->0.7	1	1	50.00	2
>0.7	10	2	16.67	12
Total	4360	123	2.74	4483

Table 3: Correct and incorrect matches by bands of DIS-SUDA scores for the pre-SDC SAR.

DIS-SUDA Threshold	False Matches	Correct matches	%correct
>0	4360	123	2.7
>0.1	460	42	8.4
>0.2	242	34	12.3
>0.3	131	23	14.9
>0.4	54	14	20.6
>0.5	21	6	22.2
>0.6	11	3	21.4
>0.7	10	2	16.7

Table 4: Match rates achieved by a hypothetical intruder using various thresholds of confidence as measured by the DIS-SUDA score. For the pre-SDC SAR

So what is the impact of the SDC process on these results? Table 5 shows the frequency of true and false matches for various bands of DIS-SUDA scores for the post-SDC match file. Again the match rates do seem to increase as the DIS SUDA rates get higher but as Table 6 shows the impact is nowhere near as marked as with the pre-SDC file. This is undoubtedly mostly the effect of the perturbation, which specifically targeted these high risk records (although not using the particular combination of key variables that were used in this experiment).

DIS-SUDA Band	False matches	Correct matches	% Correct	Total
0->0.1	1577	28	1.74	1605
0.1->0.2	394	8	1.99	402
0.2->0.3	85	2	2.30	87
0.3->0.4	52	8	13.33	60
0.4->0.5	45	2	4.26	47
0.5->0.6	14	2	12.50	16
0.6->0.7	6	0	0.00	6
>0.7	11	1	8.33	12
Total	2184	51	2.28	2235

Table 5: Correct and incorrect matches by bands of DIS-SUDA scores for the post-SDC SAR.

DIS-SUDA Threshold	False Matches	Correct matches	%correct
>0	2184	51	2.3
>0.1	607	23	3.7
>0.2	213	15	6.6
>0.3	128	13	9.2
>0.4	76	5	6.2
>0.5	31	3	8.8
>0.6	17	1	5.6
>0.7	11	1	8.3

Table 6: Match rates achieved by a hypothetical intruder using various thresholds of confidence as measured by the DIS-SUDA score.

It is also noteworthy that the correct matching rate becomes non-monotonic with respect of the DIS-SUDA score above the threshold of 0.2, this again is the result of the targeting of the high risk records and in essence there is no value in a hypothetical intruder - in this example – operating at a confidence threshold greater than 0.2. To put it another way, the SDC that was employed on the 2001 SARs appears to seriously undermine intrusion attempts based on the fishing method of attack.

So from the point of view of the SARs the results look reassuring. Whether the residual correct matching rates (2.3% overall and less than 10% on the broad fishing type of attack) fall within the boundary of "undue effort" criterion is something for ONS to decide.

4 Discussion

There are various caveats that need to be placed on the interpretation of these results. The first concerns data divergence. The two datasets concerned here are both produced by ONS. Therefore, in general we would expect that there would be lower rates of data divergence, than either would have with a non-ONS dataset. This will tend to inflate the match rates compared to what an intruder would achieve. Against this a very particular form of data divergence would have been present in this case, arising from the data collection. The SARs were generated from the census which was collected on one day in April 2001, whereas the LFS file we used was collected from March to May 2001. This has an impact on the data divergence of the age variable in particular. Someone whose birthday fell between the census date and the collection of their LFS data would have a different age recorded on the two datasets. This form of data divergence affects the pre-SDC SARs (which is coded in single years) more than the post-SDC SARs (which is not). The impact of this on match rates is difficult to estimate exactly but it would be reasonable to assume that approximately one eighth of records would be affected.

A second point to be born in mind in interpreting these results is that we have simulated only one sort of attack, a database cross match of common variables. A sophisticated, determined intruder could use this initial match to pursue further information about potential matches.

Given that, in a real data intrusion, identification information would be present on the attack file, the intruder could go through the list of potential matches and try to extend the match keys for those cases, by attempting to gather further information about the individuals (which would confirm or deny the matches). When we are talking about 6000 possible matches that would obviously be quite onerous but by

focusing on the matches with high SUDA scores this might enable the intruder to differentiate the false from true matches[2] quite effectively. This again emphasises the importance of the perturbation in masking the high risk matches.

The third caveat is that an intruder is unlikely to have access to a dataset which has exactly the same coverage to that of the SARs. A different rate of coverage would change the absolute matching rates and would also affect the information available from the data structure of the attack file and therefore the confidence of an intruder in any given match. An attack file with a significantly larger coverage than the SARs 3% would represent a much increased risk.

A final point is that the SUDA algorithm is only one method for assessing record level disclosure risk. Evidence suggests that others such as that of Skinner and Holmes(1998) maybe superior although Shlomo and Barton (2006) found that the SUDA metric produced similar correlations to that of Skinner and Holmes' model based approach, so this is not likely to have affected the overall pattern of results greatly.

Overall then, by comparing the results for the pre and post-SDC SARs files it has been possible to estimate the effect of the SDC process. Generally, the recoding appears to have substantially reduced the total number of matches and the perturbation has effectively masked the higher risk records on the SARS.

There are three processes which would cause variation in the results obtained here; (i) differences in the data divergence rates between the files used in the simulation and that used in an actual intrusion attempt (ii) An intruder using secondary intrusion techniques to confirm matches (or not) (iii) an intruder using an attack file with different coverage than the LFS. Nevertheless, the headline result of this study is that the targeted disclosure control used with the 2001 SARS appears to be very effective in protecting against an attack based on identifying high risk records.

References

Bycroft. C. and Merrett, K. (2005) 'Experience of using post randomisation method at the Office for National Statistics', *Monographs of Official Statistics, UNECE and Eurostat Work Session on Statistical Data Confidentiality Geneva*.

[2] Of course, extending the key would also increase the data divergence, so the intruder would need to be operating at a high degree of sophistication.

Elliot, M. J. and Manning A. (2004) A Special Uniques Analysis on the 2001 Samples of Anonymised Records. Report to the Office for National Statistics, with Manning A.

Elliot, M.J. (2006) Scenario Keys version 3, CAPRI group working paper. University of Manchester.

Elliot, M. J. and Dale A (1998) Disclosure Risk for Microdata. Report to the European Union ESP/ 204 62/DG III.

Muller, W; Blien, U.; and Wirth, H. (1992). Disclosure risks of anonymous individual data. Paper presented at the 1st International Seminar for Statistical Disclosure. Dublin 1992.

Skinner, C. J. and Holmes, D. J.(1998) Estimating the Re-identification Risk per Record in Microdata. *Journal of Official Statistics*. Vol. 14 (4) 361-372.

II

Tabular data protection

New Implementations of Noise for Tabular Magnitude Data, Synthetic Tabular Frequency and Microdata, and a Remote Microdata Analysis System

Laura Zayatz

U.S. Census Bureau[1], Commerce/Census/SRD/5K011, 4600 Silver Hill Road, Washington, DC 20233, Laura.zayatz@census.gov

Abstract: The U.S. Census Bureau collects its survey and census data under Title 13 of the U.S. Code which promises confidentiality to its respondents. The agency also has the responsibility of releasing data for the purpose of statistical analysis. In common with most national statistical institutes, our goal is to release as much high quality data as possible without violating the pledge of confidentiality. We apply disclosure avoidance techniques prior to publicly releasing our data products to protect the confidentiality of our respondents and their data (Willenborg and de Waal, 2001). This paper discusses three areas of current disclosure avoidance research: noise for tabular magnitude data, synthetic tabular frequency and microdata, and a remote access system. It also discusses how the methods developed needed to be altered when we applied them to real data, and how they are currently being used on real data products.

Key Words: Disclosure Avoidance, Confidentiality, Public Use Data Products

1 Introduction

The U.S. Census Bureau collects its survey and census data under Title 13 of the U.S. Code. This prevents the Census Bureau from releasing any data "...whereby the data furnished by any particular establishment or individual under this title can be identified." In addition to Title 13, the Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA) requires the protection of information collected or acquired for exclusively statistical purposes under a pledge of confidentiality. In addition, the agency has the responsibility of releasing data for the purpose of statistical analysis. In common with most national statistical institutes, our goal is to release as

[1] This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed are those of the author and not necessarily those of the U.S. Census Bureau.

much high quality data as possible without violating the pledge of confidentiality (Duncan, Keller-McNulty, and Stokes, 2003; Kaufman, Seastrom, and Roey, 2005)). We apply disclosure avoidance techniques prior to publicly releasing our data products to protect the confidentiality of our respondents and their data. This paper discusses three areas of current disclosure avoidance research: noise for tabular magnitude data, synthetic tabular frequency and microdata, and a remote access system. For each technique, we give an introduction to the method. We describe what happened when we applied the method to real data. We discuss how we needed to alter the method for use on real data products. We then list the public use data products that currently use the method.

2 Noise for Tabular Magnitude Data

2.1 Introduction to the Method

This technique is an alternative to cell suppression which we have used for decades for tabular magnitude data. Noise is added to the underlying microdata prior to tabulation (Evans, Zayatz, and Slanta, 1998; Massell, Zayatz, and Funk, 2006). Each responding company's data are perturbed by a small amount, say 10% (the actual percent is confidential), in either direction. Noise is added in such a way that cell values that would normally be primary suppressions, thus needing protection, are changed by a large amount, while cell values that are not sensitive are changed by a small amount. Noise has several advantages over cell suppression. It enables data to be shown in all cells in all tables. It eliminates the need to coordinate cell suppression patterns between tables. It is a much less complicated and less time-consuming procedure than cell suppression. Because noise is added at the microdata level, additivity of the table is guaranteed.

To perturb an establishment's data by about 10%, we multiply its data by a random number that is close to either 1.1 or 0.9. We could use any of several types of distributions from which to choose our multipliers, and the distributions remain confidential within the agency. The overall distribution of the multipliers is symmetric about 1. The noise procedure does not introduce any bias into the cell values for census or survey data. Because we protect the data at the company level, all establishments within a given company are perturbed in the same direction. The introduction of noise causes the variance of an estimate to increase by an amount equal to the square of the difference between the original cell value and the noise added value. One could incorporate this information into published coefficients of variation.

2.2 Applications to Real Data

A problem we found with the noise technique is that it can add excessive amounts of distortion to cells that would be shown much more precisely under deterministic methods such as cell suppression and controlled tabular adjustment. A natural question to ask then is whether or not there is a way to modify the method such that it adds less noise to the non-sensitive cells, while retaining the amount of protection provided to the sensitive cells. One of the primary benefits of the noise technique over cell suppression is that it allows for the release of more usable data. Suppression has the advantage that all published cell values are the exact estimates collected by an agency, and so if noisy cell values are highly distorted then the benefit of noise over suppression is significantly reduced. As a result, we investigated possible methods to reduce the overall amount of noise add to the data without compromising the level of protection (Massell and Funk, 2007a).

The Census Bureau's magnitude data is almost always published in some rounded form, often in integer form representing thousands or millions of dollars. This type of rounding can be done at the record level prior to any tabulations, or applied directly to the unrounded table values. Noise is designed to protect individual respondents by changing their response values by small percentages. Rounding can therefore systematically remove the effect of noise on small response values. In some cases, this may not be an issue of concern, but under certain circumstances, this occurrence could result in serious damage to the level of protection provided by noise. We investigated a few types of rounding methods that could work to sustain the adequate protection provided by noise (Massell and Funk, 2007b).

2.3 Modifying the Method

In order to decrease the amount of distortion to non-sensitive cells, we developed a balanced noise technique, the details of which are described in (Massell and Funk, 2007a). The technique involves choosing one or more tables for the balancing application. This choice is determined after experimenting with several options and can be different for different surveys and censuses, but it is typically at a lower level in the hierarchy of related tables and has a trickle up effect. In step 1, random noise multipliers are applied to establishments in cells with only 1 or 2 contributors and to establishments from companies represented in more than 1 cell. In the second step, multipliers are applied to all other establishments (single unit establishments not in cells with only 1 or 2 contributors) taking into account the multipliers that have been previously assigned in such a way that the distortion in non-sensitive cells is minimized. If multipliers assigned in step one would lead to an increase (or decrease) in a non-sensitive cell's value, then multipliers assigned in step two would be created to make the final cell value have the least amount of distortion possible.

In (Massell and Funk, 2007b), one can find the results of testing various modifications to the standard rounding techniques. They include rounding of the underlying microdata values and rounding of tabulated cell values to ensure that standard rounding does not undo the protection offered to small cells by the noise. In general, ceiling/floor techniques seem to work best, but all of the techniques should be tested for each survey or census to identify the best technique for a given set of data.

2.4 Current Uses on Public Use Data Products

A few years ago, John Abowd (Cornell University and the Census Bureau) was developing a new data product: Quarterly Workforce Indicators. John was not a fan of cell suppression (he did not like holes in the data), so he decided to use the noise technique and the Disclosure Review Board approved it. Since then, he has been using noise and recently added a small amount of synthetic data.

Then, staff at the Census Bureau who work on the Commodity Flow Survey were considering adding many more detailed tables to their public data products but did not want them filled with suppressions, so they decided to use the noise technique. The method has since caught on. We have used it for our Non-Employer data products, and plan to use it for our Census of Island Areas,

Survey of Business Owners, and Commodity Flow Survey. The Associate Director for our Economic Programs says that he sees this as the future technique for most of our tabular magnitude data products.

3 Synthetic Tabular Frequency and Microdata

3.1 Introduction to the Method

Given a data set, one can develop posterior predictive models to generate synthetic data that have many of the same statistical properties as the original data (Abowd and Woodcock, 2001). Generating the synthetic data is often done by sequential regression imputation, one variable in one record at a time (Rubin, 1993). Using all of the original data, we develop a regression model for a given variable (Raghunathan, Reiter, and Rubin, 2003). Then, for each record, we blank the value of that variable and use the model to impute for it. Then, we go to the next variable and repeat the process (Reiter, 2003 and Reiter, 2004).

Synthesizing data can be done in different ways and for different types of data products. One can synthesize all variables for all records (full synthesis) or a subset of variables for a subset of records (partial synthesis). If doing partial synthesization, we target records that have a potential disclosure risk and those variables that are causing this risk. We can synthesize demographic data and establishment data, though demographic data are easier to model and synthesize. We can synthesize data with a goal of releasing the synthetic microdata or some tabulation or other type of product (such as a map) generated from the synthetic microdata. And finally, we can generate one implicate (one synthetic data set) which looks similar to the original file, but with synthetic data; or we can generate several implicates (several different synthetic data sets) that could be released together. Multiple synthetic implicates can be analyzed using multiple imputation analysis techniques.

3.2 Applications to Real Data

The biggest problems we have encountered when trying to generate synthetic data from real data are relationships between variables within a data set. For example, for many of our microdata files, we have records for households which are linked with records from all people within those households (think of a family with a mother, father, son, and daughter). Many values for many of

the variables will be structurally missing (or blank) because of skip patterns in the survey instrument. For example, the number of children ever born to the father and son will be missing. The income of anyone under 15 will be missing. Other combinations of variables would not make sense. For example, we cannot have a mother who is only 7 years older than her own child after synthesizing the age variable.

3.3 Modifying the Method

We often impute some of the structurally missing values, but at the end of the synthesis procedure, restore them to missing for standard imputation and edits. For the SIPP/SSA file described in the next section, synthesizing the logical structure of the complicated longitudinal survey was not possible using the existing methods. Imposing structural zeros involving the interactions of several variables on the feasible statistical models we were using required an additional layer of programming that eventually became a nine-level collection of parent-child relationships that enforced all of the constraints.

3.4 Current Uses on Public Use Data Products

John Abowd (Cornell University) lead a group in developing a public use microdata file containing linked Social Security Administration earnings data and the Census Bureau's Survey of Income and Program Participation (SIPP) data with the goal of releasing multiple synthetic implicates. If we want to begin releasing public use files that link our data with data from other agencies, synthetic data are probably our only choice. Other statistical avoidance techniques are not sufficient to protect the confidentiality of such files. The vast majority of the variables on the file are synthesized. The two agencies were responsible for judging the quality of the final data product. The Census Bureau's Disclosure Avoidance Research Group used record linkage software to ensure the resulting data cannot be linked to any of our SIPP public use microdata files.

John Abowd developed another product called "On The Map," which is a set of maps of transportation data. See. http://lehd.did.census.gov. The maps are based on partially synthetic data. The Disclosure Review Board looked at the data underlying the maps and decided that the synthetic data were sufficiently different from the original data, especially in small geographic areas. John compared the resulting maps and decided they looked almost identical, so everyone was pleased with the product. In developing this product, it helped

knowing its intended use, and one should also note that only a handful of variables needed to be synthesized.

We are using partially synthetic data to protect both frequency tabular and microdata for Group Quarters data in the American Community Survey (ACS). We are conducting research to see if we should use the method to protect more or even all ACS data products (Hawala and Funk, 2007), and to see if the synthesized data are an improvement over currently imputed values for missing data.

4 Remote Microdata Analysis System

4.1 Introduction to the Method

The American FactFinder (Rowland and Zayatz, 2001) was developed to allow for broader and easier access to the standard Summary Files (frequency count data) from Census 2000 and to allow data users to generate their own tabular data products from Census 2000. See http://factfinder.census.gov/home/saff/main.html.
One part of American FactFinder is the Advanced Query System (AQS). The goal of the AQS is to allow users to submit requests for user-defined tabular data electronically. A request passes through a firewall to an internal Census Bureau server, which holds a previously swapped, recoded, and topcoded microdata file. The table is created and electronically reviewed for disclosure problems. If it is judged to have none, the table is sent back electronically to the user.

The AQS accepts queries only for tables and only from Census 2000 data. We would like to see if we can expand its capabilities to handle data from other demographic surveys and other types of statistical analysis. We are currently developing a prototype of a Microdata Analysis System (MAS) that would do just that. It is a web-based system. The user selects the data set, the geography, the universe, the type of analysis, and the variables (or transformations thereof). The web site generates the SAS code needed to arrive at the desired results. The user may see the SAS code but may not alter it. The generated code is run against the data and the results are verified. If the output passes the results filter (we are working on this now), it is returned to the user.

4.2 Applications to Real Data

Perhaps the biggest decision that we needed to make about the Microdata Analysis System is whether we wanted a "disabled" or an "enabled" type of system. We could allow users to write and submit their SAS own code and disable some procedures such as PROC PRINT and PROC LIST (anything that would allow users to see the underlying microdata). The other option was to enable users to choose, from various menus, what they wanted to do and have the system generate the SAS code. We chose the second (enabled) option. There are advantages and disadvantages to both. A disabled system gives many more options to a data user, but it may require quite a bit of "babysitting" in that an agency should probably restrict who has access to such a system and strictly monitor the requests coming in and the results going out. An enabled system restricts the types of analyses that can be done, but could be made available to the general public without strict monitoring.

Also, in the early part of our work, we were focusing on the model statements in terms of looking for disclosure problems, but we soon realized that we needed to look at the underlying data tables that would be used in, for example, a regression. We also took a lesson already learned from the Advanced Query System that we will need to offer short, medium, and long lists of categories for certain variables so that users can obtain as much detail as possible in their analyses without running into potential disclosure problems (add AQS reference here).

4.3 Modifying the Method

We are now modifying the system to look at the tables underlying any type of analysis, and in particular to look at the marginal totals in the tables, because marginal totals of size 1 could potentially be used through several queries to put together a microdata record. We are also conducting research to find the best ways of identifying "cut points" in our short, medium, and long lists of continuous variables, again so that users can obtain as much detail as possible without disclosure problems. This includes automatic treatment of negative values, missing values, and non-monetary continuous values.

4.4 Current Uses on Public Use Data Products

The AQS is currently available to the Census Bureau's State Data Centers and Census Information Centers as well as a group of beta testers. Data users can

contact these Centers to request free tabulations (Weinberg, et.al. 2007). The Microdata Analysis System is still under development. It is being tested with American Community Survey and Current Population Survey data (Steel and Zayatz, 2006).

5 Conclusion

There have been several recent developments in disclosure avoidance at the Census Bureau. We are using the noise addition technique for tabular magnitude data for several data products. We have released several data products on based on partially synthetic data. The Advanced Query System was completed and is being widely used by State Data Centers and Census Information Centers, and we will continue our work on the Microdata Analysis System. The new techniques all took a few years to develop conceptually, and then took more years to adapt for use with real data. The work proved well worth it when we began using the techniques on real, public use data products.

References

Abowd, J. M. And Woodcock, S. D. (2001), "Disclosure Limitation in Longitudinal Linked Data," Confidentiality , Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies, Doyle, P., Lane, J., Zayatz, L., and Theeuwes, J., eds.,Elsevier Science, The Netherlands, pp. 215-277.

Duncan, G. T., Keller-McNulty, S., and Stokes, S. L. (2003), "Disclosure Risk vs. Data Utility: The R-U Confidentiality Map," Technical Report 2003-6, Heinz School of Public Policy and Management, Carnegie Mellon University.

Evans, B. T., Zayatz, L., and Slanta, J. (1998), "Using Noise for Disclosure Limitation for Establishment Tabular Data," *Journal of Official Statistics*, Vol. 14, No. 4, pp. 537-552.

Hawala, S. and Funk, J. (2007), "Model Based Disclosure Avoidance for Data on Veterans," *Proceedings of the Research Conference of the Federal Committee on Statistical Methodology*, Arlington, Virginia, November 5-7, 2007.

Kaufman, S., Seastrom, M., and Roey, S. (2005), "Do Disclosure Controls to

Protect Confidentiality Degrade the Quality of the Data?" American Statistical Association, Proceedings of the Section on Survey Research.

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

Massell, P. and Funk, J. (2007b), "Recent Developments in the Use of Noise for Protecting Magnitude Data Tables: Balancing to Improve Data Quality and Rounding that Preserves Protection," *Proceedings of the Research Conference of the Federal Committee on Statistical Methodology*, Arlington, Virginia, November 5-7, 2007.

Massell, P., Zayatz, L., and Funk, J. (2006), "Protecting the Confidentiality of Survey Tabular Data by Adding Noise to the Underlying Microdata: Application to the Commodity Flow Survey," Privacy in Statistical Databases, CENEX-SDS Project International Conference, PSD 2006, Proceedings, Lecture Notes in Computer Science (LNCS) 4302, Springer 2006, ISBN 3-540-49330-1.

Raghunathan, T. E., Reiter, J. P., and Rubin, D. B. (2003), "Multiple Imputation for Statistical Disclosure Limitation," *Journal of Official Statistics*, 19, pp. 1-16.

Reiter, J. P. (2003), "Inference for Partially Synthetic, Public Use Microdata Sets", *Survey Methodology*, 29, pp. 181-188.

Reiter, J. P. (2004), "Simultaneous Use of Multiple Imputation for Missing Data and Disclosure Limitation," *Survey Methodology*, 30, pp. 235-242.

Rowland, S. and Zayatz, L. (2001), "Automating Access with Confidentiality Protection: The American FactFinder," Proceedings of the Section on Government Statistics, American Statistical Association.

Rubin, D. B. (1993), "Discussion of Statistical Disclosure Limitation," *Journal of Official Statistics*, 9, pp. 461-468.

Steel, P. and Zayatz, L. (2006), "Description of a Microdata Access System"

for Presentation to the Census Advisory Committee of Professional Associations, US Census Bureau, October 27, 2006.

Weinberg, D., Abowd, J., Rowland, S., Steel, P., and Zayatz, L. (2007), "Access Methods for United States Microdata," *Proceedings of the Workshop on Data Access to Microdata, Nurembourg,* Germany, August 20-21, 2007. Also found on the Social Science Research Network http://hq.ssrn.com and US Census Bureau Center for Economic Studies Paper No. CES-WP-07-25.

Willenborg, L. and de Waal, T. (2001), Elements of Statistical Disclosure Control, Springer-Verlag New York, Inc.

An Examination of Two Methods for Controlled Tabular Adjustment of Tabular Data That Preserve Data Quality

Lawrence H. Cox

National Center for Health Statistics, Centers for Disease Control and Prevention
Hyattsville, MD 20782 USA
LCOX@CDC.GOV

Keywords. Linear programming, iterative proportional fitting, Kullback-Leibler, QP-CTA, MDI-CTA

Abstract. Two methods for balancing data quality and data confidentiality for tabular data have recently emerged. Each limits disclosure through controlled tabular adjustment (CTA). The first, Quality-Preserving (QP-) CTA, achieves CTA through a linear programming model, and preserves quality by controlling change to selected parameters and statistics associated with the original distribution via specialized capacities and constraints incorporated into the linear program. The second, Minimum Discrimination Information (MDI-) CTA, is an iterative procedure that at each stage employs iterative proportional fitting (IPF) to achieve a current CTA solution exhibiting minimum Kullback-Leibler distance (MDI) from the original distribution (table) for fixed inputs, and at the next stage refines the inputs and repeats the IPF in an attempt to produce a next CTA solution exhibiting smaller MDI. Based on limited testing, both methods appear to perform well on a practical basis. The strengths and limitations of the two methods are often opposite to each other. This paper explores these properties with an eye towards an enhanced CTA methodology.

1 Introduction

Tabular data are ubiquitous. Standard forms include count data as in population and health statistics, concentration or percentage data as in financial or energy statistics, and magnitude data such as retail sales in business statistics or average daily air pollution in environmental statistics. Tabular data remain a staple of official statistics. Data confidentiality was first investigated for tabular data [1, 2]. Tabular data are additive and expressible as specialized systems of linear equations: $\mathbf{Tx} = \mathbf{v}$, where $\mathbf{x}$ represents the *tabular cells,* $\mathbf{T}$ the *tabular equations*, and $\mathbf{v}$ fixed values. Entries of $\mathbf{T}$ are in the set {-1, 0, +1}, and each row of $\mathbf{T}$ contains at most one -1.

For decades, the prevailing *statistical disclosure limitation* (SDL) [3] method for tabular magnitude data and, to a lesser but considerable extent, tabular count data, was *complementary cell suppression* (CCS) [1, 4, 5]. Recently, an alternative SDL method named *controlled tabular adjustment* (CTA) has emerged [6, 7]. This development was motivated by computational complexity, analytical obstacles, and user dissatisfaction with CCS [8].

Disclosure in tabular data is typically defined by a *threshold rule* (count data) or a *dominance rule* (magnitude data), and more generally by a *linear sensitivity measure* [9]. Counts of 1 or 2 are likely to identify individuals, or sales data may be dominated by contributions of n = 1 or 2 contributors, and thereby provide a close estimate of the larger contributor's confidential

business information to the other contributor (n = 2) or to the public (n = 1). The sensitivity measure also determines minimal distances from the sensitive cell value to *safe values* above and below it. The *protection interval* for the sensitive cell value is defined to be the set of all unsafe values—values between its lower and upper safe values. See [9] for details.

Complementary cell suppression removes from publication the values of all sensitive cells and in addition removes sufficiently many nonsensitive cell values to ensure that the linear system $\mathbf{Tx} = \mathbf{v}$ does not reveal a sensitive cell value or locate it within an interval finer than its protection interval ([1, 9]). Drawbacks of cell suppression for statistical analysis include removal of useful and otherwise harmless information and consequent difficulties analyzing tabular systems with cell values missing not-at-random.

Controlled tabular adjustment replaces sensitive cell values with safe values and, because these adjustments almost certainly throw the tabular system $\mathbf{Tx} = \mathbf{v}$ out of kilter, CTA adjusts some or all of the nonsensitive cells by small amounts to rebalance the additive system. In terms of ease-of-use, controlled tabular adjustment is unquestionably an improvement over cell suppression. As CTA changes sensitive and other cell values, the data quality issue is then: Can CTA be accomplished while preserving important data analytical properties of original data?

A preliminary discussion of quality aspects of CTA was introduced in [8]. A methodology for preserving distributional parameters of linear models for univariate distributions was introduced in [6], and named *quality-preserving controlled tabular adjustment* (QP-CTA). In [10], QP-CTA was extended to multivariate distributions. QP-CTA is based on mathematical (mostly, linear) programming. A statistical approach based on iterative proportional fitting to preserve the original data distribution, as measured by Kullback-Leibler distance ([11]) between the original and adjusted tables, was introduced in [12], and named *minimum discrimination information controlled tabular adjustment* (MDI-CTA).

QP-CTA and MDI-CTA each appear to perform well and efficiently based on limited testing. The strengths and limitations of the two methods are in many cases opposite to each other. The purpose of this paper is to examine and compare these properties, with an eye towards an enhanced methodology that exploits their combined strengths. In Section 2, we specify the two methods mathematically. In Section 3, we examine their quality characteristics and, in Section 4, compare them. Section 5 provides concluding comments.

2 Controlled Tabular Adjustment and Data Quality

2.1 The Basic CTA Methodology

CTA is applicable to all tabular data, but for convenience we focus on magnitude data. A simple paradigm for statistical disclosure in magnitude data follows. Tabulation cell i comprises k respondents (e.g., retail clothing stores in a county) and a statistic of interest (e.g., retails sales). The NSO assumes that any respondent is aware of the identity of the other respondents. The cell value is the total value of the statistic of interest, viz., the sum of nonnegative values of this statistic (called *contributions*) by each respondent in the cell. Denote the cell value $v^{(i)}$ and the contributions $v_j^{(i)}$, ordered from largest to smallest. It is possible for any respondent j to compute

$v^{(i)} - v_j^{(i)}$ which is an upper estimate of the contribution of any other respondent. This estimate is closest, in percentage terms, when the target is the largest respondent and j = 2. The *p-percent rule* declares that the cell value represents disclosure whenever this estimate is less than (100 + p)-percent of the largest contribution. The sensitive cells are those failing this condition. This is a standard rule and has a corresponding linear sensitivity measure ([9]).

The NSO may also assume that any respondent can use public knowledge to estimate the contribution of any other respondent to within q-percent (q > p, e.g., q = 50%). This information allows the second largest to estimate $v^{(i)} - v_1^{(i)} - v_2^{(i)}$ to within q-percent. This upper estimate provides the second largest a lower estimate of $v_1^{(i)}$. This is referred to as the *p/q-ambiguity rule*, also associated with a linear sensitivity measure ([9])

The *lower* and *upper protection limits* for the cell value equal, respectively, the minimum amount that must be subtracted from (added to) the cell value so that these lower (upper) estimates are at least p-percent away from the true value $v_1^{(i)}$. Numeric values outside the protection limit range of the true value are its safe values. A common NSO practice is to assume that both protection limits are equal to a common value p_i. Complementary cell suppression suppresses all sensitive cells from publication, replacing sensitive values by variables in the tabular system **Tx = v**. Because, almost surely, one or more suppressed sensitive cell values can be estimated via linear programming to within its unsafe range, it is necessary to suppress additional nonsensitive cell values until no sensitive estimates can be obtained. This yields a mixed integer linear programming (MILP) problem over binary suppression variables ([4, 5]).

Controlled tabular adjustment replaces each sensitive value with a safe value. This is an improvement over complementary cell suppression as it replaces a suppression symbol by an actual value. However, safe values are not necessarily unbiased estimates of true values. To minimize bias, it is often desirable to replace the true value by either of its nearest safe values, $v^{(i)} - p_i$ or $v^{(i)} + p_i$. Because these assignments almost surely throw the tabular system out of kilter, CTA adjusts nonsensitive values to restore additivity. Because choices to adjust each sensitive value down or up are binary, combined these steps define a MILP. Typically, its *linear programming relaxation* is solved. Even as a MILP, CTA is superior to CCS because, for CCS, the MILP assigns one binary variable to each nonsensitive cell while, for CTA, binary variables are assigned to the sensitive cells, which almost certainly are far fewer in number.

A MILP by itself will not assure that analytical properties of original and adjusted data are comparable. Three simple procedures aimed at preserving quality were introduced in [8]. First, sensitive values are replaced by nearest safe values to reduce statistical bias. Second, lower and upper bounds (*capacities*) are imposed on adjustments to nonsensitive values to keep individual adjustments acceptably small, e.g., capacities might be based on estimated cell value measurement error e_i. Third, the objective function for the linear program is an overall measure of data distortion such as minimum sum of absolute, or percent, adjustments. An (arbitrarily large) upper bound on adjustment of sensitive cell value v_i is denoted m_i. The MILP model for CTA subject to the second and third criteria--while relaxing the first--is as follows ([10]).

Assume there are *n* tabulation cells of which the first *s* are sensitive, original data are represented by the nx1 vector **a**, adjusted data by $a + y^+ - y^-$; and $y = y^+ - y^-$. The MILP is:

$$\min \sum_{i=1}^{n} (y_i^+ + y_i^-) \quad \text{subject to:}$$

$$T\, y = 0 \tag{1}$$

$$p_i(1 - I_i) \le y_i^- \le m_i(1 - I_i), \quad p_i I_i \le y_i^+ \le m_i I_i \;;\; I_i \;\; binary \qquad i = 1,...s$$

$$0 \le y_i^-,\; y_i^+ \le e_i \qquad i = s+1, ..., n$$

If the capacities on adjustments to nonsensitive cells are too tight, it is possible that problem (1) be *infeasible* (lack solutions), requiring that some capacities be increased. A companion strategy, allows sensitive cell adjustments smaller than p_i in well-defined situations. This is justified mathematically because the intruder does not know if the adjusted value lies above or below the original value, but nevertheless is perceived by some as controversial.

Problem (1) is a MILP. The integer part can be solved by exact methods for small to medium-sized problems or via heuristics which first fix the integer variables and subsequently use linear programming to solve the linear programming relaxation. The remainder of this paper focuses on the problem of preserving data quality under CTA, and is not concerned with how the integer portion is being or has been solved.

2.2 QP-CTA: Using CTA to Preserve Parameters of Linear Models

For univariate data, we seek to preserve approximately mean and variance of original data and also correlation and regression slope between original and adjusted data, while maintaining additivity. For multivariate data, in addition adjusted data should preserve approximately covariance, correlation and regression coefficients between variables in original data.

Preserving mean values is straightforward. Any cell value a_i can be held fixed by forcing its corresponding adjustment variables y_i^+, y_i^- to zero, viz., set each variable's upper capacity to zero. Means are averages over sums. So, for example, to fix the grand mean, simply fix the grand total. Or, to fix means over all or a selected set of rows, columns, etc., in the tabular system, simply capacitate to zero adjustments to the corresponding totals.

In [6], it is shown how data quality objectives---variance, correlation and regression slope--- can be achieved by forcing covariance between original data **a** and adjustments **y** to them to be close to zero, while preserving the corresponding means, viz., over the indices comprising the mean, sum of upper adjustments equals sum of lower adjustments. Define

$$L(y) = \mathrm{Cov}(a, y)/\mathrm{Var}(a) \tag{2}$$

For any (renumbered) subset of *t* cells, as $\bar{y} = 0$: $L(y) = (1/(t\mathrm{Var}(a)))\sum_{i=1}^{t}(a_i - \bar{a})\, y_i$.

For variance, $Var(a + y) = (1/t)(\sum((a_i + y_i - (\bar{a} + \bar{y}))^2) = \mathrm{Var}(a) + (2/t)\sum(a_i - \bar{a})y_i + \mathrm{Var}(y)$.

So, $|Var(\text{a} + \text{y})/\text{Var(a)} - 1| = |2L(\text{y}) + (\text{Var(y)}/\text{Var(a)})|$ and relative change in variance can be minimized by minimizing the right-hand side. As $Var(\text{y})/\text{Var(a)}$ is typically small, it suffices to minimize $|L(\text{y})|$. This is accomplished by:

a) adjoin two additional linear constraints to (1):
$w \geq L(\text{y})$, $w \geq -L(\text{y})$, and (3)
b) minimize w

For univariate correlation, we seek Corr (a, a + y) = 1 approximately. As $\bar{y} = 0$,

$Corr(a,a+y) = Cov(a,a+y) / \sqrt{Var(a)Var(a+y)} = (1 + L(y)) / \sqrt{\text{Var(a + y)/ Var(a)}}$

As Var(y)/Var(a) is typically small, $\min |L(\text{y})|$ will typically suffice.

Finally, to preserve ordinary least squares regression $Y = \beta_1 X + \beta_0$ of adjusted data Y = a + y on original data X = a, we want β_1 near one and β_0 near zero:

$$\beta_1 = Cov(\text{a} + \text{y}, \text{a}) / \text{Var(a)} = 1 + \text{L(y)}, \quad \beta_0 = (\bar{\text{a}} + \bar{\text{y}}) - \beta_1 \bar{\text{a}}$$

As $\bar{y} = 0$, then $\beta_0 = 0$, $\beta_1 = 1$ if L(y) = 0 is feasible. Again, min $|L(\text{y})|$ suffices.

For multivariate data, in place of a single data set organized in tabular form, viz., **Ta = v**, to which adjustments **y** are to be made for confidentiality purposes, we have multiple data sets, each organized within a common tabular structure **T**. This is typical in official statistics where, e.g., tabulations would be shown at various levels of geography and industry classification for a range of variables such as total retail sales, cost of goods, number of employees, etc.

For concreteness, we focus on the bivariate case. Original data are denoted **a**, **b** and corresponding adjustments to original values are denoted by variables **y** and **z**. In the univariate case, the key to preserving variance, correlation and regression slope was to force Cov (**a, y**) = 0. It is easy to overlook in the univariate case that as Var (**a**) = Cov (**a, a**), then preserving variance via Cov (**a, y**) = 0 is equivalent to requiring Cov (**a, a + y**) = Cov (**a, a**). In the multivariate situation, however, preserving covariance (and variance) is of key importance and not to be overlooked. Namely, if we can preserve mean values and the variance-covariance matrix of original data, then we have preserved essential properties of the original data, particularly in the case of linear statistical models. We also would like to preserve simple linear regression of original data **b** on original data **a** in the adjusted data. And, of course, we wish to preserve the univariate properties of each variable.

To preserve Cov (**a, b**), we require: Cov (a, b) = Cov (a + y, b + z) = Cov (a, b) + Cov (a, z) + Cov (b, y) + Cov (y, z). Consequently, we seek:

$$\min \{|\text{Cov}(\mathbf{a}, \mathbf{z}) + \text{Cov}(\mathbf{b}, \mathbf{y}) + \text{Cov}(\mathbf{y}, \mathbf{z})|\}, \text{ subject to (1, 2, 3)} \quad (4)$$

The last term in the objective function is quadratic. For some problems, use of quadratic programming would be acceptable computationally. A linear approach to solving (4) heuristically is: perform successive alternating linear optimizations, viz., solve (2) for $\mathbf{y} = \mathbf{y_0}$, substitute $\mathbf{y}_0$ into (4) and solve for $\mathbf{z} = \mathbf{z_0}$, and continue in this fashion until an acceptable solution is reached.

The next objective is to preserve the estimated regression coefficient under simple linear regression of **b** on **a**. We seek approximately:

Cov (a, b) / Var (a) = Cov (a + y, b + z) / Var (a + y)

Var (a + y) / Var (a) = Cov (a + y, b + z) / Cov (a, b)

= 1 + Cov (a, z) / Cov (a, b) + Cov (b, y) / Cov (a, b) + Cov (y, z) / Cov (a, b)

Observe: $\mathrm{Var}(a + y)/\mathrm{Var}(a) = 2L(y) + 1 + \mathrm{Var}(y)/\mathrm{Var}(a)$

2L(y) + Var(y)/Var(a) = Cov (a, z) / Cov (a, b) + Cov (b, y) / Cov (a, b) + Cov (y, z) / Cov (a, b)

To preserve the regression coefficient, then we solve the linear program:

$$\min |(\mathrm{Cov}(\mathbf{a}, \mathbf{z}) + \mathrm{Cov}(\mathbf{b}, \mathbf{y})) / \mathrm{Cov}(\mathbf{a}, \mathbf{b})|, \text{ subject to (4)} \qquad (5)$$

To preserve correlation, we seek: Corr (**a, b**) = Corr (**a + y, b + z**). Equivalently:

$$\sqrt{\frac{Var(a+y)}{Var(a)}}\sqrt{\frac{Var(b+z)}{Var(b)}} = \frac{Cov(a+y,b+z)}{Cov(a,b)}$$

2.3 MDI-CTA: Achieving CTA While Preserving the Distributions Subject to Kullback-Leibler Divergence

Kullback-Leibler minimum discrimination information (MDI) ([11]) is a measure of distance between two statistical distributions. Often, the first distribution is known and the unknown second distribution is the closest distribution to the first within a predefined class of distributions, where "close" is measured by MDI. In our setting, the first distribution is the original distribution (table) and the class is the set of tables satisfying prespecified fixed marginal totals (*minimal sufficient statistics* = MSS). It is well-known that the iterative proportional fitting (IPF) procedure can be used to compute a unique solution that minimizes MDI. IPF permits fixing a subset of the cell values. In our setting, this subset includes sensitive cells set at selected safe values and structural zeroes. In [12], it is shown how to apply IPF to preserve distributions under CTA, as follows.

CTA fixes the values of the sensitive cells to safe values. Typically, these values are set equal to either the maximum lower or minimum upper safe value, viz., $v^{(i)} - p_i$ or $v^{(i)} + p_i$. This results in a binary choice for each sensitive cell, resulting in 2^s possible choices. Conditional on any one of these choices and on fixed MSS, it is possible to compute the IPF solution. The IPF solution maintains additivity to the MSS and, conditional on choices for the safe values, minimizes MDI relative to the original table. Based on a heuristic algorithm for updating choices of safe values to improve the MDI solution, the procedure of [12] iterates the IPF until the difference in MDI between the original and adjusted tables is statistically insignificant.

3 Data Quality Characteristics of the Two CTA Methods

In [13], two broad classes of data quality indicators for tabular data were introduced—local quality and global quality. *Local quality* refers to preserving or remaining close to original data values and relationships between them. *Global quality* refers to preserving the original distribution, its properties, and characteristics. A third category, arguably subsumed under both, is *structural quality*, e.g., preserving additivity.

3.1 Characteristics of QP-CTA

Operational characteristics of QP-CTA are as follows.

Pro

- preserves additivity
- relies on standard linear programming software
- capacities, constraints and objective are easily modified
- typically computationally efficient
- applicable to arbitrary tabular structure, dimension, and size
- can be performed in a multivariate setting
- does not require that marginal totals be fixed

Con

- can be solved exactly for certain structures, dimension, and size, but typically relies on heuristics to assign safe values to the sensitive cells
- objective function(s) and heuristics not tied to statistical criteria

Data quality characteristics of QP-CTA are as follows.

Local Quality

- adjustments to individual sensitive cells can be minimized
- adjustments to individual nonsensitive cells can be limited in size
- structural zeroes and other selected cell values can be exempt from change
- nonstructural zero cells can be adjusted away from zero

Global Quality

- can minimize global distance or average distance
- preserves univariate properties: mean, variance, correlation, regression
- preserves multivariate properties: covariance, regression
- can preserve these quantities for arbitrarily defined subsets of cells

3.2 Characteristics of MDI-CTA

Operational characteristics of MDI-CTA are as follows.

Pro

- preserves additivity
- relies on standard statistical algorithms available as software
- typically computationally efficient
- objective function(s) and heuristics tied to statistical criteria

Con

- relies on a heuristic for assigning safe values to sensitive cells
- not easily applied to arbitrary tabular structure
- no clear generalization to a multivariate setting
- requires (some) marginal totals to be fixed

Data quality characteristics of MDI-CTA are as follows.

Local Quality

- structural zeroes and other selected cell values can be exempt from change

Global Quality

- preserves original distribution, not just selected parameters or statistics
- nonstructural zero cells remain fixed at zero

4 Comparison of QP-CTA and MDI-CTA for Preserving Data Quality

Global data quality is concerned with preserving the original distribution and its properties. MDI-CTA preserves the original distribution conditional on the safe values. That is, if the number of sensitive cells is small relative to the total number of cells, and if safe values are not exceptionally large relative to nonsensitive values and estimated measurement errors, then it is reasonable to expect that MDI-CTA will preserve the original distribution. Whether this is the case or not can be verified by computing MDI or another test statistic to detect a statistically significant distance between original and adjusted data. If MDI-CTA has preserved the original distribution, then it is reasonable to expect that it also preserved important distributional parameters and statistics. This is not guaranteed but also can be verified. When releasing the adjusted data, it would be useful for the NSO also to release estimates of these quantities computed from original data.

QP-CTA preserves means, variances, covariances and regressions, so it is reasonable to expect that original and adjusted distributions are not too far apart, conditional on the safe values. Furthermore, if it is possible to limit adjustments to nonsensitive cells to within estimated measurement error, and if sensitive cells are adjusted to values at or near minimal safe values, then, conditional on the safe values, it is reasonable to expect that original and QP-CTA adjusted distributions are similar. This is not guaranteed, but can be verified by testing for a statistically significant MDI between original and adjusted tables.

Local data quality is concerned with changes to individual cell values and relationships between them. QP-CTA preserves local data quality directly via capacity constraints on adjustments to individual cell values, and in addition preserves covariances, correlations and regressions. Both QP-CTA and MDI-CTA can exempt selected values, including structural zeroes, from change. However, MDI-CTA does not control local changes to nonexempt cells, and there does not appear to be a way to modify IPF to incorporate capacity constraints. QP-CTA is able to adjust nonstructural zeroes away from zero. Replacing nonstructural zeroes with small “epsilon” values would enable MDI-CTA to do likewise.

Both methods preserve additivity to marginal totals. In some applications, marginal totals can be sensitive, and therefore adjusted, and in addition likely to require adjustment of additional marginal totals. In other cases, it may be desirable to permit adjustment of all or some nonsensitive marginal totals, e.g., to improve global quality. Conversely, if no marginal totals are sensitive, it may be desirable not to adjust any marginal total, e.g., when totals have been published previously. QP-CTA enables all of these choices. Currently, MDI-CTA does not enable any of them. This could be overcome if a set of MSS involving only nonsensitive marginals is identified. IPF then is performed based on the MSS, and each marginal total set equal to the sum of its constituent internal entries. As designed, MDI-CTA applies easily to a single, standard multi-dimensional table but not to arbitrary tabular structures.

Both methods employ heuristics for selecting the safe values. More work, e.g., [14], is needed on developing appropriate, effective heuristics. Regarding assessing goodness of fit, whereas MDI is convenient to perform the CTA and preserve the original distribution, it is not clear what is the most appropriate test statistic and work on that is needed also. Research into ways to combine these rather different methods into a stronger, combined method is indicated.

5 Concluding Comments

QP-CTA and MDI-CTA are two methods based on controlled tabular adjustment for producing a quality-preserving, disclosure-limited set of tabulations from an original set of tabulations that contains disclosure. We have presented mathematical/statistical models for applying these methods and examined their respective strengths and weaknesses operationally and for preserving data quality. We observed that often a strength of one method is a weakness of the other, and vice-versa, which motivated our comparison of their respective quality characteristics.

Disclaimer This paper represents the work of the author and is not intended to represent the policies or practices of the Centers for Disease Control and Prevention or any other organization.

References

1. Cox, L.H.: Suppression Methodology and Statistical Disclosure Control. Journal of the American Statistical Association. 75(1980) 377-385
2. Fellegi, I.P.: On the Question of Statistical Confidentiality. Journal of the American Statistical Association. 67 (1972) 7-18
3. U.S. Department of Commerce.: Statistical Disclosure and Disclosure Limitation Methods, Statistical Policy Working Paper 22, Washington, DC: Federal Committee on Statistical Methodology.(1994)
4. Cox, L.H.: Network Models for Complementary Cell Suppression. Journal of the American Statistical Association. 90(1995) 1153-1162.
5. Fischetti, M. and J.J. Salazar-Gonzalez: Models and Algorithms for Optimizing Cell Suppression in Tabular Data with Linear Constraints. Journal of the American Statistical Association. 95 (2000) 916-928.

6. Cox, L.H. and J.P. Kelly: Balancing Data Quality and Confidentiality for Tabular Data. Proceedings of the UNECE/Eurostat Work Session on Statistical Data Confidentiality, Luxembourg, 7-9 April 2003, Monographs of Official Statistics. Luxembourg: Eurostat (2003) 11-23.
7. Cox, L.H.: Discussion. ICES II: The Second International Conference on Establishment Surveys: Survey Methods for Businesses, Farms and Institutions. Alexandria, VA: American Statistical Association. (2000) 905-907.
8. Cox, L.H. and R.A. Dandekar: A New Disclosure Limitation Method for Tabular Data that Preserves Data Accuracy and Ease of Use. Proceedings of the 2002 FCSM Statistical Policy Seminar, Washington, DC: U.S. Office of Management and Budget (2003) http://www.fcsm.gov/working-papers/wp35_1.pdf
9. Cox, L.H.: Linear Sensitivity Measures in Statistical Disclosure Control. Journal of Statistical Planning and Inference 5 (1981) 153-164.
10. Cox, L.H., J.P. Kelly and R. Patil: Balancing Quality and Confidentiality for Multivariate Tabular Data. in: Privacy in Statistical Databases, Lecture Notes in Computer Science 3050 (J. Domingo-Ferrer and V. Torra, eds.), Berlin: Springer-Verlag (2004) 87-98.
11. Kullback, S. and R.A. Leibler: On Information and Efficiency. Annals of Mathematical Statistics, Volume 86 (1951) 79-86.
12. Cox, L.H., J.G. Orelien and B.V. Shah: A Method for Preserving Statistical Distributions Subject to Controlled Tabular Adjustment. in: **P**rivacy and Statistical Data Bases 2006, Lecture Notes in Computer Science 4302 (J. Domingo-Ferrer and L. Franconi, eds.), Heidelberg: Springer-Verlag (2006) 1-11.
13. Cox, L.H.: Balancing Quality and Confidentiality of Statistical Data. Proceedings of the 56th Session of the International Statistical Institute: Invited Papers, Voorburg: International Statistical Institute (2007), CD-ROM.
14. Glover, F., L.H. Cox, J.P. Kelly and R. Patil: Exact, Heuristic and Metaheuristic Methods for Confidentiality Protection by Controlled Tabular Adjustment, International Journal of Operations Research (2007), under review.

Comparative Evaluation of Four Different Sensitive Tabular Data Protection Methods Using a Real Life Table Structure of Complex Hierarchies and Links

Ramesh A Dandekar
Statistics and Methods Group, EI-70, U. S. Department of Energy, Washington DC 20585
Ramesh.dandekar@eia.doe.gov

Abstract The practitioners of tabular data protection methods in national statistical agencies have some familiarity with commonly used table structures. However, they require some guidance on how to evaluate appropriateness of various sensitive tabular data methods when applied to their own table structure. With that in mind, we use a real life "typical" table structure of moderate hierarchical and linked complexity and populate it with synthetic micro data to evaluate the relative performance of four different tabular data protection methods. The methods selected for the evaluation are: 1) lp-based classical cell suppression 2) lp-based CTA (Dandekar 2001), 3) network flow-based cell suppression as implemented in DiAna, a software product made available to other Federal statistical agencies by the US Census Bureau and 4) a micro data level noise addition method documented in a US Census Bureau research paper. The outcome from the comparative evaluation is available from http://mysite.verizon.net/vze7w8vk/

1 Introduction

To allow comparison of various sensitive tabular data protection methods on a consistent basis, the statistical disclosure control/limitation (SDC/SDL) researchers have long used public domain artificial (synthetic) data sets available from http://webpages.ull.es/users/casc/ website. The format used by these data sets, however, fails to convey visualization aspects of inherent complexities associated with various structural details typical of public use tables. The practitioners of tabular data protection methods are usually familiar with their own table structures. However, they require some assistance to evaluate appropriateness of proposed SDL methods when applied to their own table structure. As a first step to get around this problem, in this paper we use a real life table structure of moderate hierarchical and linked complexity and populate it with artificial (synthetic non-real) data to evaluate the relative performance of four different tabular data protection methods.

2 Table Structure

The templates for the hypothetical linked tables containing hierarchical structure selected for the comparative evaluation are in the Appendix. Appendix A shows column headings of these tables. The rows of the actual tables show geography in a hierarchical structure. However, only US totals are shown in Appendix A. The tables appear as two separate three-dimensional tables: Table 1 "Volumes by Grade, Sales Type, PAD District, and State" and Table 2 "Volumes by Formulation, Sales Type, PAD District, and State". For analytical purposes, these two tables could be considered as two three-dimensional linked portions of a four-dimensional table with

missing two-way interactions between grade and formulation. The tables consist of two independent (separable) components, namely, "Sales to End Users" and "Sales for Resale". The later component offers a far greater challenge for the sensitive data protection task and therefore is selected for the comparative evaluation.

The four-dimensional table template, without "Sales to End Users" stand alone part of the published table, is populated with artificial micro-data to create this example. The resulting table contains a total of 1556 non-zero cells. The p percent rule with p=10% is used to identify sensitive tabular cells. There are a total of 78 sensitive cells requiring protection from statistical disclosure. Appendix C illustrates the size relationship of sensitive cells and non-sensitive cells in the table. Appendix B is in two parts and contains only a partial listing of Table 1 to illustrate the format used to display the outcome from the four different data protection methods. The entire populated table structure containing artificial data is available in the public domain to SDL researchers from the web site http://mysite.verizon.net/vze7w8vk/ . Part 1 of Appendix B, displays the outcome from classical lp-based cell suppression method in first four columns. Sensitive cells are identified by a symbol 'w'. Non sensitive cells requiring suppression are identified by a symbol 's'. The last four columns of the table display the cell value adjustments from the CTA method. In these columns, the controlled tabular adjustment values to sensitive cells have been shown by a symbol 'w'. Adjustments to non-sensitive values are displayed by using symbol 'A'. Similarly, part 2 of Appendix B displays the outcome from DiAna software (network flow model) in the first four columns. Sensitive cells are identified by the symbol 'p'. Suppressed non-sensitive cells are identified by symbol 'c'. The last four columns of the table display the adjustments to cell value from the noise method.

3 Tabular Data Protection Methods

The methods selected for the comparative evaluation are: 1) classical lp-based cell suppression 2) lp-based CTA 3) DiAna's network flow-based cell suppression and 4) micro data level noise addition method described in a US Census Bureau research paper.

The classical lp-based cell suppression method used for the evaluation is similar to that used by CONFID at Statistics Canada since the mid-80. The selection of the complementary cell suppression pattern is done by using a cost proportional to the table cell value as an objective function. This results in higher preference for smaller tabular cells as complementary suppression cells.

The controlled tabular adjustments (CTA) a.k.a. synthetic tabular data method used is the one documented in Dandekar (2001) and Dandekar/Cox (2002) . Large size non-sensitive table cells are targeted for adjustments by using a cost function which is a reciprocal of the table cell value. Such an approach results in relatively small

percentage changes in the cell values and therefore, reduces the overall degradation in the accuracy of the statistical information imbedded in table cell values.

The network flow model in the DiAna software uses a minimal cost flow (mcf) based algorithm from the University of Texas to develop a complementary cell suppression pattern. The PC version of the software used for this evaluation targets smaller sized cells to develop a complementary cell suppression pattern.

The micro data level noise addition method as described in the paper http://www.census.gov/srd/papers/pdf/bte9601.pdf is used for this evaluation. Micro data is perturbed by an average of 10% and standard deviation of 0.005 by using a normal distribution.

4 Comparative Evaluation – Cell Suppression Methods

Complementary cell suppression methods have been used by statistical agencies for many years. Both network flow (DiAna) and classical simplex-based linear programming (Statistics Canada) methods have been used to develop cell suppression patterns. There are pros and cons associated with both methods. Network flow methods are computationally far more efficient than simplex based LP methods and therefore are preferred for large tasks. Auditing of a cell suppression pattern to identify potential problems arising from either insufficient or lack of protection from disclosure is a recommended follow-up procedural step to both cell suppression methods.

Our comparative evaluation of the two suppression methods shows that the DiAna's network flow based procedure results in 479 cells (31% of total non-zero cells) being suppressed. The classical LP based procedure results in 294 cell (19% of total non-zero cells) suppressions. The suppression count includes 78 sensitive cells. A relatively large number of cell suppressions associated with the network flow model is due to the sequential "one two-dimensional section at a time" procedure used by the network flow model. The software also lacks the capability to identify and remove un-necessary secondary cell suppressions. The classical LP-based procedure in the first pass suppresses 321 cells. The second pass through the procedure, which is commonly referred to as a "clean-up" procedure, reduces the suppressions to 294.

5 Comparative Evaluation – Noise vs CTA

The ultimate objective of the noise method and the CTA method is to protect the sensitive tabular data by a sufficient distortion of sensitive tabular cell values without adversely affecting the overall quality of the published non-sensitive tabular cells. The noise method takes an ***indirect approach*** in an ***attempt*** to achieve that objective by a systematic distortion of related micro data records. The CTA method, on the

other hand, takes a ***direct approach*** to achieve that objective by first adjusting the values for sensitive tabular cells by a ***precise*** amount determined by use of the linear cell sensitivity rule. The non-sensitive tabular cells are adjusted "minimally" by using some predetermined criteria. For the noise method, there is no known systematic procedure to determine a direct one-to-one mathematical/statistical relationship between micro data distributional characteristics and the highly aggregated multi-variate public use table structure.[1] As a result of the "ad hoc" nature of the noise method, it does not guarantee enough distortion (therefore, protection from statistical disclosure) of sensitive tabular cells. The noise method also results in unnecessary changes in values for non-sensitive tabular cells. In theory, one advantage of applying methods, like noise addition, directly to micro data is that all tables produced from the micro data will be protected. This would preclude the need for table specific analysis required of the other methods. However, in practice extensive quality control measures are required to ensure adequate protection from statistical disclosure of sensitive tabular cells and to avoid excessive adjustments to non-sensitive tabular cells. We have used the histogram of cell count by percent change in cell value to evaluate relative performance of the noise and the CTA method when applied to two 3-D Linked tables.

CTA vs NOISE - TABULAR DATA QUALITY

CTA frequency Distribution

% From	% To	Non-Sensitive	Sensitive
.00 -	.10	1235	0
.10 -	.50	137	1
.50 -	1.00	60	0
1.00 -	1.50	15	0
1.50 -	2.00	13	1
2.00 -	5.00	15	50
5.00 -	10.00	3	26
10.00 -	15.00	0	0
15.00 -	30.00	0	0
30.00 -	100.00	0	0

1432

Noise Frequency Distribution

% From	% To	Nonsensitive	Sensitive
.00-	.10	96	1
.10-	.50	272	0
.50-	1.00	265	0
1.00-	1.50	215	0
1.50-	2.00	164	0
2.00-	5.00	439	2
5.00-	10.00	27	51
10.00-	15.00	0	24
15.00-	30.00	0	0
30.00-	100.00	0	0

633

Based on 1% or less error as good data quality acceptance criteria, the CTA procedure provides 1432 (92% of total non-zero cells) good quality cells. The noise method, on the other hand, provides 633 (41% of total non-zero cells) good quality cells. Based on these statistics, it is clear that the CTA outperforms noise-based cell perturbation.

[1] Highly disaggregated multi-variate table structure in "limiting case" approaches related micro-data and therefore exhibits micro-data characteristics.

6 Comparative Evaluation – Cell Suppression vs Perturbation

Ease of implementation issues aside, in general in addition to protecting sensitive information, the overall objective for the cell suppression method is to "minimize" the information loss. Cell perturbation methods such as CTA or the noise method, on the other hand, are implemented to provide overall high quality information to the end users after adequately protecting imbedded sensitive information. Due to such inherent differences in the strategy, the cell suppression methodology usually targets smaller cells for complementary suppression, while perturbations are "***preferred***" to be targeted on adjusting larger non-sensitive cells. Such a preferential criterion is easy to implement in the CTA method by using an appropriate selection of the objective function. The noise method, unfortunately, does not allow for preferential treatment of tabular cells. This is further confirmed by comparing across four methods selected for the evaluation.

The network flow method performs better than the noise method (69% published cells vs. 41% good quality cells). The CTA method performs better than the classical lp-based cell suppression method (92% good quality cells vs. 81% published cells)

7 Expanding Table Structure to Include Missing Two Way Interactions

If for whatever reason, the agency decides to include for publication a missing two-way interaction between grades and formulation in this example, it would need to create and protect four-dimensional table structures. In Appendix D we provide a summary performance statistics related to four different tabular data protection methods when used on two 3-D linked tables and one 4-D table structure. Based on the summary performance statistics, the relative ranking of the four methods selected for the evaluation remains the same. The detailed output for the 4-D table is available on the website http://mysite.verizon.net/vze7w8vk/.

8 Dream or Reality?

In an ideal situation, a cell suppression method of choice should have a computational speed which is typical of network flow models and should create a cell suppression pattern which is typical of classical lp-based cell suppression methods. A preliminary research performed by this author, during a time frame from 1996 to 1997, shows that lp-based shrinking hypercube method (abstract Dandekar 2002) has a potential to offer such an alternative. In the table below we provide a comparative evaluation of multiple exploratory runs from the lp-based shrinking hypercube method, targeted towards suppressing smaller non-sensitive tabular cells.

LP_BASED HYPERCUBE RELATIVE TO CLASSICAL LP BASED SUPPRESSION

SUPPRESSION METHOD	CELL COUNT	QUANTITY SUPPRESSED
CLASICAL LP	294	886128
HYP(4,2,0.1)	287	954602
HYP(2,5,0.5)	319	1037975
HYP(1,1,0.1)	302	961529
HYP(4,0,0.5)	292	1085394

Conclusion

In this paper we have evaluated the outcome from four different tabular data protection methods by using a common table structure of moderate hierarchical and linked complexity. Our comparative evaluation ranks the CTA highest, followed by classical lp-based cell suppression in second place, network flow based method in third place and the noise based procedure in last place.

The choice of an "***appropriate***" tabular data protection method depends on multiple factors. Factors, such as available technical skills and resources play a critical role in the selection of a method of choice for a statistical agency. We hope that the information presented in this paper will be useful for the statistical agencies in deciding on the appropriateness of their selected tabular data protection method.

References

Dandekar R. A. (2001) "Synthetic Tabular Data: A Better Alternative To Complementary Data Suppression - Original Manuscript Dated December 2001". Energy Information Administration, U. S. Department of Energy. Also available from CENEX-SDC Project International Conference, PSD2006, Rome, Italy, December 13-15, 2006, Companion CD Proceedings ISBN: 84-690-2100-1.

Dandekar R. A. and Cox L. H. (2002), Synthetic Tabular Data: An Alternative to Complementary Cell Suppression, 2002. Manuscript, Energy Information Administration, U. S. Department of Energy.

Dandekar, R.A (2003), Cost Effective Implementation of Synthetic Tabulation (a.k.a. Controlled Tabular Adjustments) in Legacy and New Statistical Data Publication Systems, working paper 40, UNECE Work session on statistical data confidentiality (Luxembourg, 7-9 April 2003)

Dandekar Ramesh A. (2004), Maximum Utility-Minimum Information Loss Table Server Design for Statistical Disclosure Control of Tabular Data, pp 121-135, Lecture Notes in Computer Science, Publisher: Springer-Verlag Heidelberg, ISSN:

0302-9743, Volume 3050 / 2004**,** Title: Privacy in Statistical Databases: CASC Project International Workshop, PSD 2004, Barcelona, Spain, June 9-11, 2004.

Fischetti, M. and J. J. Salazar (2000), "Models and Algorithms for Optimizing Cell Suppression Problem in Tabular Data with Linear Constraints", *Journal of the American Statistical Association* **95**, 916-928.

Evans T., Zayatz L., Slanta j. (1998),"Using Noise for Disclosure Limitation of Establishment Data", USBC paper available from http://www.census.gov/srd/papers/pdf/bte9601.pdf

Appendix A—Templates, Column Headings From Two Linked Tables, Used With Geography

Table 1: Volumes by Grade, Sales Type, PAD District, and State

(Thousand Gallons per Day)

Geographic Area / Month	Regular						Midgrade					
	Sales to End Users		Sales for Resale				Sales to End Users		Sales for Resale			
	Through Retail Outlets	Total[1]	DTW	Rack	Bulk	Total	Through Retail Outlets	Total[1]	DTW	Rack	Bulk	Total
United States												
November 2006	45,169.5	46,307.2	34,864.4	185,329.5	31,904.2	252,218.1	6,049.4	6,116.4	2,261.5	11,172.0	–	13,524.4
October 2006	46,389.9	47,471.1	33,856.4	194,761.3	35,954.4	264,572.0	6,174.6	6,245.6	2,244.1	11,318.4	–	13,562.4
November 2005	47,061.6	48,145.3	34,836.8	190,758.1	45,057.6	270,652.6	5,602.7	5,678.0	2,779.3	12,955.6	–	15,734.9

Table 1 continue ..... Volumes by Grade, Sales Type, PAD District, and State

(Thousand Gallons per Day) — Continued

Geographic Area / Month	Premium						All Grades					
	Sales to End Users		Sales for Resale				Sales to End Users		Sales for Resale			
	Through Retail Outlets	Total[1]	DTW	Rack	Bulk	Total	Through Retail Outlets	Total[1]	DTW	Rack	Bulk	Total
United States												
November 2006	5,014.9	5,164.0	7,281.6	20,893.2	2,722.2	30,897.0	56,173.8	57,627.5	44,717.5	227,395.7	34,626.3	306,649.5
October 2006	5,118.9	5,256.2	7,236.7	20,710.1	2,945.0	30,900.7	57,681.3	59,072.9	43,337.2	226,798.7	38,899.3	309,035.2
November 2005	5,523.1	5,671.3	7,469.0	20,249.4	2,177.3	29,896.7	57,727.4	59,094.7	45,085.9	223,883.3	47,234.9	316,204.1

Table 2: Gasoline Volumes by Formulation, Sales Type, PAD District, and State

(Thousand Gallons per Day)

Geographic Area / Month	Conventional						Oxygenated					
	Sales to End Users		Sales for Resale				Sales to End Users		Sales for Resale			
	Through Retail Outlets	Total[1]	DTW	Rack	Bulk	Total	Through Retail Outlets	Total[1]	DTW	Rack	Bulk	Total
United States												
November 2006	31,687.1	32,542.6	7,582.5	164,118.7	32,280.1	203,980.3	2,305.5	2,456.6	1,687.8	8,257.9	–	11,895.8
October 2006	32,609.1	33,779.2	7,605.0	164,130.6	35,224.7	207,160.9	1,849.7	1,947.0	1,330.7	9,854.5	–	10,385.2
November 2005	32,962.9	33,939.7	8,023.9	162,294.3	38,193.8	209,511.6	2,752.0	2,882.1	1,933.7	8,932.9	–	10,499.7

Table 2: continue ..... Volumes by Formulation, Sales Type, PAD District, and State

(Thousand Gallons per Day) — Continued

Geographic Area / Month	Reformulated						All Formulations					
	Sales to End Users		Sales for Resale				Sales to End Users		Sales for Resale			
	Through Retail Outlets	Total[1]	DTW	Rack	Bulk	Total	Through Retail Outlets	Total[1]	DTW	Rack	Bulk	Total
United States												
November 2006	22,230.5	22,578.3	35,147.2	53,476.1	2,337.2	90,960.5	56,173.8	57,627.5	44,717.5	227,395.7	34,626.3	306,649.5
October 2006	22,912.6	23,246.7	34,200.0	53,095.6	3,674.6	91,481.1	57,681.3	59,072.9	43,337.2	226,798.7	38,899.3	309,035.2
November 2005	22,812.9	22,721.0	34,229.7	53,157.6	6,643.1	94,029.8	57,727.4	59,094.7	45,085.9	223,883.3	47,234.9	316,204.1

Appendix B

Part 1 – Complete Table available at: http://mysite.verizon.net/vze7w8vk/tableofcontents.pdf

Classical LP-Based Cell Suppresion vs CTA

Classical LP/CTA 01 regular

	DTW	Rack	Bulk	Total	<--- CTA Solution --->			
United States	188668.0	218471.0	170021.0	577160.0	-130.A	113.A	-61.A	-78.A
PAD District I	64625.0	72994.0	65620.0	203239.0	-8.A	69.A	-143.A	-82.A
Subdistrict IA	25314.0	28780.0	16952.0	71046.0	-8.A	8.A	0.	0.
Connecticut	6258.0	1494.0	1700.0	9452.0	0.	0.	0.	0.
Maine	3936.0	4719.0	4429.0	13084.0	0.	0.	0.	0.
Massachusetts	172.0 w	3840.0 s	.0	4012.0	-8.w	8.A	0.	0.
New Hampshire	7879.0	.0	3188.0	11067.0	0.	0.	0.	0.
Rhode Island	1748.0 s	6224.0 s	3976.0	11948.0	0.	0.	0.	0.
Vermont	5321.0	12503.0	3659.0	21483.0	0.	0.	0.	0.
Subdistrict IB	19417.0	16493.0	22335.0	58245.0	0.	48.A	-61.A	-13.A
Delaware	6978.0	2400.0	4272.0	13650.0	0.	48.A	0.	48.A
District of Columbia	2253.0	5070.0	11338.0	18661.0	0.	0.	0.	0.
Maryland	3111.0	1936.0 c	1079.0 w	6226.0	0.	0.	-60.w	-60.A
New Jersey	6875.0	.0	144.0	7019.0	0.	0.	0.	0.
New York	.0	648.0 s	784.0 w	1432.0	0.	0.	-39.w	-39.A
Pennsylvania	.0	6539.0	4719.0	11257.0	0.	0.	39.A	39.A
Subdistrict IC	19894.0	27721.0	26333.0	73948.0	0.	13.A	-82.A	-69.A
Florida	.0	10857.0	1847.0	12704.0	0.	0.	-17.A	-17.A
Georgia	9961.0	.0	.0	9961.0	0.	0.	0.	0.
North Carolina	2268.0 s	7226.0 s	8464.0 s	17958.0	0.	13.A	-65.A	-52.A
South Carolina	1195.0	5887.0	7582.0	14664.0	0.	0.	0.	0.
Virginia	3560.0	.0	3625.0	7185.0	0.	0.	0.	0.
West Virginia	2910.0 s	3751.0 s	4815.0 s	11476.0	0.	0.	0.	0.
PAD District II	76174.0	62147.0	54796.0	193117.0	-71.A	0.	126.A	55.A
Illinois	4128.0	.0	.0	4128.0	0.	0.	0.	0.
Indiana	4613.0 s	.0	3846.0 s	8459.0 s	-14.A	0.	14.A	0.
Iowa	1149.0	4196.0	4216.0 s	9561.0 s	0.	0.	0.	0.
Kansas	11996.0	10330.0	1948.0	24274.0	-57.A	0.	112.A	55.A
Kentucky	5826.0 c	2787.0 c	6523.0	15136.0	0.	0.	0.	0.
Michigan	2022.0 s	.0	6668.0 s	8690.0	0.	0.	0.	0.
Minnesota	6400.0	3694.0	1332.0	11426.0	0.	0.	0.	0.
Missouri	5915.0	10385.0	3934.0	20234.0	0.	0.	0.	0.
Nebraska	2652.0	7667.0	942.0	11261.0	0.	0.	0.	0.
North Dakota	4671.0	8286.0	.0	12957.0	0.	0.	0.	0.
Ohio	7197.0	.0	3477.0	10674.0	0.	0.	0.	0.
Oklahoma	4030.0	1854.0	4339.0	10233.0	0.	0.	0.	0.
South Dakota	24.0	11013.0	5526.0	16563.0	0.	0.	0.	0.
Tennessee	2242.0	645.0	8325.0	11212.0	0.	0.	0.	0.
Wisconsin	13309.0	1280.0 s	3720.0 s	18309.0	0.	0.	0.	0.
PAD District III	15248.0	23726.0	26417.0	65391.0	0.	-19.A	0.	-19.A
Alabama	3504.0	259.0 w	2856.0 s	6619.0	0.	25.w	0.	25.A
Arkansas	1598.0	5628.0	6358.0	13584.0	0.	0.	0.	0.
Louisiana	.0	3088.0 s	4667.0 s	7755.0	0.	0.	0.	0.
Mississippi	666.0	8925.0	2980.0	12571.0	0.	0.	0.	0.
New Mexico	9410.0	4928.0	6696.0	20034.0	0.	0.	0.	0.
Texas	1070.0	898.0 w	2860.0 s	4828.0	0.	-44.w	0.	-44.A
PAD District IV	13561.0	23112.0	8479.0	45152.0	-51.A	132.A	-44.A	37.A
Colorado	.0	8772.0	5637.0	14409.0	0.	0.	0.	0.
Idaho	925.0 s	940.0 w	890.0 w	2755.0 s	0.	94.w	-44.w	50.A
Montana	514.0 w	7358.0 s	.0	7872.0 s	-51.w	0.	0.	-51.A
Utah	5676.0 s	382.0 w	.0	6058.0 s	0.	38.w	0.	38.A
Wyoming	6446.0 s	5660.0 s	1952.0 s	14058.0	0.	0.	0.	0.
PAD District V	19060.0	36492.0	14709.0	70261.0	0.	-69.A	0.	-69.A
Alaska	.0	7949.0	4300.0	12249.0	0.	0.	0.	0.
Arizona	2721.0	828.0	2189.0	5738.0	0.	0.	0.	0.
California	3792.0	3728.0	2251.0	9771.0	0.	-69.A	0.	-69.A
Hawaii	1038.0	6141.0	327.0	7506.0	0.	0.	0.	0.
Nevada	2555.0	3522.0	.0	6077.0	0.	0.	0.	0.
Oregon	.0	14325.0	3040.0	17365.0	0.	0.	0.	0.
Washington	8954.0	.0	2602.0	11556.0	0.	0.	0.	0.

Appendix B

Part 2- Complete Table available at: http://mysite.verizon.net/vze7w8vk/tableofcontents.pdf

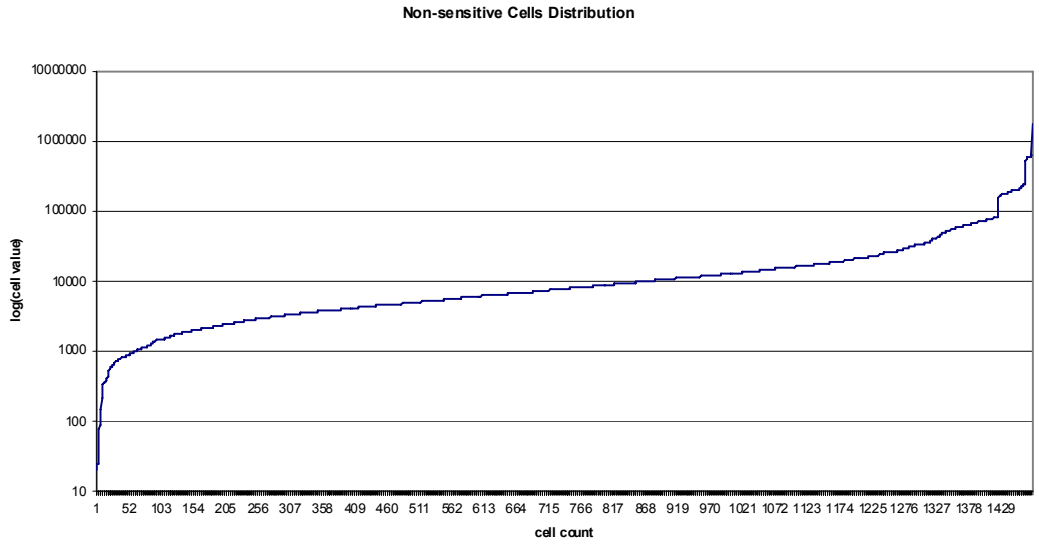
DiAna Cell Suppression Pattern vs Noise Addition

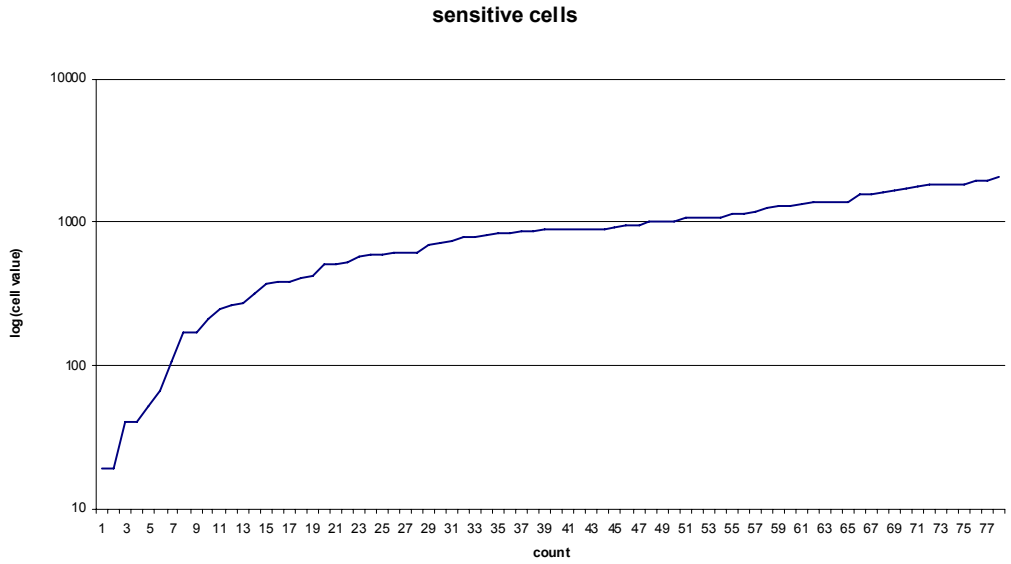
Diana/Noise 01 regular

	DTW		Rack		Bulk		Total		<--- Census' Noise Method-->			
United States	188668.		218471.		170021.		577160.		-16.	357.	81.	422.
PAD District I	64625.		72994.		65620.		203239.		-144.	219.	-53.	22.
Subdistrict IA	25314.		29780.		16952.		71046.		-267.	-115.	83.	-300.
Connecticut	6258.	C	1494.	C	1700.		9452.		-161.	-26.	-35.	-222.
Maine	3936.		4719.		4429.		13084.		150.	-69.	64.	145.
Massachusetts	172.	P	3840.	C	.		4012.		-9.	-65.	0.	-74.
New Hampshire	7879.		.		3188.		11067.		-159.	0.	81.	-78.
Rhode Island	1740.		6224.		3976.		11940.		-53.	143.	79.	168.
Vermont	5321.	C	12503.	C	3659.		21483.		-34.	-98.	-106.	-238.
Subdistrict IB	19417.		16493.		22335.		58245.		123.	51.	-137.	38.
Delaware	6978.	C	2400.	C	4272.		13650.		133.	19.	47.	200.
District of Columbia	2253.		5070.		11338.		18661.		-53.	85.	86.	118.
Maryland	3311.	C	1836.	C	1079.	P	6226.		-120.	-46.	-110.	-275.
New Jersey	6875.	C	.		144.	C	7019.		163.	0.	5.	168.
New York	.		648.	C	784.	P	1432.		0.	22.	-41.	-19.
Pennsylvania	.		6539.		4718.		11257.		0.	-30.	-124.	-154.
Subdistrict IC	19894.		27721.		26333.		73948.		0.	283.	1.	284.
Florida	.		10857.	C	1847.	C	12704.		0.	15.	-15.	1.
Georgia	9961.		.		.		9961.		93.	0.	0.	93.
North Carolina	2268.		7226.	C	8464.	C	17958.		94.	205.	-49.	250.
South Carolina	1196.		5887.		7502.		14664.		25.	-5.	165.	186.
Virginia	3560.		.		3625.		7185.		-187.	0.	-112.	-298.
West Virginia	2910.		3751.	C	4815.	C	11476.		-26.	68.	10.	52.
PAD District II	76174.		62147.		54796.		193117.		12.	186.	22.	220.
Illinois	4128.		.		.		4128.		154.	0.	0.	154.
Indiana	4613.		.		3846.	C	8459.	C	158.	0.	-112.	46.
Iowa	1149.	C	4196.	C	4216.	C	9561.		50.	-49.	60.	62.
Kansas	11996.		10330.		1948.		24274.		-62.	-156.	21.	-197.
Kentucky	5926.	C	2787.	C	6523.	C	15136.		106.	61.	-73.	94.
Michigan	2022.	C	.		6668.	C	8690.	C	73.	0.	-158.	-85.
Minnesota	6400.		3694.		1332.		11426.		-145.	110.	37.	2.
Missouri	5915.		10385.		3934.		20234.		-195.	46.	13.	-136.
Nebraska	2652.		7667.		942.		11261.		-80.	148.	32.	100.
North Dakota	4671.		8286.		.		12957.		30.	-18.	0.	12.
Ohio	7197.		.		3477.		10674.		-101.	0.	-40.	-141.
Oklahoma	4030.		1864.		4339.		10233.		46.	-52.	82.	76.
South Dakota	24.		11013.		5526.		16563.		1.	75.	94.	170.
Tennessee	2242.	C	645.	C	8325.		11212.		-70.	-14.	157.	73.
Wisconsin	13309.		1280.		3720.		18309.		47.	35.	-91.	-9.
PAD District III	15248.		23726.		26417.		65391.		30.	-233.	181.	-22.
Alabama	3504.	C	259.	P	2856.		6619.		-94.	-26.	-51.	-171.
Arkansas	1598.		5628.		6358.		13584.		11.	-43.	80.	48.
Louisiana	.		3088.	C	4667.	C	7755.		0.	-21.	115.	94.
Mississippi	666.	C	8925.	C	2980.		12571.		-23.	-218.	-155.	-396.
New Mexico	8410.		4928.		6696.		20034.		158.	121.	111.	389.
Texas	1070.	C	898.	P	2860.	C	4828.		-22.	-46.	81.	14.
PAD District IV	13561.		23112.		8479.		45152.		-106.	-133.	228.	-11.
Colorado	.		8772.		5637.		14409.		0.	-116.	129.	13.
Idaho	925.	C	940.	P	890.	P	2755.	C	-19.	100.	49.	131.
Montana	514.	P	7358.	C	.		7872.		53.	-23.	0.	30.
Utah	5676.	C	382.	P	.		6058.		-99.	-39.	0.	-138.
Wyoming	6446.		5660.	C	1952.	C	14058.	C	-41.	-54.	49.	-46.
PAD District V	19060.		36492.		14709.		70261.		192.	318.	-297.	213.
Alaska	.		7948.		4300.		12248.		0.	36.	-87.	-51.
Arizona	2721.	C	828.	C	2189.	C	5738.		82.	29.	-67.	44.
California	3792.		3728.	C	2251.	C	9771.		-84.	97.	-61.	-48.
Hawaii	1038.	C	6141.		327.	C	7506.		76.	-196.	11.	-108.
Nevada	2555.	C	3522.	C	.		6077.		-19.	56.	0.	37.
Oregon	.		14325.		3040.		17365.		0.	297.	-64.	233.
Washington	8954.		.		2602.		11556.		136.	0.	-30.	106.

Appendix C

Linked Tables 1 and 2 Cell Distribution





APPENDIX D

SUMMARY PERFORMANCE STATISTICS

TWO 3-D LINKED TABLES — 4-D ENTIRE TABLE

NETWORK FLOW MODEL

- 28 Column Variables
- 15 Column Relations
- 78 Sensitive Cells
- 479 Suppressions
- 31% Suppressions
- 1077 Published

- 64 Column Variables
- 48 Column Relations
- 267 Sensitive Cells
- 1844 Suppressions
- 62% Suppressed
- 1149 Published

CLASSICAL CELL SUPPRESSION

- 1556 Non-Zero Cells
- 2707 Equations
- 78 Sensitive Cells
- 294 Suppressions
- 19% Suppressions
- 1282 Published

- 2993 Non-Zero Cells
- 6273 Equations
- 267 Sensitive Cells
- 1143 Suppressions
- 38% Suppressions
- 1850 Published

CONTROLLED TABULAR ADJUSTMENT

% From	% To	Nonsensitive	Sensitive
.00-	.10	1235	0
.10-	.50	137	1
.50-	1.00	60	0
1.00-	1.50	15	0
1.50-	2.00	14	1
2.00-	5.00	14	50
5.00-	10.00	3	26
10.00-	15.00	0	0
15.00-	30.00	0	0
30.00-	100.00	0	0
1432 Good Quality		124 Poor Quality	

% From	% To	Non-Sensitive	Sensitive
.00 -	.10	1803	0
.10 -	.50	438	1
.50 -	1.00	214	0
1.00 -	1.50	97	1
1.50 -	2.00	59	0
2.00 -	5.00	103	171
5.00 -	10.00	12	94
10.00 -	15.00	0	0
15.00 -	30.00	0	0
30.00 -	100.00	0	0
2455 Good Quality		538 Poor Quality	

MICRO DATA LEVEL NOISE ADDITION

% From	% To	Nonsensitive	Sensitive
.00-	.10	96	1
.10-	.50	272	0
.50-	1.00	265	0
1.00-	1.50	215	0
1.50-	2.00	164	0
2.00-	5.00	439	2
5.00-	10.00	72	51
10.00-	15.00	0	24
15.00-	30.00	0	0
30.00-	100.00	0	0
633 Good Quality		923 Poor Quality	

% From	% To	Nonsensitive	Sensitive
.00-	.10	137	1
.10-	.50	416	0
.50-	1.00	400	0
1.00-	1.50	334	0
1.50-	2.00	322	0
2.00-	5.00	1069	2
5.00-	10.00	48	172
10.00-	15.00	0	92
15.00-	30.00	0	0
30.00-	100.00	0	0
953 Good Quality		2040 Poor Quality	

Assessing the Impact of SDC Methods on Census Frequency Tables

Natalie Shlomo*

* Southampton Statistical Sciences Research Institute, University of Southampton, Highfield, Southampton, UK SO17 1BJ, e-mail: N.Shlomo@Soton.ac.uk

Abstract: Statistical Agencies are faced with increasing demands by users to release more detail and high quality statistical data. This requires examining the trade off between managing disclosure risk below tolerable thresholds and disseminating "fit for purpose" data with as much information as possible. In particular, protecting Census data containing whole population counts is one of the greatest SDC challenges and confidentiality requirements and codes of practice are constantly changing to meet demands for high quality small area data. The impact of SDC methods on whole population counts causes much information loss and hence the need to evaluate a wide range of SDC methods. In this paper we take an in depth look at one particular large table from the UK 2001 Census with respect to measuring disclosure risk, implementing SDC methods and comparing their impact on information loss through measures based on distortion to distributions, measures of association and other statistical analysis tools.

1 Introduction

Statistical Agencies are facing increasing demands to disseminate more detail and high quality statistical data for small areas based on Census or administrative sources. The standard mode of dissemination for whole population counts are frequency tables. Protecting these tables is more difficult than protecting tables from a survey sample since the sampling introduces ambiguity into the frequency counts and as a result it is more difficult to identify statistical units without response knowledge nor infer what the true count may be in the population.

This paper provides a review of common Statistical Disclosure Control (SDC) methods for protecting tabular outputs containing whole population counts from Censuses or register-based data. Since more invasive SDC methods are needed to protect against disclosure risk in a Census context, this has a negative impact on the utility of the data. The SDC methods will be compared using quantitative disclosure risk and information loss measures which focus on the effects on statistical analysis (see: Shlomo, 2007 and references therein for more details). The aim is to strike a balance between managing disclosure risk while maximizing the amount of information that can be released to users. The analysis of the SDC methods will be demonstrated on one typical table selected from the 2001 UK Census.

It is well known that Census and register-based data have errors due to data processing, coverage adjustments, non-response and edit and imputation procedures, although much effort is devoted to minimizing these errors. When assessing disclosure risk, it is essential to take into account measurement errors and the

protection that is already inherent in the data. For example, a quantitative measure of disclosure risk should take into account the amount of imputation and adjust parameters of the SDC methods accordingly to be inversely proportional to the imputation rate. This ensures that the data are not overly protected causing unnecessary loss of information. It should be noted that once statistical results are disseminated, they are typically perceived and used by the user community as accurate counts.

SDC methods implemented at Statistical Agencies for Census tables include pre-tabular methods, post-tabular methods and combinations of both. Pre-tabular methods are implemented on the microdata prior to the tabulation of the tables. The most commonly used method is record swapping between a pair of households matching on some control variables (Willenborg and de Waal, 2001 and references therein). This method has been used for protecting Census tables at the United States Bureau of the Census and the Office for National Statistics (ONS) in the United Kingdom. Record swapping can be seen as a special case of a more general pre-tabular method based on a Post-Randomization Method (PRAM) (Gouweleeuw, Kooiman, Willenborg and De Wolf, 1998). This method changes values of categorical variables for a small number of records according to a prescribed probability matrix and a stochastic process based on the outcome of a random multinomial draw. PRAM can also be carried out in such a way as to ensure marginal distributions. In practice, Statistical Agencies prefer record swapping since the method is easy to implement and explain to users.

Post-tabular methods are implemented on the entries of the tables after they are computed and typically take the form of rounding, either on the small cells or on all entries of the tables. The method of small cell adjustments has been carried out on Census tables at the Australian Bureau of Statistics (ABS) and the UK ONS, and full random rounding has been carried out at Statistics Canada and Statistics New Zealand. A fully controlled rounding option has recently been added to the Tau-Argus SDC software package (Hundepool, 2002, Salazar-Gonzalez, Bycroft and Staggemeier, 2005), although this has yet to be implemented for full scale Census outputs. Tau-Argus also has cell suppression modules among which we implement the heuristic Hypercube Method (Giessing, 2004) in order to cope with the large Census tables. A new technique for cell perturbation is the Controlled Tabular Adjustment (CTA) (Dandekar and Cox, 2002) which involves “imputing” values for the suppressed cells under additivity and other constraints. This method is still under development and will not be considered further in this paper.

Section 2 describes the table which will be used to illustrate the disclosure risk-data utility assessment and Section 3 the SDC methods applied. In Section 4 we examine the quantitative disclosure risk and information loss measures that will be implemented and carry out an analysis of the SDC methods. A discussion and conclusions are presented in Section 5.

2 Table Description

We examine a typical table extracted from one estimation area (EA) of the unperturbed 2001 UK Census data. The table is disseminated by Output Areas (OA) which are the smallest Census tracts that are published for the UK Census. The number of OAs in the EA is 1,487 and includes on average about 125 households. For each OA, the table is defined as follows (the number of categories is given in parenthesis): Economic Activity (9) × Sex (2) × Long-Term Illness (2), i.e. a total of 36 categories. The table includes 317,064 individuals between the ages of 16 and 74 in 53,532 internal cells. The average cells size is 5.92 although the table is skewed with very large columns and very small columns. There are 17,915 (33.5%) zeros in the table and 14,726 (27.5%) cells with 1 or a 2.

3 SDC Methods

In this analysis, we will examine the following SDC methods:

3.1 Record Swapping

The most common pre-tabular method of SDC for frequency tables containing whole population counts is record swapping on the microdata prior to tabulation where variables are exchanged between pairs of households. In order to minimize bias, pairs of households are determined within strata defined by control variables, such as a large geographical area, household size and the age-sex distribution of the individuals in the households. In addition, record swapping can be targeted to high-risk households found in small cells of Census tables thereby ensuring that households that are most at risk for disclosure are likely to be swapped. In a Census context, geography variables are often swapped between households.

For this analysis, random record swapping was carried out on households from extracts of the 2001 UK Census at the following swapping rates: 10%, and 20%. The control variables that defined the strata were the number of persons in the household according to sex and three broad age groups and a "hard-to-count" index of the household based on the 1991 UK Census enumeration. The record swapping was carried out within a large geographical area (Local Authority) and households were swapped in and out of small geographical areas (Output Areas). In addition, targeted record swapping was carried out by defining an additional control variable based on a "flag" for the household that had at least one person in a small cell in a range of Census tables. On average, about 0.15% of the households selected for swapping were not swapped because no paired record was found for them. In general, those records would have to be swapped outside the large geographical area.

3.2 Rounding

The most common post-tabular method of SDC for Census tables is based on variations of rounding as follows:

Unbiased Random Rounding: Let $Floor(x)$ be the largest multiple k of the base b such that $bk < x$ for an entry x. In addition, define $res(x) = x - Floor(x)$. For an unbiased rounding procedure, x is rounded up to $(Floor(x) + b)$ with probability $\frac{res(x)}{b}$ and rounded down to $Floor(x)$ with probability $(1 - \frac{res(x)}{b})$. If x is already a multiple of b, it remains unchanged. Each cell is rounded independently in the table, i.e. a random uniform number u between 0 and 1 is generated for each cell. If $u < \frac{res(x)}{b}$ then the entry is rounded up, otherwise it is rounded down. As mentioned, the expectation of the rounding is zero and no bias should remain in the table. However, the realization of this stochastic process on a finite number of cells in a table may lead to overall bias since the sum of the perturbations (i.e. the difference between the original and rounded cell) going down may not equal the sum of the perturbations going up.

The method can be carried out on small cells only. In this case, margins of the tables are obtained by aggregating rounded and non-rounded cells, and therefore tables with the same population base will have different totals. While this provides ambiguity in the marginal totals, the users of Census tables generally object to inconsistent totals across tables. For full random rounding, margins are rounded separately from the internal cells because of the large number of perturbations and therefore tables are not additive.

The stochastic rounding methods are transparent and users can take the rounding into account when carrying out statistical analysis. The random rounding procedure (for all cells or only on small cells) is typically carried out independently for each cell based on a random draw, i.e. sampling with replacement. The algorithm however can be improved by preserving the stochastic unbiased properties but placing more control in the selection of the entries to round up or down. First the expected number of entries that are rounded up is predetermined (for the entire table or for each row/column of the table). Based on this expected number, a random sample of entries is selected (without replacement) and rounded up. The other entries are rounded down. This process ensures a bias of zero and the rounded internal cells aggregate to the controlled rounded total. For this analysis, we carried out the full random rounding to base 3 and to base 5 under the following methods: independent rounding in each cell and semi-controlled to the overall total. In addition, we assess the impact of combining the SDC methods of record swapping and random rounding with respect to disclosure risk and information loss in the Census table.

Controlled Rounding: We implemented the controlled rounding feature in Tau-Argus on the Census table. The procedure uses linear programming techniques to round entries up or down and in addition ensures that all rounded entries add up to the rounded totals. It should be noted that the method is not unbiased and cells can jump a base in order to meet the constraints of the program. We implemented the method to base 3 and base 5.

Cell Suppression: Cell Suppression in Tau-Argus for frequency tables is determined by a minimum threshold for identifying primary suppressions. Secondary suppressions are then chosen in order to avoid the recalculation of primary suppressions through the margins. For an optimal selection of secondary suppressions, a target function based on information loss is minimized within a linear program framework subject to constraints of pre-defined protection intervals for each suppressed cell. Because of the size of Census tables, we implemented the heuristic Hybercube method based on suppressing corners of a hypercube under optimality conditions. The minimum threshold for primary suppressions was 3.

4 Analysis of SDC Methods

4.1 Disclosure Risk

The main type of disclosure risk arises from small cells in tables (or small cells appearing in potential slithers of differenced tables) as well as the amount and placement of the zeros. This can lead to identification and attribute disclosure when many tables are disseminated from one database.

Pre-tabular methods of disclosure control, and in particular record swapping, will not prevent small cells and therefore a quantitative disclosure risk measure is needed which reflects whether the small counts in tables are true values. The quantitative disclosure risk measure for assessing the impact of record swapping is the proportion of records in small cells that have not been perturbed. The perturbation comes from two sources: record swapping and imputation. In general, imputed records can be viewed as protected records and therefore we need to take them into account in the quantitative risk measures. Imputation is typically carried out for item non-response, unit non-response and for Census coverage adjustments.

Let R_i represent the record i, I the indicator function having a value 1 if true and 0 if false, C_1 the set of cells with a value of 1, C_2 the set of cells with a value of 2, $n_{C_1 \cup C_2}$ the number of small cells with a value of 1 or 2. The disclosure risk measure is: $DR1 = \dfrac{\sum_{i \in C_1 \cup C_2} I(R_i \ \ not \ \ perturbed \ \ or \ \ imputed)}{n_{C_1 \cup C_2}}$. Table 1 presents

results of the disclosure risk measure *DR1* for the table in the analysis under record swapping.

Original	Random Swap		Targeted Swap	
	10%	20%	10%	20%
0.83	0.65	0.54	0.49	0.33

Table 1 Percentage of Records in Small Cells not Swapped or Imputed (*DR1*)

Based on Table 1, without any disclosure control method, imputation provides some protection to small cells: 17% of the records in small cells in the table had some imputation carried out. For both swapping rates (10% and 20%), lower levels of disclosure risk are obtained, especially if records to be swapped are targeted from among unique records. In general, the probability that a small cell is indeed a true value for random record swapping is about: 1-2×(swap rate). For example, for the 10% random record swapping in EA1, the probability of a true small value is approximately 0.8 (i.e., 1-2×0.10). The level of imputation was 0.17 and therefore we obtained a final probability of 0.65. The targeted record swapping at higher swapping rates gives better protection by lowering the probability of a true small value.

Post-tabular forms of rounding or cell suppression eliminate all small cells in the table and therefore disclosure risk is minimal with respect to attribute disclosure. In addition, in contrast to record swapping, the perception of disclosure risk is also minimal since no small cells appear in the tables.

Another disclosure risk measure comparable across all the SDC methods is the percentage of true zeros out of the total number of zeros (perturbed and not-perturbed) in the table. The more ambiguity introduced into the zero counts, the more the table is protected. Let C_0^{orig} be the number of true zero counts and C_0^{pert} the number of perturbed zero counts. The disclosure risk measure is defined as: $DR2 = \frac{C_0^{orig}}{C_0^{orig} + C_0^{pert}}$. Table 2 presents the *DR2* measures for the SDC methods evaluated on the Census table.

Record Swapping				Rounding		Cell Suppression
10%		20%		Base 3	Base 5	
Random	Targeted	Random	Targeted			
0.92	0.86	0.89	0.81	0.69	0.58	1.00

Table 2 Proportion of true zeros (*DR2*) in Table 4 of EA1

Based on Table 2, a zero in the table will be a true zero about 90% of the time for the record swapping whereas this proportion is greatly reduced with random or

controlled rounding to base 3 or base 5. The cell suppression does not introduce any ambiguity in the zero counts since these are not usually suppressed and the users know the true zeros.

Some forms of rounding can be deciphered by linking and differencing tables with common margins. To minimize this risk of disclosure, Statistics Agencies often disseminate only one set of geographies and variables, ensure minimum population thresholds and carry out auditing to evaluate the protection levels.

4.2 Information Loss

In this analysis we look at four main areas for measuring information loss: distortion to distributions, the impact on a measure of association (Cramer's V) for 2 dimensional tables, the impact on the variance of the cell counts and a "between" variance that is used in an ANOVA. All of the results are presented in Table 3.

When assessing information loss for cell suppression we need to implement a method of imputation for the suppressed cells which would typically be carried out by an average user. The simplest case would be to replace the suppressed cells by the average information loss in each row or column. More formally:

Let m_{ij} be a cell count in a two way table $i = 1,...,I$ rows and $j = 1,...J$ columns. Let marginal totals be defined as: $m_{i.}$ and $m_{.j}$. The margins appear in the table without perturbation unless they have a small value and are suppressed. In that case, we define the margin to take a value of 1 for the imputation scheme. Let z_{ij} be an indicator taking on the value of 1 if the cell was suppressed (primary or secondary) and a 0 otherwise. Each suppressed cell in row i is replaced by $\hat{y}_i = \dfrac{m_{i.} - \sum_{j=1}^{J} m_{ij}(1 - z_{ij})}{\sum_{j=1}^{J} z_{ij}}$. For example: Two cells are suppressed in a row where the known marginal total is 500. The total obtained by adding up non-suppressed cells is 400, and therefore the total information loss in the row is 100. Each of the two suppressed cells is replaced with a value of 50.

Information loss will be defined as follows:

- **Distance Metric**

We examine distortions to the internal and marginal cells of the Census table. Since the basic unit for most Census tables are small geographies, i.e. OAs, a measure of distortion at this level of geography is preferred. The distance metric between original and protected cells of the table (including zero cells) are calculated separately for each OA. The final utility measure is the overall average of the distance metric across the OAs.

Following the notation of Gomatam and Karr (2003), let D^k represent a table for OA k, $D^k(c)$ be the cell frequency c in the table and n_{OA} the number of OAs under analysis, i.e. $n_{OA} = 1{,}487$ in the Census table. We define the Hellinger's Distance metric as follows: $$HD(D_{orig}, D_{pert}) = \frac{1}{n_{OA}} \sum_{k=1}^{n_{OA}} \sqrt{\sum_{c \in k} \frac{1}{2} \left(\sqrt{D^k_{pert}(c)} - \sqrt{D^k_{orig}(c)} \right)^2}$$

In addition, we examine distortions to the marginal totals of the Census table for both rows and columns. Denote by M the margin under consideration, n_M the number of categories in the margin and N^l the total number of persons in the *l-th* category of margin M. The Hellinger's Distance metrics is:

$$HDM(N_{orig}, N_{pert}) = \sum_{l=1}^{n_M} \sqrt{\frac{1}{2} \left(\sqrt{N^l_{pert}} - \sqrt{N^l_{orig}} \right)^2}$$

- **Impact on Measures of Association**

A very important statistical tool that is frequently carried out on contingency tables is the Chi-Square test for independence based on the Pearson Chi-Squared Statistic χ^2 which tests the null hypothesis that the criteria of classification, when applied to a population, are independent. The Pearson Statistic for a two-dimensional table $i = 1,..,I$ and $j = 1,..,J$ is defined as: $\chi^2 = \sum_i \sum_j \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$ where under the null hypothesis of independence: $e_{ij} = \frac{n_{i.} \times n_{.j}}{n}$, $n_{i.}$ is the marginal row total and $n_{.j}$ is the marginal column total.

In order to assess the impact of the SDC methods on tests for independence, the Pearson statistic obtained from a perturbed contingency table is compared to the Pearson statistic obtained from the original contingency table. In particular, we focus on the measure of association, Cramer's V defined as: $CV = \sqrt{\frac{\chi^2 / n}{\min(I-1),(J-1)}}$.

The information loss measure is the percent relative difference:

$$RCV(D_{orig}, D_{pert}) = \frac{100 \times (CV(D_{pert}) - CV(D_{orig}))}{CV(D_{orig})} .$$

For this analysis, the rows of the table are the OAs and the columns of the table are the Economic Activity × Sex × Long-Term Illness indicator. It should be noted that random rounding rounds the margins separately from the internal cells and tables are not additive. Nevertheless, using a standard statistical package, the expected cell

frequency e_{ij} is calculated by aggregating internal cells and not obtained from known margins. A large Cramer's V represents a high level of association between the rows and the columns of the two-way table.

- **Impact on Variance of Cell Counts**

SDC methods impact on the variances that are calculated for estimates based on the frequency tables. The focus in this analysis is on the variance of the average cell count calculated at the OA level of geography in the table. The overall information loss measure is obtained by the percent difference between the average variance across all of the OAs for the original table and the same average variance for the perturbed table.

Let: $V(D_{orig}) = \frac{1}{n_{OA}} \sum_{k=1}^{n_{OA}} \frac{1}{n_k - 1} \sum_{c \in k} (D^k_{orig}(c) - \overline{D}^k_{orig})^2$ where n_k is the number of columns, i.e. $n_k = 36$ in the Census table, Let $V(D_{pert})$ be similarly calculated. The utility measure is the percent relative difference:

$$RDV(D_{orig}, D_{pert}) = \frac{100 \times (V(D_{pert}) - V(D_{orig}))}{V(D_{orig})}.$$

- **Impact on "Between" Variance**

We assess the impact of SDC methods on the goodness of fit criterion R^2 of a regression analysis or ANOVA and in particular on the "between" variance which is the numerator in R^2. For example, in an ANOVA, we test whether a continuous dependent variable has the same means across groupings defined by categorical explanatory variables. The goodness of fit criterion R^2 is based on a decomposition of the variance of the mean of the dependent variable. The total sum of squares *SST* can be broken down into two components: the "within" sum of squares *SSW* which measures the variance within groupings defined by explanatory variables and the "between" sum of squares *SSB* which measures the variance between the groupings. R^2 is the ratio of *SSB* to *SST*. By perturbing the statistical data, the groupings may lose their homogeneity, *SSB* becomes smaller, and *SSW* becomes larger. In other words, the means of each of the groupings are shrinking towards the overall mean. On the other hand, *SSB* may become artificially larger showing more association within the groupings than in the original variable.

We define information loss based on the "between" variance of a proportion on cell *c*: Let $P^k_{orig}(c)$ be a target proportion for cell *c* in OA *k*, i.e. $P^k_{orig}(c) = \frac{D^k_{orig}(c)}{\sum_{c \in k} D^k_{orig}(c)}$

and let $P_{orig}(c) = \frac{\sum_{k=1}^{n_{OA}} D^k_{orig}(c)}{\sum_{k=1}^{n_{OA}} \sum_{c \in k} D^k_{orig}(c)}$ be the overall proportion across all the OAs of the table. The "between" variance for the proportion is defined as: $BV(P_{orig}(c)) = \frac{1}{n_{OA} - 1} \sum_{k=1}^{n_{OA}} (P^k_{orig}(c) - P_{orig}(c))^2$ and the information loss measure is:

$$BVR(P_{pert}(c), P_{orig}(c)) = \frac{100 \times (BV(P_{pert}(c)) - BV(P_{orig}(c)))}{BV(P_{orig}(c))}.$$

For this analysis, we chose the proportion of full-time male students with no long-term illness.

Based on Table 3, the greatest impact on distortion to cell counts for both internal and marginal cells is the random rounding to base 5. Putting some control in the random rounding procedure seems to cause slight improvements. The full controlled rounding has less distortion to cell counts for both base 3 and base 5 due to its similarity to deterministic rounding. As expected, rounding to base 3 has less distortion compared to rounding to base 5. The cell suppression with the simple imputation method has the least distortion since marginal totals and original cell counts above the value of 3 (that are not secondary suppressed) are disseminated without any perturbation. Record swapping has less distortion to distributions compared to the rounding methods. The distortion is greater as the swapping rates increase. Targeted record swapping has larger distance metrics for the internal cells but not necessarily for the OA margins since records were swapped across OAs in both cases. It should be noted that for the margin based on sex, long-term illness and economic activity, the record swapping did not cause any distortion. This is likely due to the control variables that were used for the selection of swapping pairs. In general, there is more distortion when unique records are targeted for swapping. When combining a rounding procedure with record swapping, all distance metrics are higher. The increased distortion to distributions therefore needs to be weighed against the extra protection that record swapping may provide to Census tables by introducing ambiguity when differencing and linking tables.

Table 3 also demonstrates the loss in association and attenuation when swapping records across geographical areas. The two-way Census table in the example is now leaning towards independence since the counts are "flattening" out in the table (this is seen by the negative value of the *RCV* measure). With higher swapping rates the loss in association is more severe. Targeted record swapping has less impact on the loss of association compared to the random record swapping. We also see in Table 3 that the rounding procedures have the opposite effect. By eliminating small cells through the rounding procedures and introducing more zeros into the table, the level of association based on the observed cell counts has artificially increased. This effect

Method	HD	HDM (Cols.)	HDM (Rows)	RCV Original Cramer's V (0.121)	RDV Original Average Variance (188.3)	BVR Original Between Variance (0.000233)
RR Base 3	2.03	1.48	6.36	11.58	0.52	11.4
RR Base 3 (controlled to total)	2.04	2.27	5.19	11.88	0.54	13.1
Controlled Rounding Base 3	1.95	0.07	1.53	9.97	0.39	12.9
RR Base 5	3.02	3.39	9.87	27.52	1.64	36.6
RR Base 5 (controlled to total)	3.03	3.26	5.43	27.65	1.62	39.4
Controlled Rounding Base 5	2.58	0.09	3.20	26.93	1.27	34.5
Cell Suppression	0.42	0	0	0.22	-0.04	-0.64
Swap Random 10%	1.39	0	2.46	-3.65	-1.31	-4.82
Swap Random 20%	1.98	0	3.59	-6.27	-2.10	-8.25
Swap Targeted 10%	1.58	0	2.38	-1.93	-0.59	-3.49
Swap Targeted 20%	2.19	0	3.16	-4.37	-1.51	-7.61
Swap 10% and RR Base 3	2.53	2.17	6.90	10.39	-0.78	9.45
Swap 20% and RR Base 3	2.91	1.86	7.16	7.66	-1.57	5.60

Table 3 Results of Information Loss Measures on Census Table

however is less severe with the controlled rounding method. When combining rounding procedures with record swapping, there are opposing effects on Cramer's V and therefore the *RCV* is smaller compared to the *RCV* based on the rounding procedures alone.

These same conclusions are seen with respect to the impact on the variance of the average cell counts (*RDV)* and the "between" variance (*BVR*). We obtain a clear pattern of decreasing variances (as noted by the negative values) as higher swapping rates are introduced, i.e. the cell counts are "flattening", the variance has become smaller in the *RDV* and the proportions within OA groups are moving towards the overall proportion in the *BVR*. The targeted record swapping has slightly less reduction in the two variances compared to the random record swapping. As seen for

the analysis on the Cramer's V above, the opposite effect occurs with the rounding procedures and the two variances are increasing. There is more of an increase in the variances with semi-controlled rounding but less of an increase for the full controlled rounding. The impact on the variances when combining rounding procedures with record swapping depends on the direction and magnitude of the variances of each procedure separately, although it is clear that the opposing effects are cancelling out.

5 Discussion

In this analysis, we examine some common approaches to SDC for Census tabular outputs: pre-tabular methods based on variations of record swapping and post-tabular methods based on forms of rounding and cell suppression. In addition, we assessed the impact when combining SDC methods.

From this analysis, it was shown that using record swapping as a sole SDC method for Census tables results in high probabilities that small cells in tables are true values and can be identified. Targeted record swapping lowers disclosure risk but distance metrics show that there is more distortion to distributions. Higher swapping rates raise the level of protection but also cause more distortion to the data. The overall distortion on cell counts is higher with the rounding procedures compared to the swapping methods. Placing controls in the rounding procedure preserves additivity and causes less distortion to cell counts and therefore raises the utility of the tables. It should be noted that rounding procedures protect against the perception of disclosure risk compared to record swapping where the effects are hidden to users. Combining rounding with record swapping raises the level of protection but increases the loss of utility to the Census tables. For some statistical analysis, the combination of record swapping and rounding may balance to some degree opposing effects that the methods have on the utility of the tables. For example, record swapping "flattens" out cell counts, reduces measures of association and homogenizes the data while rounding procedures introduce more dependencies, increase measures of association and raise the levels of dispersion. These conclusions on the impact of record swapping and rounding procedures are consistent across all tables containing whole population counts and not just the specific Census table that was used for this analysis.

We have demonstrated in this paper how a Statistical Agency should carry out an assessment of SDC methods by examining both sides of the SDC decision problem: managing disclosure risk while maximizing the utility and quality of the outputs. The final decision on what SDC methods to employ depends on whether the disclosure risk is below tolerable thresholds and if the utility of the outputs meets the demands for "fit for purpose" data by the user community. SDC methods should be combined, adapted and modified in order to ensure higher utility in the outputs. A correct balance must be found between the use of non-perturbative transparent SDC methods and perturbative SDC methods which have hidden effects and introduce

bias that cannot be accounted for. Clear guidance and quality measures need to be disseminated with the Census tables in order to inform users of the impact of the SDC methods and how to analyze disclosure controlled statistical data.

Future dissemination strategies for Censuses will include more use of flexible table generating software where users can design and generate their own Census tables. Therefore, the development of SDC methods needs to be directed to these types of online dissemination strategies. Improved GIS systems may advance the research for developing SDC methods that protect nested geographies thus allowing more flexibility for online dissemination. Finally, Statistical Agencies are relying more on safe settings, remote access and license agreements to provide alternative SDC strategies which limit the access to the data, especially when dealing with highly disclosive Census sample microdata and Origin-Destination tables.

References

Dandekar, R.A. and Cox, L. (2002) Synthetic Tabular Data – an Alternative to Complementary Cell Suppression. *Unpublished manuscript*.

Giessing, S. (2004) Survey on Methods for Tabular Data Protection in Argus. In Domingo-Ferrer, J. and Torra, V. (eds.): *Privacy in Statistical Databases, LNCS 3050*, Springer-Verlag.

Gomatam, S. and Karr, A. (2003) Distortion Measures for Categorical Data Swapping. *Technical Report Number 131*, National Institute of Statistical Sciences.

Gouweleeuw, J., Kooiman, P., Willenborg, L.C.R.J., and De Wolf, P.P. (1998) Post Randomisation for Statistical Disclosure Control: Theory and Implementation. *Journal of Official Statistics, 14*, 463-478.

Hundepool, A. (2002) The CASC Project. In Domingo-Ferrer, J. (eds.): *Inference Control in Statistical Databases: From Theory to Practice, LNCS* 2316, Springer-Verlag.

Salazar-González, J. J., Bycroft, C. and Staggemeier, A.T. (2005) Controlled Rounding Implementation. *Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality*, Geneva.

Shlomo, N. (2007) Statistical Disclosure Control Methods for Census Frequency Tables. *International Statistical Review, Vol. 75*, Number 2.

Willenborg, L. and De Waal, T. (2001) *Elements of Statistical Disclosure Control in Practice*. Lecture Notes in Statistics, 155. New York: Springer-Verlag.

Differentially Private Marginals Release with Mutual Consistency and Error Independent of Sample Size

Cynthia Dwork, Frank McSherry, Kunal Talwar
Microsoft Research, Silicon Valley,
1065 La Avenida Mountain View, CA, 94043, USA.
{dwork,mcsherry,kunal}@microsoft.com

Abstract. We report on a result of Barak *et al.* on a privacy-preserving technology for release of mutually consistent multi-way marginals [1]. The result ensures *differential privacy*, a mathematically rigorous notion for privacy-preserving statistical data analysis capturing the intuition that essentially no harm can befall a respondent who accurately reports her data beyond that which would befall her should she refuse to respond, or respond completely inaccurately [7, 5].

In addition to differential privacy, the techniques described herein ensure consistency among released tables and, in many cases, excellent accuracy.

1 Introduction

In this note we describe a method developed by Barak *et al.* for privacy-preserving contingency table release [1]. This was part of an ongoing project in Microsoft Research on privacy-preserving data analysis, and relies on earlier project contributions both for the definition of privacy and for some of the techniques used. An important feature of this ongoing effort is that domain-specific knowledge or expertise is *not* required for ensuring privacy. Thus, using our techniques, one does not need to be an expert on the American population, or an expert on the availability of other databases produced by official and commercial parties, and so on, to "safely" compute and release useful statistics.

Our approach to privacy-preserving data mining has its roots in cryptography. In specifying a cryptographic primitive one must formally define what it means to break the primitive – intuitively, what is the adversary's goal? – and delineate to what resources – computataional power and auxiliary information – the adversary may have access.A rigorous pursuit of an *ad omnia* definition of privacy, taking into consideration auxiliary information, led to the discovery that (at least one natural formalization of) Dalenius' goal, to wit, that anything learnable about a respondent, given access to a statistical database, should be learnable without access to the database [3], is provably not achievable [5]. This led us to an alternative, but still

ad omnia, goal, *differential privacy*, which captures the intuition that essentially no harm can befall a respondent who accurately reports her data beyond what would befall her should she refuse to respond, or should she respond completely inaccurately [7, 5].

Since our language may be non-standard in the statistics community, we begin with an informal description of the problem, clarifying what *we* mean by the terms "contingency table" and "marginal."

1.1 Contingency Table Release

Informally, a contingency table is a table of counts. In the context of a census or other survey, we think of the data of an individual as a *row* in a database. For the present, each row consists of k bits describing the values of k binary attributes $a_1, \ldots, a_k$.[1] Formally, the contingency table is a vector in $\mathbb{R}^{2^k}$ describing, for each setting of the k attributes, the number of rows in the database with this setting of the attribute values.

Commonly, the contingency table itself is not released. Instead, for various sets of attributes, one releases the projection of the contingency table onto each such subset of the attributes, *i.e.*, the counts for each of the possible settings of the restricted set of attributes. These counts are called marginals, each marginal being named by a subset of the attributes. A marginal named by a set of j attributes, $j \leq k$, is called a *j-way* marginal. The data curator will typically release many sets of low-order marginals for a single contingency table, with the goal of revealing correlations between many different, and possibly overlapping, sets of attributes.

2 Differential Privacy

Many papers in the literature attempt to formalize Dalenius' goal by requiring that the adversary's prior and posterior views about an individual (*i.e.*, before and after having access to the statistical database) shouldn't be "too different," or that access to the statistical database shouldn't change the adversary's views about any individual "too much." Of course, this is clearly silly, if the statistical database teaches us anything at all. For example, suppose the adversary's (incorrect) prior view is that everyone has 2 left feet. Access to the statistical database teaches that almost everyone has one left foot and one right foot. The adversary now has a very different view of whether or not any given respondent has two left feet. Even when used correctly, in a way that is decidedly not silly, this prior/posterior approach suffers from definitional awkwardness [9, 8, 2].

The real difficulty in achieving Dalenius' goal is posed by what cryptographers call *auxiliary information*. This is any information available to the adversary/user

[1] Typically, attributes are non-binary. Any attribute with m possible values can be decomposed into $\log(m)$ binary attributes; see the discussion in Section 5.

other than what is in the statistical database. Statisticians worry about auxiliary information too: this is precisely what is exploited in a linkage attack.

Suppose we have a statistical database that teaches average heights of population subgroups, and suppose further that it is infeasible to learn this information (perhaps for financial reasons) any other way (say, by conducting a new study). Consider the auxiliary information "The radio talk show host Terry Gross is two inches shorter than the average Lithuanian woman." Access to the statistical database teaches Terry Gross' height. In contrast, someone without access to the database, knowing only the auxiliary information, learns much less about Terry Gross' height.[2]

This brings us to an important observation: Terry Gross did not have to be a member of the database for the attack described above to be prosecuted against her. This suggests a new notion of privacy: minimize the increased risk to an individual incurred by joining (or leaving) the database. That is, we move from comparing an adversary's prior and posterior views of an individual to comparing the risk to an individual when included in, versus when not included in, the database. This new notion is called *differential privacy.*

2.1 The Definition

In the sequel, the randomized function $\mathcal{K}$ is the algorithm applied by the curator (*e.g.*, census bureau) when releasing information. So the input is the data set, and the output is the released information, or *transcript.*

Recall that we think of a database as a set of rows. We say databases D_1 and D_2 *differ in at most one element* if one is a subset of the other and the larger database contains just one additional row.

Definition 1 *A randomized function $\mathcal{K}$ gives* ε-differential privacy *if for all data sets D_1 and D_2 differing on at most one element, and all $S \subseteq Range(\mathcal{K})$,*

$$\Pr[\mathcal{K}(D_1) \in S] \quad \leq \quad \exp(\varepsilon) \times \Pr[\mathcal{K}(D_2) \in S], \tag{1}$$

where the probability space in each case is over the coin flips of the mechanism $\mathcal{K}$.

A mechanism $\mathcal{K}$ satisfying this definition addresses all concerns that a respondent might have about accurately contributing her personal information: even if the participant removed her data from the data set, no outputs (and thus consequences of outputs) would become significantly more or less likely. For example, if the database were to be consulted by an insurance provider before deciding whether or not to insure a given individual, then the presence or absence of that individual's data in the database will not significantly affect her chance of receiving coverage.

[2] A rigorous impossibility result generalizes and formalizes this argument, extending to essentially any notion of privacy compromise [5].

Differential privacy is therefore an *ad omnia* guarantee. It is also a very strong guarantee, since it is a statistical property about the behavior of the mechanism and therefore is independent of the computational power and auxiliary information available to the adversary/user.

Remarks: *1. The parameter ε is public. The choice of ε is essentially a social question and is beyond the scope of our work. That said, we tend to think of ε as, say,* 0.01, 0.1, *or in some cases,* $\ln 2$ *or* $\ln\ 3$. *If the probability that some bad event will occur is very small, it might be tolerable to increase it by such factors as 2 or 3, while if the probability is already felt to be close to unacceptable, then an increase of $e^{0.01} \approx 1.01$ might be tolerable, while an increase of e, or even only $e^{0.1}$, would not.*

2. Definition 1 extends to group privacy as well (and to the case in which an individual contributes more than a single row to the database). A group of c participants might be concerned that their collective data might leak information, even when a single participant's does not. Using this definition, we can bound the dilation of any probability by at most $\exp(\varepsilon c)$, *which may be tolerable for small c. Of course, the point of the statistical database is to disclose aggregate information about large groups (while simultaneously protecting individuals), so we should expect privacy bounds to disintegrate with increasing group size.*

Differential privacy provides much stronger guarantees than other privacy definitions of which we are aware [11, 12, 13, 10] (see [1] for a discussion). Differential privacy also rules out the practice of publishing a random subsample of the database. For a given row x, consider two datasets D_1 and $D_2 = D_1 \setminus \{x\}$. A sampling procedure that conceivably chooses x when the dataset is D_1 can *never* choose x when the dataset is D_2, since $x \notin D_2$, yielding a zero in the denominator in 1. Even if the subsampled row is somewhat altered, its release could violate differential privacy.

Our techniques work best, that is, introduce the least relative error, when n, the size of the dataset, is large. This is because the amount of distortion introduced for protecting privacy depends only on the set of marginals requested, and not on n. However, privacy is *always* guaranteed. This is in contrast with the case of subsampling, where the unexpressed but implicit privacy guarantee depends on n.

Strong as it is, differential privacy is not an absolute guarantee of privacy. As we have seen, any statistical database with any non-trivial utility can be used to compromise privacy, even of people not in the database. However, in a society that has decided that the benefits of certain databases outweigh the costs, differential privacy ensures that only a limited amount of additional risk is incurred by participating in the (socially beneficial) databases.

3 Differentially Private Data Analysis

The *sensitivity* of a function f acting on a data set is defined in [7] as

Definition 2 [7]. *For $f : \mathcal{D} \to \mathbb{R}^d$, the* L_1-sensitivity *of f is*

$$\Delta f = \max_{D_1, D_2} \|f(D_1) - f(D_2)\|_1 \quad (2)$$

for all D_1, D_2 differing in at most one element.

Note that sensitivity is a property of the function alone, and is independent of the actual database held by the curator.

Intuitively, the sensitivity of the function f describes the degree of uncertainty that must be present in the released approximation to the value of f when applied to the specific database held by the curator, in order to hide the presence or absence of any individual. This is captured by Theorem 1 below, connecting sensitivity to the amount of noise that suffices to ensure ε-differential privacy.

Theorem 1 [7]. *For any $f : \mathcal{D} \to \mathbb{R}^d$, the addition of independent, symmetric, exponential noise with variance $2\sigma^2$ ensures $(\Delta f/\sigma)$-differential privacy.*

Remark: Taking $\sigma = \Delta/\epsilon$ ensures ϵ-differential privacy for a query of sensitivity Δ.

Note that application of the theorem yields a *process* for releasing statistics in general, and marginals in particular. The *process* ensures differential privacy. It makes no sense to speak of the privacy of a specific set of marginals that are released.

An application of Theorem 1 that is of particular interest in the context of contingency tables is to the class of *histogram* queries. A histogram query is an arbitrary partitioning of the domain of database rows into disjoint "cells," and the true answer is the set of counts describing, for each cell, the number of database rows in this cell. Although a histogram query with d cells may be viewed as d individual counting queries, the addition or removal of a single database row can affect the entire d-tuple of counts in at most one location (the count corresponding to the cell to (from) which the row is added (deleted); moreover, the count of this cell is affected by at most 1, so by Definition 2, every histogram query has sensitivity 1.

Since a contingency table is a histogram, this means that we can add independently generated noise proportional to ε^{-1} to each cell of the contingency table to obtain an ε-differentially private (non-integer and not necessarily non-negative) table. We will address the question of integrality and non-negativity later. For now, we simply note that any desired set of marginals can be computed directly from this noisy table, and consistency among the different marginals is immediate. A drawback of this approach, however, is that while the noise in each cell of the contingency table is relatively small, the noise in the computed marginals may be large. For example, the variance in the 1-way table describing attribute a_1 is $2^{k-1}\varepsilon^{-2}$. We consider this unacceptable, especially when $n \ll 2^k$.

Marginals are also histograms. A second approach, with much less noise, but not offering consistency of marginals, works as follows. Let C be the set of marginals to

be released. We can think of a function f that, when applied to the database, yields the desired marginals. Now apply Theorem 1 with this choice of f, (adding noise to each cell in the collection of tables independently), as directed in Theorem 1, with sensitivity $\Delta f = |C|$. When n (the number of rows in the database) is large compared to $|C|\varepsilon^{-1}$, this also yields excellent accuracy. Thus we would be done, and there would be no need for the Barak *et al.* paper, if the small table-to-table inconsistencies caused by independent randomization of each (cell in each) table are not of concern, and if the user is comfortable with occasionally negative and typically non-integer cell counts.

We have no philosophical or mathematical objection to these artifacts of the privacy-enhancing technology, but in practice they can be problematic. For example, the cell counts may be used as input to other, possibly off-the-shelf, programs that anticipate positive integers, giving rise to type mismatch. It may also be confusing to lay users, say, ordinary citizens accessing the American FactFinder website.

4 Release of Marginals

The material in this Section appears in [1]. We first highlight key elements of the approach, then introduce formal notation, and finally state the results. Proofs may be found in the original paper.

4.1 Key Steps in The Solution

Apply Theorem 1 and Never Look Back. We *always* obtain privacy by applying Theorem 1 to the raw data or a possibly reversible transformation of the raw data. This gives us an intermediate object, on which we operate further, but we never again access the raw data. Since anything obtained via Theorem 1 is differentially private, any quantity computed from the intermediate object also enjoys differential privacy.

Move to the Fourier Domain. When adding noise, two natural solutions present themselves: adding noise to entries of the source table (this was our first proposal; accuracy is poor when k is large), or adding noise to the reported marginals (our second proposal; consistency is violated). A third approach begins by transforming the data into the Fourier domain. This is just a change of basis. Were we to compute all 2^k Fourier coefficients we would have a non-redundant encoding of the entire consistency table. If we were to perturb the Fourier coefficients and then convert back to the contingency table domain, we would get a (different, possibly non-integer, possibly negative) contingency table, whose "distance" (for example, L_2 distance) from the original is determined by the magnitude of the perturbations. The advantage of moving to the Fourier domain is that if only a set C of marginals is desired then we do not need the full complement of Fourier coefficients. For example,

if C is the set of all 3-way marginals, then we need only the Fourier coefficients of weight at most 3 (see Section 4.4), of which there are $\binom{k}{3} + \binom{k}{2} + k + 1$. This will translate into a much less noisy set of marginals.

The Fourier coefficients needed to compute the marginals C form a model of the dataset that captures everything that can be learned from the set C of marginals. Adding noise to these coefficients as indicated by Theorem 1 and then converting back to the contingency table domain yields a procedure for generating synthetic datasets that ensures differential privacy and yet to a great (and measurable) extent captures the information in the model. This is an example of a concrete method for generating synthetic data with provable differential privacy.

Strictly speaking, we don't really need to move to the Fourier domain: we can perturb the marginals directly and then use linear programming to find a positive fractional data set, which can then be rounded. See [1] for a discussion.

Use Linear Programming We employ linear programming to obtain a non-negative, but likely non-integer, data set with (almost) the given Fourier coefficients, and then round the results to obtain an integer solution. Interestingly, the marginals obtained from the linear program are no "farther" (made precise below) from those of the noisy measurements than are the true marginals of the raw data. Consequently, the additional error introduced by the imposition of consistency is no more than the error introduced by the privacy mechanism itself.

When k is Large The linear program requires time polynomial in 2^k. When k is large this is not satisfactory. However, somewhat surprisingly, non-negativity (but not integrality) can be achieved by adding a relatively small amount to the first Fourier coefficient before moving back to the data domain. No linear program is required, and the error introduced is pleasantly small. Thus if polynomial in 2^k is an unbearable cost and one can live with non-integrality then this approach serves well. (In this case we construct the output marginals directly from the Fourier coefficients, rather than reconstructing the contingency table. See [1] for additional details.) We remark that non-integrality was a non-issue in the prototyped system mentioned above, since answers were anyway converted to percentages.

4.2 Notation and Preliminaries

For all positive integers d, $\forall x \in \mathbb{R}^d$, the L_1 norm of x is $\|x\|_1 = \sum_{i=1}^{d} |x_i|$. As noted above, and letting k denote the number of (binary) attributes, we think of the data set as a vector $x \in \mathbb{R}^{2^k}$, indexed by attribute tuples. For each $\alpha \in \{0,1\}^k$ the quantity x_α is the number of data elements with this setting of attributes. We let $n = \|x\|_1$ be the total number of tuples, or rows, in our data set.

For any $\alpha \in \{0,1\}^k$, we use $\|\alpha\|_1$ for the number of non-zero locations. We write $\beta \preceq \alpha$ for $\alpha, \beta \in \{0,1\}^k$ if every zero location in α is also a zero in β.

4.3 The Marginal Operator

We think of the computation of a set of marginals as the result of applying a *marginal operator* to the contingency table vector x. The operator $C^\alpha : \mathbb{R}^{2^k} \to \mathbb{R}^{2^{\|\alpha\|_1}}$ for $\alpha \in \{0,1\}^k$ maps contingency tables to the marginal of the attributes that are positively set in α (there are $2^{\|\alpha\|_1}$ possible settings of these attributes). We abuse notation, and only define $C^\alpha x$ at those locations β for which $\beta \preceq \alpha$: for any $\beta \preceq \alpha$, the outcome of $C^\alpha x$ at position β is the sum over those coordinates of x that agree with β on the coordinates described by α:

$$(C^\alpha(x))_\beta \quad = \quad \sum_{\gamma:\gamma\wedge\alpha=\beta} x_\gamma \tag{3}$$

Notice that the operator C^α is linear for all α.

It is common to consider the ensemble of marginal operators C^α for all α with a fixed value of $\|\alpha\|_1 = j$, referred to as the j-way marginals. For example, when $j = 3$ this is the ensemble of marginal operators C^α for all $\alpha \in \{0,1\}^k$ containing exactly 3 ones, *i.e.*, the ensemble of all 3-way marginals.

4.4 The Fourier Basis

We will find it helpful to view our contingency table x in an alternate basis; rather than a value for each position α, we will project onto a set of 2^k so-called *Fourier basis* vectors, each of which aggregates across the table in a different way. Our motivation lies in the observation, made formally soon, that while a marginal depends on all coordinates of the contingency table, a low-order marginals (that is, C^α when $\|\alpha\|_1$ is small) depends on only a few of the new coordinates in the Fourier basis.

The Fourier basis for real vectors defined over the Boolean hypercube is the set of vectors f^α for each $\alpha \in \{0,1\}^k$, defined coordinate-wise as

$$f^\alpha_\beta \quad = \quad (-1)^{\langle\alpha,\beta\rangle}/2^{k/2} \, . \tag{4}$$

That is, each vector comprises coordinates of the form $\pm 1/2^{k/2}$, with the sign determined by the parity of the intersection between α and β.

The following theorem is well known.

Theorem 2 *The f^α form an orthonormal basis for $\mathbb{R}^{2^k}$.*

The projection $\langle f^\alpha, x\rangle f^\alpha$ of a vector x onto a Fourier basis vector f^α is referred to as a Fourier coefficient. The following theorem says that any marginal over few attributes requires only a few Fourier coefficients.

Theorem 3 *$C^\beta f^\alpha \neq \boldsymbol{0}$ if and only if $\alpha \preceq \beta$.*

Consequently, we are able to write any marginal as the small summation over relevant Fourier coefficients:

$$C^\beta x \quad = \quad \sum_{\alpha\preceq\beta} \langle f^\alpha, x\rangle C^\beta f^\alpha \, . \tag{5}$$

4.5 Algorithms

To apply Theorem 1 in the Fourier domain, we need only bound the sensitivity of each Fourier coefficient, since a straightforward argument shows that the sensitivity of a collection of coefficients is bounded by the number of coefficients in the collection times the sensitivity of any one coefficient. Formally, let $Lap(\sigma)$ be a random variable with density at x proportional to $\exp(-|x|/\sigma)$.

Theorem 4 *Let $B \subseteq \{0,1\}^k$ describe a set of Fourier basis vectors. Releasing the set $\phi_\beta = \langle f^\beta, x\rangle + Lap(|B|/\epsilon 2^{k/2})$ for $\beta \in B$ preserves ϵ-differential privacy.*

Proof: Each tuple contributes exactly $\pm 1/2^{k/2}$ to each output coordinate, and consequently the L_1 sensitivity of the set of $|B|$ outputs is at most $|B|/2^{k/2}$. By Theorem 1, the addition of symmetric exponential noise with standard deviation $|B|/\epsilon 2^{k/2}$ gives ϵ-differential privacy. QED.

Remark: To get a sense of scale, we could achieve a similar perturbation to each coordinate by randomly adding or deleting $|B|^2/\epsilon$ individuals in the data set, which can be much smaller than n.

4.5.1 Non-Negative Integrality

Consider the set of (now noisy) Fourier coefficients $\{\phi_\beta \,|\, \beta \in B\}$ released in Theorem 4. While there certainly exists a real valued contingency table whose Fourier coefficients equal these released values, it is unlikely that there is a non-negative, integral contingency table with these coefficients. We use linear programming to find a non-negative, but likely fractional, contingency table with nearly the correct Fourier coefficients, which we round to an integral table with little additional error.

Imagining that we observed (noisy) Fourier coefficients ϕ_β, the linear program in the next section minimizes, over all contingency tables w, the largest error b between its Fourier coefficients $\langle f^\beta, w\rangle$ and the observed ϕ_β's.

4.5.2 Putting the Steps Together

We now collect the various steps. Recall that to compute a marginal $\alpha \in \{0,1\}^k$ we require the Fourier coefficients $\{f^\beta \,|\, \beta \preceq \alpha\}$. Thus, to compute a set A of marginals, we need all the Fourier coefficients f^β for β in the downward closure of A uner $\preceq$.

Marginals$(A \subseteq \{0,1\}^k, D)$:

1. Let B be the downward closure of A under $\preceq$.
2. For $\beta \in B$, compute $\phi_\beta = \langle f^\beta, D\rangle + Lap(|B|/\epsilon 2^{k/2})$.

3. Solve for w_α in the following linear program, and round to the nearest integral weights, w'_α.

$$\begin{array}{rcll} \text{minimize} & b & & \\ \text{subject to:} & & & \\ w_\alpha & \geq & 0 & \forall \alpha \\ \phi_\beta - \sum_\alpha w_\alpha f_\alpha^\beta & \leq & b & \forall \beta \in B \\ \phi_\beta - \sum_\alpha w_\alpha f_\alpha^\beta & \geq & -b & \forall \beta \in B \end{array}$$

4. Using the contingency table w'_α, compute and return the marginals for A.

Theorem 5 *Using the notation of* ***Marginals***(A)*, with probability* $1-\delta$*, for all* $\alpha \in A$*,*

$$\|C^\alpha x - C^\alpha w'\|_1 \leq 2^{\|\alpha\|_1} 2|B| \log(|B|/\delta)/\epsilon + |B| \,. \tag{6}$$

Why the *Fourier* Basis? The Fourier coefficients exactly describe the information required by the marginals. By measuring exactly what we need, we add the least amount of noise possible using the techniques of [7]. Moreover, the Fourier basis is particularly attractive because of the natural decomposition according to sets of attribute values. In particular, even tighter bounds than in Theorem 5 can be placed on sub-marginals (that is, lower order marginals) of a marginal C^α, by noting that the bounds for the marginals C^β, where $\beta \preceq \alpha$, are obtained at no additional cost. No more Fourier coefficients are required, so $|B|$ is not increased, but $\|\beta\|_1 \leq \|\alpha\|_1$.

4.5.3 Simple Non-Negativity

The solution of the linear programs we have described is an expensive process, taking time polynomial in 2^k. In many settings, but not all, this is an excessive amount, and must be avoided. We now describe a very simple technique for arriving at Fourier coefficients corresponding to a non-negative, but fractional, contingency table with high probability, without the solution of a linear program. As noted above, we construct the output marginals directly from the Fourier coefficients, rather than reconstructing the contingency table.

Ensuring the existence of a non-negative contingency table with the observed Fourier coefficients turns out to be easy: we simply add a small amount to the first Fourier coefficient. Intuitively, any negativity due to the small perturbation we have made to the Fourier coefficients is spread uniformly across all elements of the contingency table. Consequently, very little needs to be added to make the elements non-negative.

Theorem 6 *Let x be a non-negative contingency table with d Fourier coefficients ϕ_α. If the Fourier coefficients are perturbed to ϕ'_α, then the contingency table*

$$x' \;=\; x + \sum_\alpha (\phi'_\alpha - \phi_\alpha) f^\alpha + \|\phi' - \phi\|_1 f^{\overline{0}} \qquad (7)$$

is non-negative, and has $\langle f^\alpha, x' \rangle = \phi'_\alpha$ for $\alpha \neq \overline{0}$.

It is *critical* that we not disclose the actual L_1 norm of the perturbation, but we can add a value for which the negativity probability is arbitrarily low:

Corollary 1 *By adding $t \times d^2/\epsilon 2^{k/2}$ to the first Fourier coefficient, the resulting contingency table is non-negative with probability at least $1 - \exp(-t)$.*

5 Discussion

Non-Binary Attributes. We have assumed that attributes are binary-valued. While it is possible to convert a non-binary valued attribute to vector of binary attributes, this introduces some inefficiency, for example, we need 2 bits to describe a three-valued attribute, and 5 bits to describe a 17-valued attribute. The linear program can be modified to force an outcome of 0 in these "structurally zero" locations, but for each such attribute we may be increasing the size of the linear program by a factor of almost 2. Even without linear programming, some care must be taken to redistribute any non-zero weight assigned to these locations (after adding noise to the Fourier coefficients and converting back to the space of marginals) to the "real" locations.

When the number of attributes is large, the error introduced in each cell in the set of all low-order marginals, although independent of n, the size of the population, is still substantial. For example, in the case of the set of all 3-way marginals, we get a value in $\Omega(k^3)/\varepsilon$, even if the attributes are binary. When k is on the order of 30, 50, or 100 the distortion is on the order of thousands or (a small number of) millions. There are several possible approaches to improving the accuracy.

Gaussian Noise. First, we can use a Gaussian distribution on noise. The analysis in this case is more involved, and the nature of the privacy guarantee is slightly different. Instead of ε-differential privacy we obtain (ε, δ)-differential privacy (see [6], where this is called *δ-approximate ε-differential privacy*). Roughly speaking, this says that, with all but probability δ, ε-differential privacy is ensured. Using Gaussian noise with distribution $G(x) \propto \exp(-x^2/2R)$ we get (ε, δ) differential privacy when $\varepsilon \geq [2\log(1/\delta)/R]^{1/2}$. The advantage of this approach is that it allows us to improve the dependence on k of the distortion in each cell to $O(k^{3/2})$.

Partitioning into "Sensitive" and "Insensitive" Attributes. The literature frequently discusses "sensitive" and "insensitive" attributes. We prefer to avoid such distinctions; they are often flawed and they rely on domain-specific information. However, there is a way to exploit the distinction, were one to accept its validity: only add noise to Fourier coefficients in the downward closure of any requested marginal α containing at least one sensitive attribute. For example, suppose only attribute a_1 is sensitive. If the requested set C is the ensemble of all 3-way marginals, we need only add noise to $\binom{k-1}{2} + \binom{k}{2} + k + 1$ Fourier coefficients, rather than to all $\binom{k}{3} + \binom{k}{2} + k + 1$, and we may scale down the magnitude of the noise accordingly.

Lower Bounds on Noise. Finally we remark that, at least theoretically, there is no way to avoid the dependence on k in the presence of even one sensitive attribute. In the admittedly contrived setting in which there are $k = n \log^2 n + 1$ attributes, and (for some reason) for each tuple independently, for each attribute independently, the attribute is set with probability $1/2$, it is possible to compute, in time polynomial in n, a candidate vector c that agrees with the vector v of values of the secret, kth, attribute, in all but $o(n)$ locations, assuming only that the magnitude of the noise added to each marginal is $o(\sqrt{n})$ [4]. If the adversary is not restricted to polynomial time then the attack works whenever the noise in each marginal is $o(n)$.

References

[1] B. Barak, K. Chaudhuri, C. Dwork, S. Kale, F. McSherry, and K. Talwar. Privacy, accuracy, and consistency too: a holistic solution to contingency table release. In L. Libkin, editor, *PODS*, pages 273–282. ACM, 2007.

[2] A. Blum, C. Dwork, F. McSherry, and K. Nissim. Practical privacy: the SuLQ framework. In C. Li, editor, *PODS*, pages 128–138. ACM, 2005.

[3] T. Dalenius. Towards a methodology for statistical disclosure control. *Statistisk. tidskrift*, 3:213–225, 1977.

[4] I. Dinur and K. Nissim. Revealing information while preserving privacy. In *PODS '03: Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 202–210, New York, NY, USA, 2003. ACM.

[5] C. Dwork. Differential privacy. In M. Bugliesi, B. Preneel, V. Sassone, and I. Wegener, editors, *ICALP (2)*, volume 4052 of *Lecture Notes in Computer Science*, pages 1–12. Springer, 2006.

[6] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor. Our data, ourselves: Privacy via distributed noise generation. In S. Vaudenay, editor, *EUROCRYPT*, volume 4004 of *Lecture Notes in Computer Science*, pages 486–503. Springer, 2006.

[7] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In S. Halevi and T. Rabin, editors, *TCC*, volume 3876 of *Lecture Notes in Computer Science*, pages 265–284. Springer, 2006.

[8] C. Dwork and K. Nissim. Privacy-preserving datamining on vertically partitioned databases. In M. K. Franklin, editor, *CRYPTO*, volume 3152 of *Lecture Notes in Computer Science*, pages 528–544. Springer, 2004.

[9] A. Evfimievski, J. Gehrke, and R. Srikant. Limiting privacy breaches in privacy preserving data mining. In *PODS '03: Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 211–222, New York, NY, USA, 2003. ACM.

[10] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. In L. Liu, A. Reuter, K.-Y. Whang, and J. Zhang, editors, *ICDE*, page 24. IEEE Computer Society, 2006.

[11] P. Samarati and L. Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression.

[12] L. Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):557–570, 2002.

[13] X. Xiao and Y. Tao. M-invariance: towards privacy preserving re-publication of dynamic datasets. In C. Y. Chan, B. C. Ooi, and A. Zhou, editors, *SIGMOD Conference*, pages 689–700. ACM, 2007.

A Measure of Disclosure Risk for Aggregate Data

Duncan Smith[1], Mark Elliot[1]

[1] Confidentiality and Privacy Group, Group,Cathie Marsh Centre for Census and Survey Research, University of Manchester, England, M13 9PL{duncan.g.smith, M.Elliot}@man.ac.uk

Summary. The paper describes a new method for assessing disclosure risk for tables of counts; the *subtraction - attribution probability* (SAP) method. The SAP score is the probability of an intruder recovering a 'risky' subpopulation table given a quantity of information about the individuals in a population table. The method can be applied to exact or perturbed individual tables and sets of tables. The method can also be used to compare the risk impact of different disclosure control regimes.

1 Introduction

Releases of population data can be used by so-called data intruders to glean sensitive information about individuals in the population. Disclosure occurs when a data intruder makes reliable inferences (i.e., with a high degree of confidence) about one or more population units. Statistical agencies need to guard against disclosure in order to meet their legal obligations to safeguard respondent confidentiality and to maintain public trust. Lack of trust can result in individuals refusing to complete, for example, census forms or returning forms with false or missing information. Most statistical agencies are mainly concerned with the risk of an intruder identifying a population unit, although this is not a requirement for disclosure of information about the individual concerned.

The need for appropriate measures of disclosure risk has been well discussed. Many authors have indicated that such measures should as far as possible take a data intruder's perspective of the risk (see e.g. Paass 1988, Mokken et al 1992, Elliot and Dale 1999). Although intruder-based measures have been established for identification risk (Skinner and Elliot 2002), little progress has been made with generating appropriate risk metrics for the actual disclosure of information about members of the population in the absence of identification. This paper describes the "subtraction - attribution probability" (SAP) method which attempts to fill this gap.

2 Disclosure Risk

Understanding of disclosure risk has evolved over the last twenty years and there is still no unequivocal definition of the term. However, definitions of disclosure generally involve one or both of *Identification* (A one to one association between a

data unit and a target.) and *attribution* (The association of one or more variable values with a target).

Herein, a data unit is an individual or organisation contained in microdata or tabulated data that is available to a so-called data intruder; a target is an individual or organisation about which a data intruder is trying to discover information.

In some cases it is possible for an intruder to perform identification or attribution with absolute certainty. In these cases the identification or attribution is termed exact. Otherwise, identification or attribution is termed approximate. Strictly speaking there will almost always be a degree of uncertainty regarding the correctness of the data, so all inferences are approximate. However, this source of uncertainty is generally ignored for disclosure risk assessment purposes, and we follow this practice here.

In this paper we address the risk of exact attribution. Previous papers have tended to concentrate on attribution stemming from identification. Fellegi (1972) considers disclosure in terms of "sufficiently narrowly defined" populations, and goes on to state that such a population may "contain only one identifiable respondent or, at least, information can be deduced from the published estimates that can be related to a particular identifiable respondent". He then goes on to illustrate how disclosure can occur from the conditioning on known information about a target, the conditional frequency table containing only the target individual. Clearly, if an intruder can achieve this by conditioning on a subset of the variables in the tabulation, then the levels of all the remaining variables can be discovered. If the levels of the discovered variables were previously unknown to the intruder, then disclosure has taken place. Fellegi also considers that conditioning to a (sub-) population of size two can result in similar disclosure if the intruder is the other member of the conditional population. A U.S. Department of Commerce report (1978) expands this idea by considering "coalitions" of individuals within a data set who might cooperate in order to discover new information about targeted individuals. The report also considers how disclosure can take place without the requirement for identification. Their examples are reproduced below.

		Race		
County	White	Black	Other	Total
A	15	20	5	40
B	0	30	0	30

Table 2.1. Number of beneficiaries by count and race.

In Table 2.1 conditioning on a target being a resident of County B implies that the target is black. A risk of such exact disclosure exists if a marginal total (in dimension n-1) equals one of its detail cells (in dimension n). This contrasts with the example given by Fellegi which required also that the detail cell count be 1.

The U.S. Department of Commerce report contrasts this with the case when the sum of a proper subset of detail cells equals the total in the relevant margin (Table 2.2). The report does not define the implication that a target in County B is either Black or Other as disclosure, because the subset of Black or Other is not as narrowly defined as possible. Similarly, the report authors do not consider exact inferences regarding age as disclosive unless age is revealed to within a single year.

	Race			
County	White	Black	Other	Total
A	15	20	5	40
B	0	28	2	30

Table 2.2. Number of beneficiaries by count and race.

We consider this distinction to be fairly arbitrary as ethnicity can be broken down into more detailed classifications than those of the example, and any categorisation of a continuous variable such as age will involve ranges that are not as narrowly defined as possible. One approach would be to associate sensitivities with any set/range of variable levels and consider disclosure to have taken place if the sensitivity of the discovered information exceeds some predefined threshold. However, data is often collected with an unqualified assurance of confidentiality, so that it is arguable that all data should be regarded as sufficiently sensitive to warrant protection. Therefore, for the purposes of this paper we will simply consider that disclosure takes place when an intruder, by whatever means, is able to condition to a population table which contains one or more zeros. This definition encompasses the two cases illustrated above, and the additional case where an intruder can infer that a particular combination of attribute levels does not apply to a target. For example, simply conditioning on a target being a member of the population in Table 2 allows the intruder to infer that the target is not White and residing in County B (although either are individually possible). So in this strict sense, we consider a risk of disclosure to be present if a population table contains one or more zeros.

Skinner (1992) defines disclosure in the sense of Fellegi's example (requiring identification and attribution) as identification disclosure, whereas disclosure that does not require identification is prediction disclosure. He considers approximate disclosure in sample tables and develops an argument that identification disclosure is a necessary and sufficient condition for prediction disclosure. In this paper we are

concerned only with the risk of exact disclosure in population tables. Under these circumstances it is clear that identification is neither necessary nor sufficient for attribution (prediction disclosure).

2.1 Attribution risk from low population counts

Heretofore, we have only considered the risks of attribution as they stem from conditioning on known information relating to some targeted individual. It is implicit that the intruder is also conditioning on a target being a member of the population. But conditioning on known variable levels (or known absence of variable levels) is not the only way an intruder might attempt to condition down to a smaller, more disclosive population. The U.S. Department of Commerce report describes the possibility of disclosure stemming from coalitions, the main questions arising regarding the likely size of coalition, and the distribution of the coalition within the population. However, we note that the type of disclosure that can arise from coalitions does not require their existence. It is possible for an intruder to hold information on a number of population units, without their explicit cooperation. If they can be identified within the population, then their records can be removed from the data set, facilitating inferences regarding the residual subpopulation. Removal of a unique clearly leads to the presence of a zero, and a risk of attribution. Of course, records of known individuals may be removed without identification, and partially known individuals might be 'removed' from the relevant margins, placing constraints on the counts in the full cross-classification of the residual population. In essence, an intruder can use arbitrary known information about the population units in order to try to facilitate attribution. Lower counts represent a greater risk of the recovery of zeros by subtraction of known individuals.

The above requires information that can be considered external to the data set in question, and as such might not be considered an overriding issue. However, any inferences regarding a population unit require such information. Both exact identification and exact attribution require external information; at the very least an intruder must be able to condition on a target being a member of the relevant population.

2.2 Protection against attribution

Statistical agencies tend to guard against disclosure by suppressing (withholding) data or disguising the true counts by deterministic or stochastic perturbations; (see Duncan et al (2001) for a review). For example, one deterministic method is conventional rounding. A suitable non-negative odd integer is chosen as base, and each count in the cross-classification is rounded to the closest multiple of the base.

Figure 2 contains the conventionally rounded, to base 3, cross-classifications corresponding to the exact cross-classification in Figure 1.

		VAR2			
		D	E	F	
VAR1	A	1	3	0	4
	B	4	0	0	4
	C	3	2	0	5
		8	5	0	13

Fig 2.1. A 2-way cross-classification with margins.

		VAR2			
		D	E	F	
VAR1	A	0	3	0	3
	B	3	0	0	3
	C	3	3	0	6
		9	6	0	12

Fig 2.2 Conventionally rounded cross-classification.

An intruder (with knowledge of the rounding scheme) can easily generate bounds on the counts in the exact 2-way cross-classification, given the corresponding rounded cross-classification. We term these *trivial* bounds as they are based solely on the rounding scheme.

		VAR2		
		D	E	F
VAR1	A	0	2	0
	B	2	0	0
	C	2	2	0

Fig 2.3 Trivial lower bounds of figure 2.2.

		VAR2		
		D	E	F
VAR1	A	1	4	1
	B	4	1	1
	C	4	4	1

Fig 2.4 Trivial upper bounds of figure 2.2.

Here the rounding has managed to disguise the exact value of all counts. But subtraction of a known individual in cell (A,D) would recover a zero.

It is not unusual for statistical agencies releasing perturbed cross-classifications to also release perturbed, or occasionally exact, marginal tables. The presence of marginal counts places a system of linear constraints on the counts in the full (in this case 2-way) cross-classification. Solving the system of constraints via integer linear programming methods can lead to tighter bounds than those derived solely from a full rounded cross-classification. Dobra (2002) develops a method for solving cell bounds given marginal cell counts. Although his algorithm is designed to deal with exact cross-classifications it is relatively easily extended for dealing with perturbed counts (Smith and Elliot, 2003). The release of all the rounded cross-classifications

(including both 1-way margins and rounded total) in Figure 2.2 results in the following lower and upper bounds.

		VAR2		
		D	E	F
VAR1	A	0	2	0
	B	3	0	0
	C	3	2	0

Fig 2.5 Non-trivial lower bounds of figure 2.2.

		VAR2		
		D	E	F
VAR1	A	1	3	0
	B	4	1	0
	C	4	3	0

Fig 2.6 Non-trivial upper bounds of figure 2.2.

Three of the four zeros have been recovered. This stems from the fact that the trivial lower bounds for the VAR2 margin sum to 13, which is the trivial upper bound for the rounded total. Thus the total and VAR2 margin are recovered exactly. So the perturbation of the data has done little to remove the risk of attribution. Subtraction of individuals could increase the risk still further.

2.3 A measure of attribution risk

A risk of attribution exists if, and only if, one or more zeros exist in some population cross-classification. The population cross-classification in question need not necessarily have been released. In fact, it is possible to construct examples where the

exact counts in a 3-way cross-classification can be recovered from its three distinct 2-way margins. In any case, a set of population cross-classifications can be used to place bounds on any cross-classification from which they could be derived. It is enough to consider only the 'base' cross-classification with axes corresponding to the union of the variables in the released cross-classifications. Any cross-classifications over a superset of the variables in the base cross-classification contain (recovered) zeros if, and only if, the base cross-classification contains (recovered) zeros. Bounds on smaller margins can be solved, but again this is unnecessary, as any zero in a margin implies zeros in the full cross-classification.

Given the questionable distinction between inferences on the basis of the 'narrowness of definition' we propose a measure based simply on the presence of zeros in the full population cross-classification. Sensitivities are not considered for the reason given earlier, although we note that the methodology can be applied to conditional tables as easily as marginal tables, in which case we could assess risk for given population units or population cells given an assumed set of *key* variables. We also wish to take into account the additional risk stemming from intruder knowledge of the population, and to be able to apply the measure to relatively arbitrary releases of exact and / or perturbed cross-classifications. Specifically, our chosen measure is the 'probability of recovering one or more zeros in the full cross-classification given the subtraction of a random sample of n population units'. We term this the subtraction attribution probability (SAP).

Assume we have a base table of counts of arbitrary dimension with cell counts c_i, i = 1 to m. Assume that an arbitrary set of perturbed marginal tables is published, each perturbed using some independent rounding scheme (i.e. each cell is perturbed independently of the others). Then each published count, x, implies a pair of constraints of the form, $l \le c$, $c \le u$, where l and u are the trivial bounds implied by the rounding scheme and c is the total of some set of cells in the base table. Dependencies between bounds might imply that there exist tighter bounds than the trivial bounds. These may be found by integer linear programming methods. The recovery of a zero by subtraction of a known sample of the population occurs if, and only if, the sample implies that $s_i = c_i = u'_i$, where s_i is the corresponding known sample count and u' is the table of the tightest upper bounds on the base table implied by the set of all linear constraints.

The probability of recovering at least one zero for some assumed level of intruder knowledge, equivalent to a random sample of size n, is

$$SAP(n) = \frac{\sum_{s \in S} P(s \mid p) I\left(\sum_i s_i = n\right) I(0 \in u' - s)}{\sum_{s \in S} P(s \mid p) I\left(\sum_i s_i = n\right)},$$

where S is the set of all possible sample tables, p is the population table (known to the data holder), sampling is simple random sampling without replacement, subtraction of tables is pointwise, and $I(.)$ is the indicator function.

For a data release comprising of a single rounded table we have a pair of constraints, $l_i \le c_i$, $c_i \le u_i$ for each cell i. The mutual orthogonality of these pairs of constraints in R^n ensures that the trivial bounds are the tightest bounds. For a sample with corresponding sample counts, s_i, $i = 1$ to m, the SAP measure for a given sample size, n, can be calculated as follows.

2.3.1 Single rounded table

The marginal probability of recovering zeros in any set of cells with total x is simply the following Hypergeometric probability,

$$\frac{\binom{N-x}{n-x}}{\binom{N}{x}}$$ where N is the cross-classification total, $\sum c_i$.

Applying the inclusion / exclusion principle it is simple to derive an expression for the probability that at least one cell is zero given a random sample of n population units.

Let Z denote the set of all subsets of cell indices, equal to the union of the sets of n-subsets $Z(0)$, …, $Z(m)$. i.e. $Z(0) = \varnothing$, $Z(1) = \{\{1\}, \ldots, \{m\}\}$, $Z(2) = \{\{1,2\}, \{1,3\}, \ldots, \{m\text{-}1, m\}\}$, …, $Z(n) = \{\{1, \ldots, m\}\}$.

Let e.g. $c_1 + c_2$ be denoted by $c_{\{1,2\}}$.

Then,

$$SAP(n) = \sum_{i=1}^{n} \left((-1)^{i-1} \sum_{z \in Z(i)} \frac{\binom{N-c_z}{n-c_z}}{\binom{N}{n}} \right)$$

In practice many of the terms in the above summation will be equal to zero. For exact tables we have $l_i = c_i = u_i$ for all i, and all cell counts represent some risk of

recovering a zero, although for a given level of risk, n, we need only consider c_z s.t. $c_z \leq n$. For rounded tables we need only consider c_z s.t. $c_z = \sum_{i \in z} u_i \leq n$.

2.3.2 Single rounded table and rounded total

In this case we have an additional pair of constraints, $l_t \leq \sum_i c_i$, $\sum_i c_i \leq u_t$, where u_t denotes the trivial upper bound for the table total. We also have the obvious risk of subtraction where the sum of the sample counts $\sum_i s_i = u_t$, and this only occurs when $\sum_i s_i = \sum_i c_i = u_t$. But this new constraint is not mutually orthogonal to the existing constraints, and the trivial upper bounds might not be the tightest possible bounds. In this particular case the tightest possible upper bounds on any base table cell j is, $u'_j = \min\left(u_j, u_t - \sum_{i \neq j} l_i\right)$.

Lemma

If $s_i = u_i$ for any $s \in S$ and any i, then $s_i < u'_i$ for any $u'_i < u_i$. In other words, if there is any risk for the release without rounded total, then the release of the rounded total results in no increased risk.

Proof

It would be sufficient to show that $u_t - \sum_{i \neq j} l_i > c_j$ for any j. Minimum upper bounds occur when $u_t = \sum_i c_i$. So assuming the tightest possible 'new' bounds we have,

$$\sum_i c_i - \sum_{i \neq j} l_i > c_j$$

$$\sum_i (c_i - l_i) > c_j - l_j$$

So the lemma is proved true apart from the case where $c_i = l_i \; \forall \, i \neq j$, where we have equality.

In this case we have,

$$u'_j = \min\left(u_j, \sum_i (c_i - l_i) + l_j\right)$$

$$u'_j = \min\left(u_j, c_j - l_j + l_j\right)$$

$$u'_j = c_j$$

The existence of a risk (without rounded total) implies that $u_i = c_i$ for at least one c_i. So, either,

1. $u_j = c_j = u'_j$ and there is no increased risk (the bound is already tight), or

2. $l_i = c_i = u_i$ for some $i \neq j$, and we have a rounding scheme that doesn't round all counts.

The Lemma is proved for all independent rounding schemes that perturb all base table counts.□

Corollary 1

If $u_t > \sum_i c_i$, then the risk with rounded table is exactly the same as the risk without rounded table.

Corollary 2

If a rounded table represents zero risk, then the addition of a rounded total represents a risk if, and only if, $u_t = \sum_i c_i$. This risk pertains only to knowledge of the full table, unless exactly one cell count, say c_j, is not equal to its trivial lower bound. In that case all tables s.t. $s_j = c_j$ represent a risk.

So if $s_i = u_i$ for any $s \in S$ or $u_t > \sum_i c_i$, then we can use the algorithm for single rounded tables.

Otherwise, the above results lead to the following algorithm.

1. Construct a list containing the trivial lower bounds for the rounded base table counts (i.e. based solely on the rounding scheme).
2. Construct a corresponding list of counts for the exact cross-classification.
3. Find the sum, S, of those counts in the list of lower bounds that are equal to the corresponding count in the list of exact counts.
4. For all n in the range 0 to (T – S – 1) (where T is the exact cross-classification total) the SAP measure is zero.
5. For each n in the range (T – S) to T the SAP measure equals $\dfrac{\binom{S}{n-T+S}}{\binom{T}{n}}$

2.3.3 General table releases

It is hoped that the existing results can be further generalised to provide efficient means for calculating SAP measures for more general table releases. The current approach is to use an extended version of Dobra's shuttle algorithm to solve the initial bounds problem and then recursively generate all tables with non-zero risk (Smith and Elliot, 2003). Randomly sampling tables is an alternative approach for generating approximate SAP measures.

2.4 A comparison of rounding schemes

For contractual reasons we are, at present, unable to publish an extensive SAP analysis that we have conducted on the UK Neighbourhood Statistics. But Table 3 contains some results for an analysis of a set of 1200 randomly generated 2×6 cross-classifications. The cross-classification counts were generated from a Poisson distribution with mean 2. Each cross-classification was conventionally rounded to base 5, and the exact total was conventionally rounded (to base 5) to produce a rounded total. SAP measures for each cross-classification were generated for n=0 to 24. Table 1 contains the numbers of cross-classifications that had SAP scores in various ranges. SAP scores that were exactly 0 or 1 are contained in the second left and rightmost columns respectively.

For n in {0,1} the SAP measure is necessarily 0 for all cross-classifications, due to the nature of the rounding scheme. For n=2 the SAP measure could be as high as 1, given a cross-classification total of 2.

The SAP measure for any individual cross-classification and value of n must be at least that for n-1, so there tends to be a migration of SAP measures from 0 to 1 as n is

increased. But for cross-classifications with no relevant cells, the SAP measure is zero for all *n*.

SAP	=0	0-0.1	0.1-0.2	0.2-0.3	0.3-0.4	0.4-0.5	0.5-0.6	0.6-0.7	0.7-0.8	0.8-0.9	0.9-1	=1
n=0	1200	0	0	0	0	0	0	0	0	0	0	0
1	1200	0	0	0	0	0	0	0	0	0	0	0
2	26	1173	1	0	0	0	0	0	0	0	0	0
3	26	1137	36	0	0	1	0	0	0	0	0	0
4	25	869	275	28	2	0	0	0	1	0	0	0
5	25	460	497	160	40	16	1	0	0	0	0	1
6	25	234	471	267	127	48	20	6	1	0	0	1
7	23	108	332	365	169	113	59	13	14	2	0	2
8	23	60	226	266	292	115	117	50	28	17	4	2
9	23	33	144	201	254	212	115	102	64	31	17	4
10	23	14	93	146	188	203	220	93	105	75	30	10
11	23	9	58	110	158	150	205	176	125	95	72	19
12	22	4	42	63	108	149	186	154	188	113	138	33
13	22	4	21	51	84	126	126	169	211	148	186	52
14	22	3	12	37	57	104	111	138	157	234	244	81
15	22	3	8	33	32	65	104	124	159	217	303	130
16	22	3	2	18	32	46	80	108	127	199	379	184
17	22	3	0	12	30	31	46	101	118	182	395	260
18	21	4	0	7	16	28	39	73	93	144	442	333
19	20	5	0	2	13	26	25	45	84	143	409	428
20	20	5	0	0	9	12	29	40	54	105	423	503
21	20	4	1	0	7	9	26	18	43	104	365	603
22	20	3	1	0	2	10	9	27	35	74	318	701
23	20	3	1	0	0	7	10	16	26	42	290	785
24	19	3	1	1	0	5	6	10	26	33	235	861

Table 2.3. Simulation results showing for 1200 randomly generated tables the banded probabilities of producing a table containing at least one zero, given subtraction of *n* randomly selected units from the tables.

Table 2.3 demonstrates how the risk of recovering a potentially attribute-disclosive cross-classification tends to increase with greater intruder knowledge of the population. Of course, this depends on the size of the cross-classifications, the distribution of counts and the rounding scheme. But the pattern of results shown in Table 2.3 is reasonably close to that which the authors have found with real-world data sets. Analyses such as this can be used to help define threshold values for *n* for which a non-zero (or value greater than another threshold) SAP measure can be considered to constitute too great a risk for release. Similarly, analyses can be used to investigate the protection afforded by alternative perturbation schemes. Of course,

any comprehensive analysis of perturbation schemes would also consider the effect of perturbation on data quality.

3 Summary

The SAP method provides an integrated approach for assessing attribute disclosure risk for any given release of cross-classifications. It incorporates the notion of intruder knowledge and allows the same metric to be produced for single released cross-classifications and multiple released cross-classifications, whether perturbed or unperturbed. Computational constraints mean that comprehensive analyses of large cross-classification releases can be time consuming. Although the computational burden can be ameliorated through sampling to derive approximate SAP measures, further work is needed on producing exact measures. Far more efficient algorithms have been found for certain special cases. These cases were chosen for no other reason than the fact that they are common forms of release from the Office for National Statistics; in fact, there are obvious extensions to other cases that are not detailed here. Nevertheless the SAP method provides a risk measure for attribute disclosure. Most existing risk measures only concern identification risk. Concentrating on identification risk, at the expense of attribution risk, raises the possibility of real disclosure occurring as a result of the release of data that is considered 'safe' by current risk measures.

References

Dobra, A. (2002) Statistical Tools for Disclosure Limitation in Multi-Way Contingency Tables. *Ph.D. Thesis*, Department of Statistics, Carnegie Mellon University

Duncan, G.T. Fienberg, S.E., Krishnan, R. Padman, R. Roehrig, S. F.(2001) Disclosure Limitation and Information Loss for Tabular Data. In P. Doyle, J. I. Lane, J. M. Theeuwes, and Zayatz, L. M. (eds) *Confidentiality and Data Access: Theory and Practical Application for Statistical Agencies*. pp 135-183.

Elliot, M. J. (2000) DIS: A New Approach to the Measurement of Statistical Disclosure Risk. *International Journal of Risk Management*, 2(4) pp.39-48

Elliot, M. J., Dale, A. (1999) Scenarios of Attack: The Data Intruder's Perspective on Statistical Disclosure Risk. *Netherlands Official Statistics*. Spring.

Fellegi, I.P. (1972). On the Question of Statistical Confidentiality, *Journal of the American Statistical Association*, Vol. 67, No. 337, pp.7-18

Mokken, R. J., Kooiman, W. J., Pannekoek, J., Willenborg (1992) Disclosure Risks for Microdata. *Statistica Neerlandica*, Vol. 46. pp.49-67

Paass, G. (1988) Disclosure risk and Disclosure Avoidance for Microdata, *Journal of Business and Economic Statistics* 6(4) pp.487-500

Skinner, C.J. (1992). On Identification Disclosure and Prediction Disclosure for Microdata. *Statistica Neerlandica*, Vol. 46, No. 1, pp.21-32

Skinner, C. J., Elliot, M. J. (2002) A Measure of Disclosure Risk for Microdata, *Journal of the Royal Statistical Society Series* B, 64(4) pp.855-867

Smith, D., Elliot, M. (2003) An Investigation of the Disclosure Risk Associated with the Proposed Neighbourhood Statistics. *Report for the Office of National Statistics*

U.S Department of Commerce (1978) Report on Statistical Disclosure and Disclosure Avoidance Techniques, *Statistical Policy Working Paper* 2, Washington.

Cell suppression in a special class of linked tables

Peter-Paul de Wolf*

* Statistics Netherlands, Department of Methodology and Quality, P.O. Box 4000, 2270 JM Voorburg, The Netherlands. (pwof@cbs.nl)

Abstract. A heuristic approach to protect a 'stand alone' hierarchical table has already been available for some time now and is known as either HiTaS or the modular approach. At NSI's often linked (hierarchical) tables need to be dealt with. To be able to protect linked tables, one first needs a way to easily define those tables. In the current paper we will discuss a possible way to describe a simple class of linked tables that is often considers at NSI's. Moreover, we will point out some irregularities that should be kept in mind, if such a method would be implemented.

1 Introduction

Cell suppression is an often used disclosure control technique to protect tabular data at NSI's. To use that technique on a single (hierarchical) table, several methods are available. E.g., in τ-ARGUS both the modular approach and the HyperCube approach can deal with hierarchical tables. The modular approach however, is not able to deal with linked tables. To extend the modular approach to generally linked tables is very difficult.

However, NSI's often have to deal with a very special kind of linked tables. E.g., they often publish a table with turnover specified by a detailed NACE code and aggregated Regional code as well as a table with turnover specified by a detailed Regional code and aggregated NACE code, while they won't publish a table with turnover specified by detailed NACE code and detailed Regional code.

A straightforward way to deal with those linked tables, is to protect the complete hierarchical table with both detailed NACE code as well as the detailed Regional code. However, this will often lead to over suppression: unsafe cells at the lowest level (detailed NACE and detailed Region at the same time) will often lead to secondary suppressions at the higher levels.

The heuristic HiTaS (also known as the modular approach) is such that it will protect all possible non-hierarchical subtables of a hierarchical table in a special order. See e.g., De Wolf (2002). Hence, it seems reasonable to extend this heuristic in the sense that it is allowed to discard certain subtables that will not be published

anyway. However, this means that we need a way to tell HiTaS which subtables need to be protected and which can be discarded with respect to disclosure control.

In the current paper we will suggest a simple way to define the special class of linked tables that was mentioned above. Section 2 contains some definitions that we will need. In section 3 we will present some examples. Since heuristics are used, the protection of the linked tables is not necessarily watertight. In section 4 we will show some problems that might arise.

2 Hierarchies

In this section we will define our notion of hierarchies, as we will be using throughout the paper. First we will need some general definitions, taken from graph-theory.

Definition 1 *A tree is an undirected graph T in which any two vertices is connected by exactly one path.*

Definition 2 *The distance between two vertices v and w is the number of edges on the path between v and w.*

Definition 3 *A directed tree with root r is a directed graph T which would be a tree if the directions on the edges were ignored and in which each vertex $v \neq r$ in T has a path from r to v.*

Definition 4 *In a directed tree with root r, a vertex v is a ancestor of vertex w if a path exists from v to w. Moreover, w is then called a descendant of v. In case a path of length one exists from v to w, then v is the father of w and w is a child of v.*

In the remainder of this paper, we will consider a hierarchy to be a rooted, directed tree, with the categories being the vertices of the tree. Using this representation, we can define the levels of the hierarchy as well

Definition 5 *The level in a hierarchy of a category, is the distance between the corresponding vertex to the root of the directed tree.*

Note that, by definition, the root category is at level 0.

In order to be able to compare different hierarchies when considering linked tables, we will need some more definitions.

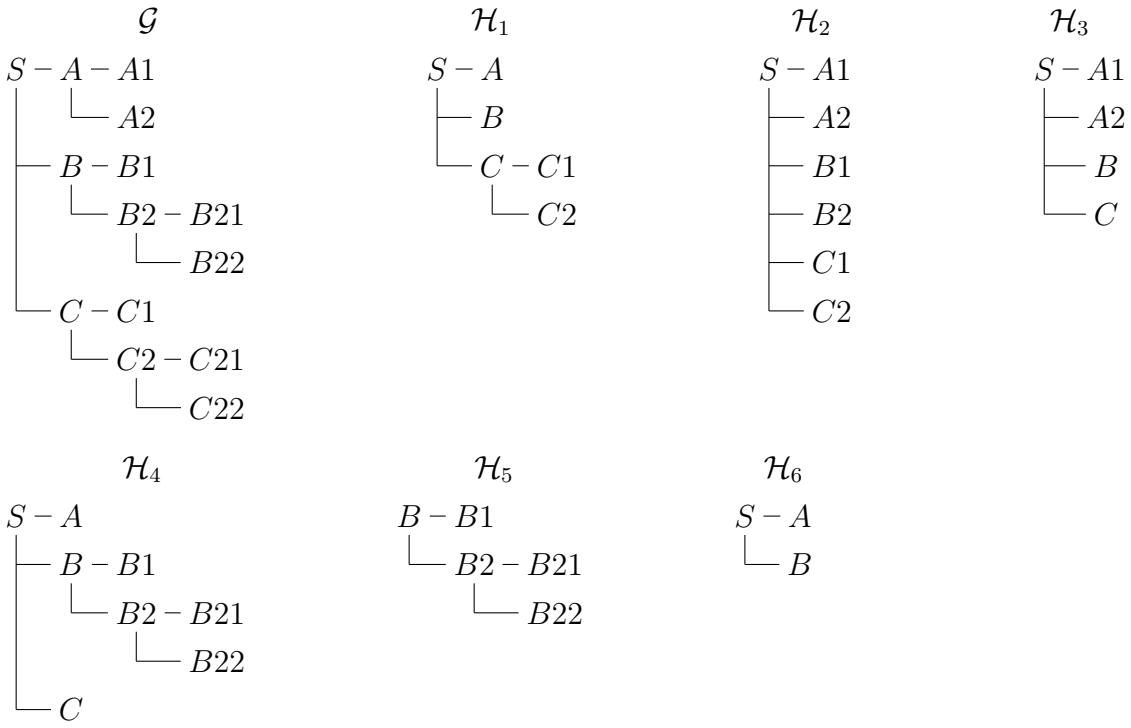
Definition 6 *A hierarchy $\mathcal{H}$ is called a pure sub hierarchy of hierarchy $\mathcal{G}$ if each path from vertex v to vertex w in $\mathcal{H}$ is also a path from v to w in $\mathcal{G}$. Notation: $\mathcal{H} \preceq \mathcal{G}$.*

Definition 7 *A set of hierarchies $(\mathcal{H}_1, \ldots, \mathcal{H}_K)$ is covered by hierarchy $\mathcal{G}$, if $\mathcal{G}$ is the hierarchy with the smallest number of vertices for which $\mathcal{H}_i \preceq \mathcal{G}$ for each $i = 1, \ldots, K$. Hierarchy $\mathcal{G}$ is then called the covering hierarchy.*

In the previous definition a covering hierarchy was defined for a set of hierarchies. In practice, we will often start with a detailed hierarchy and construct pure sub hierarchies by deleting all descendants of certain vertices. This leads to the following definition:

Definition 8 *A hierarchy $\mathcal{G}$ is a base hierarchy of hierarchy $\mathcal{H}$, in case $\mathcal{H}$ can be constructed by deleting all descendants of certain vertices in $\mathcal{G}$.*

A base hierarchy can be chosen such that it is a covering hierarchy as well. However, a covering hierarchy is not necessarily the (covering) base hierarchy of the same set of sub hierarchies. E.g., consider the following seven hierarchies:



We then have that

$$\mathcal{H}_1 \preceq \mathcal{G}, \quad \mathcal{H}_2 \not\preceq \mathcal{G}, \quad \mathcal{H}_3 \not\preceq \mathcal{G}, \quad \mathcal{H}_4 \preceq \mathcal{G}, \quad \mathcal{H}_5 \preceq \mathcal{G} \quad \text{and} \quad \mathcal{H}_6 \not\preceq \mathcal{G}.$$

Hierarchy $\mathcal{H}_6$ is not a pure sub hierarchy of $\mathcal{G}$, since root S of $\mathcal{H}_6$ equals root S of $\mathcal{G}$ if and only if the categories A and B differ from the categories A and B in $\mathcal{G}$.

Note that $\mathcal{G}$ is not a covering hierarchy of the set $(\mathcal{H}_1, \mathcal{H}_4, \mathcal{H}_5)$: vertices $A1, A2, C21$ and $C22$ of $\mathcal{G}$ are not needed to make $\mathcal{H}_1$, $\mathcal{H}_4$ and $\mathcal{H}_5$ pure sub hierarchies of $\mathcal{G}$. Additionally, $\mathcal{G}$ can be a base hierarchy of $\mathcal{H}_1$ and $\mathcal{H}_4$ but not of $\mathcal{H}_5$.

3 Example of linked tables

In this section we will provide an example of how to specify a set of linked tables and to derive a convenient cover table.

A user specifies N tables$T_1, \ldots, T_N$ that need to be protected simultaneously. Each table can have a hierarchical structure that may differ from the other hierarchical structures. However, tables that use the same spanning variables are only allowed to have hierarchies that can be covered.

Suppose that the specified tables contain M different spanning variables. Since the hierarchies are supposed to be coverable, an M-dimensional table exists having all the specified tables as subtables. The spanning variables will be numbered 1 up to M. The order only affects the way the M-dimensional table will be specified, not the way the suppression pattern is constructed.

Each spanning variable can have several hierarchies in the specified tables. Denote those hierarchies for spanning variable i by $\mathcal{H}_{i,1}, \ldots, \mathcal{H}_{i,\mathcal{I}_i}$ where $\mathcal{I}_i$ the number of different hierarchies.

Define the M-dimensional table by the table with spanning variables according to hierarchies $\mathcal{G}_1, \ldots, \mathcal{G}_M$ such that, for each $i = 1, \ldots, M$ hierarchy $\mathcal{G}_i$ covers the set of hierarchies $(\mathcal{H}_{i,j})$ with $j = 1, \ldots, \mathcal{I}_i$. This M-dimensional table will be called the cover table.

The modular approach (HiTaS) can now easily be adapted. HiTaS deals with all possible non-hierarchical subtables of a hierarchical table in a specially ordered way. We could keep this subdivision, but only consider those subtables that are also subtables of at least one of the specified tables $T_1, \ldots, T_N$ and disregard the other subtables.

Example 1 A user specifies three tables: T_1 with spanning variables $(S \times W \times R)$, T_2 with $(S \times G)$ and T_3 with $(S \times R)$. A base hierarchy for spanning variable S is given in figure 1.

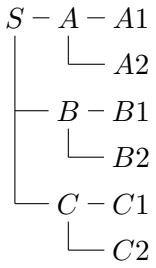


Figure 1: Base hierarchy $\mathcal{G}_S$ of spanning variable S

Since there are 4 different spanning variables, the cover table will be a four dimensional table. In figures 2–4 the hierarchies of the spanning variables are given, where the spanning variables S, R, G and W are numbered 1, 2, 3 and 4 respectively.

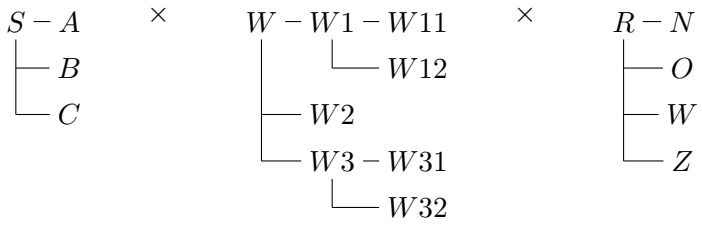


Figure 2: Table T_1: $\mathcal{H}_{1,1} \times \mathcal{H}_{4,1} \times \mathcal{H}_{2,1}$

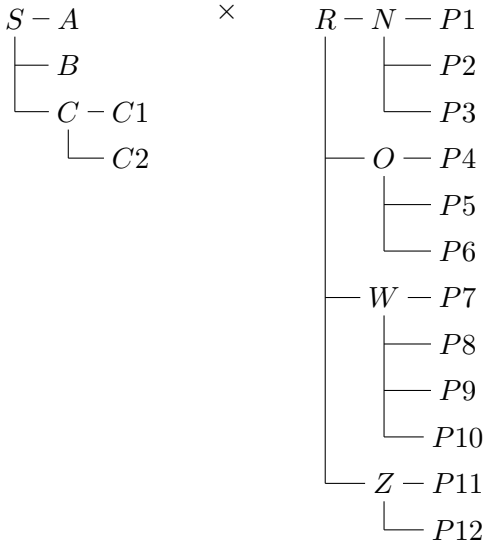


Figure 3: Table T_2: $\mathcal{H}_{1,2} \times \mathcal{H}_{2,2}$

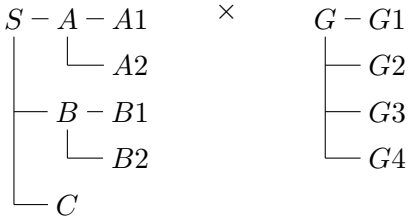


Figure 4: Table T_3: $\mathcal{H}_{1,3} \times \mathcal{H}_{3,1}$

The resulting 4-dimensional cover table will have the hierarchies $\mathcal{G}_1 = \mathcal{G}_S$ where $\mathcal{G}_S$ is the base hierarchy as given in figure 1, $\mathcal{G}_2 = \mathcal{H}_{2,2}$, $\mathcal{G}_3 = \mathcal{H}_{3,1}$ and $\mathcal{G}_4 = \mathcal{H}_{4,1}$. ◇

4 Problems

The idea of disregarding certain subtables may lead to situations in which the disclosure control is not necessarily watertight. In this section we will give two instances of problems that may arise.

In certain situations, the by the user specified tables may completely fix the internal structure of the higher dimensional table that will be considered. In case that internal structure is not included in the specified tables, it will not be considered when deriving a suppression pattern. Any, *implicitly* defined table structure will not be protected *explicitly*. The following example will show that in that way, implicitly a primary unsafe cell may be published.

Example 2 In figure 5 three tables are given that are specified by the user. These tables

	$B1$	$B2$	$B3$	Total
$A1$	30	150	70	250
$A2$	270	110	255	635
Total	300	260	325	885

	$C1$	$C2$	$C3$	Total
$A1$	140	50	60	250
$A2$	85	210	340	635
Total	225	260	400	885

	$C1$	$C2$	$C3$	Total
$B1$	110	90	100	300
$B2$	40	50	170	260
$B3$	75	120	130	325
Total	225	260	400	885

Figure 5: Three linked 2-dimensional tables

are considered to be safe on their own. However, they are the marginals of a three dimensional table $A \times B \times C$. Under the condition that all cells in that three dimensional table are non-negative, exactly one interior of that table is possible (see figure 6).

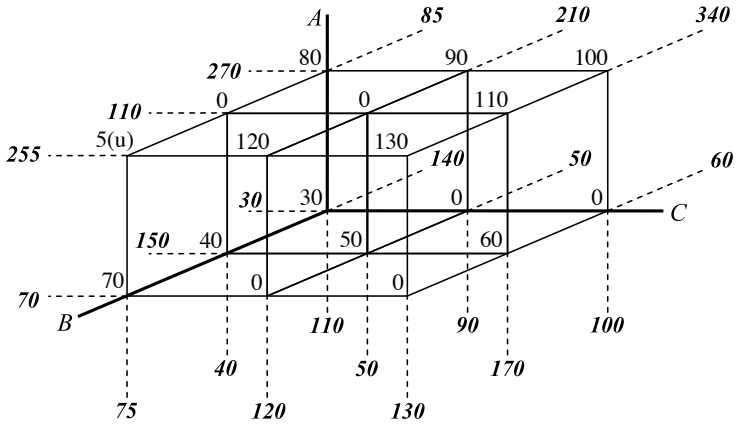


Figure 6: Graphical representation of $A \times B \times C$ of Example 2

So, if e.g., cell $(A2, B3, C1)$ would be primary unsafe (in the figure denoted by 'u'), the publication of the three 2-dimensional tables from figure 5 would hence implicitly lead to disclosure of a primary unsafe cell. ◇

Similarly, a suppression pattern can in certain situations be broken, due to the fact that the interior of the corresponding higher dimensional table is unique. See the following example.

Example 3 Figure 7 shows three tables with a possible suppression pattern. The bold cross is supposed to be a primary unsafe cell, the other crosses are secondary suppressions. Again, the underlying three dimensional table appears to be unique. See figure 8. Due to the uniqueness of the three dimensional table, the suppressed cells can easily be disclosed. ◇

	$B1$	$B2$	$B3$	Total
$A1$	10	50	x	x
$A2$	110	70	270	450
$A3$	110	250	x	x
Total	230	370	450	1050

	$C1$	$C2$	$C3$	Total
$A1$	50	20	**x**	x
$A2$	80	140	230	450
$A3$	250	120	x	x
Total	380	280	390	1050

	$C1$	$C2$	$C3$	Total
$B1$	120	50	60	230
$B2$	–	140	230	370
$B3$	260	90	100	450
Total	380	280	390	1050

Figure 7: Three linked 2-dimensional tables with suppression patterns

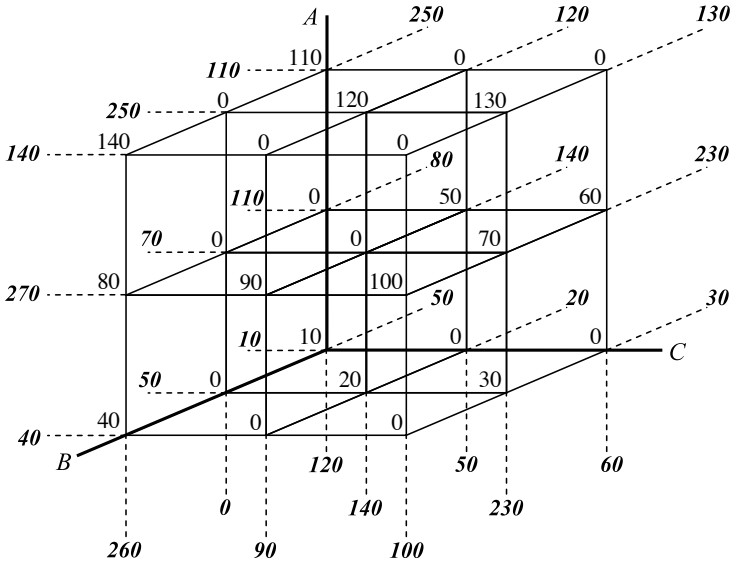


Figure 8: Graphical representation of table $A \times B \times C$ of Example 3

References

de Wolf, P.P. (2002). HiTaS: A heuristic approach to cell suppression in hierarchical tables. In J. Doming-Ferrer (Ed.), *Inference Control in Statistical Databases*, Berlin Heidelberg, pp. 74–82. Springer-Verlag.

Census tables: utility and safety via a cell threshold

Mike Camden*, Paul Cowie* and Lisa Henley*

*Statistics New Zealand, PO Box 2922, Wellington, New Zealand
mike.camden@stats.govt.nz, paul.cowie@stats.govt.nz, lisa.henley@stats.govt.nz

Abstract: Tables of counts from a population census are valued highly by users, especially users from local government. Users request detailed tables that are sometimes very sparse. Statistics New Zealand's past protections against sparseness include random rounding to base 3 and a mean cell size rule. It decided, for its 2006 Census of Population and Dwellings, to further enhance data utility and protect safety by setting a threshold for sparse tables: counts above this are released and counts at this value or below are suppressed. The confidentiality rules are applied to each geographical area separately, and our current geographical areas vary widely in population. This variation in size is a source of sparseness, but we use it in two ways. We calculate measures for utility and safety for each geographical area, and use them to evaluate the threshold approach. We also use them to help in setting the size for a possible new set of geographical areas designed for output. The results can be applied to heighten utility and safety for our 2011 Census.

1 Introduction: sparseness and methods of managing it

The New Zealand Census of Population and Dwellings is held every five years. The most recent one was in March 2006, and planning is in progress for the 2011 Census. Tables of counts are a major mode for dissemination of census information. These tables are often sparse, with large proportions of counts being 0's and 1's. The paths for providing confidentiality protection for tables like these are summarised by Wooton and Fraser (2005) and Schlomo (2005). Statistics New Zealand has taken one path: random rounding, a mean cell size rule for tables by geographical area, and now a threshold rule for cells. We refer to these rules as R, M and T, respectively, and to the combinations of interest to us as R, RM, RT and RMT. In choosing our path, acceptance by users is important.

We use the utility/risk framework (Duncan, Fienberg, Krishnan, Padman and Roehrig, 2000) to assess our set of rules. We refer to the converse of risk as safety. We calculate some measures for utility and safety for each geographical area, and use the variation in the size of the areas to assess the effect of our rules. We use this variation also to indicate the ideal size for geographical areas designed for output.

1.1 Recent history of confidentiality for the Census of Population and Dwellings

Counts in tables have been protected since the 1981 Census by random rounding to base 3 (R), and since 2001 by a mean cell size rule (M). The mean cell size for 2001 was 1 unit per cell. The 2001 rules allowed the release of sparse tables containing

many 0's and a few 3's, which are likely to arise from 1's. Jackson, Corscadden and Zeng (2005) examined uniqueness in the 2001 Census, and recommended a design perspective for small geographic areas and categories with small proportions. In preparation for 2006, a modified set of rules was created that targeted sparseness by raising the mean cell size from 1 to 2, and applied M to each geographic area separately. (The table for an area was published if mean cell size > 2.) It targeted the sensitive variable income by using an aggregated version without small frequencies.

This set of rules was designed so that, when clients submitted a request, staff could tell them whether counts would be supplied, and for which areas. The rules used information from outside the table (population in geographical area, number of cells), and so could be applied before tables were built. The rules worked well for standard tables and large geographies. As client requests for customised tables from the 2006 Census flowed in, it became clear that clients had ongoing needs for tables that contained both sparse patches and useful counts, and failed M. We needed to provide both safety and utility for these customised tables. Our solution was to keep the mean cell size rule M, but where the table for an area failed it, we would release counts above the threshold of 5, and suppress counts of 5 and below (T). We continued with random rounding (R), so secondary suppression was not needed. We assume that rules RMT reduce risk to an acceptable level.

In moving along this path, we shifted from rules that could be applied before tables were built to further modifications that would be applied after tables were built. We responded to a complex set of pressures with a complex and more finely targeted solution. This path raises questions about both utility and safety. The paper quantifies the effect of the path on utility and safety, across a range of geographical sizes.

1.2 Variation in geographical classifications: risks and opportunities

Our rule set recognises the importance of geographical location in confidentiality protection, and treats the table for any geographical area as a separate entity. The rules are applied area by area. The main geographical classifications are these:

1. 1 country, with a usually resident population (2006 Census) of 4,249,737 people
2. 16 regional authorities: min = 31,326 mean = 251,708 max = 1,303,068
3. 73 territorial authorities: min = 612 mean = 55,172 max = 404,658
4. 1,918 area units: min = 0 mean = 2,100 max = 9,027
5. 41,384 meshblocks: min = 0 mean = 97 max = 1,431.

We are concerned here with the last two classifications. Both are extremely variable, and were designed for purposes like data collection and classification-building rather than for dissemination. Fig 1 shows the size distribution for area units. The large number of areas with low sizes, on the left, is a major source of sparseness. However, we make use of this variation to assess the effects of our rules on utility and safety, and to suggest a better-distributed new set of geographies for the 2011 Census.

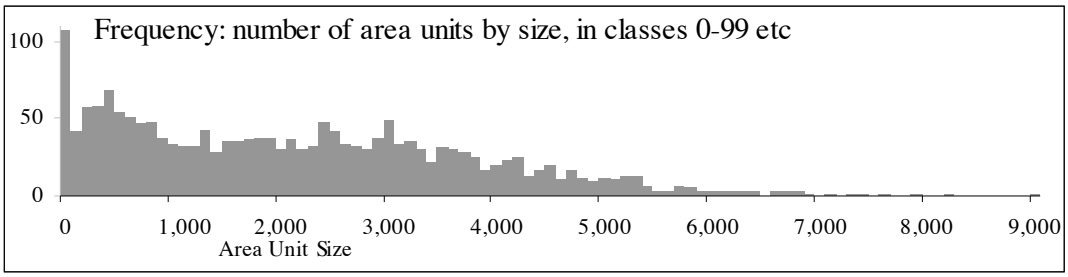


Fig 1 Frequency distribution of Area Unit Size (in classes 0–99, 100–199 etc).

1.3 Aims and contents of this paper

This paper aims to provide information about two questions. The first is about how well the path that we are taking provides utility and safety. The second is about how new classifications, especially for geographic location, can be designed. Part 2 uses a one-way table to illustrate how our rules work, and introduces the two-way table used in the analysis. It discusses table structure and a table's frequency distribution of counts. Part 3 defines some measures, applies them to our typical table and seeks answers to our two questions. It checks for new risks from a complex rule set. Part 4 concludes, with messages from the data and possibilities for the future.

1.4 Example 1: a one-way table

Table 1 concerns a rural area unit that is large in area but small in population size. Its subject population, for the 10-category variable Occupation, is 19.

Occupation:	a	b	c	d	e	f	g	h	i	j	Tot	SD
Row 1: Raw	16	0	0	1	0	0	1	0	0	1	19	–
2: R	15	0	0	3	0	0	0	0	0	0	18	–
3: noise	-1.0	0.0	0.0	2.0	0.0	0.0	-1.0	0.0	0.0	-1.0	–	0.8
4: RM	c	c	c	c	c	c	c	c	c	c	18	–
5: noise	-14.2	1.8	1.8	0.8	1.8	1.8	0.8	1.8	1.8	0.8	–	4.7
6: RMT	15	c	c	c	c	c	c	c	c	c	18	–
7: noise	-1.0	-0.3	-0.3	0.7	-0.3	-0.3	0.7	-0.3	-0.3	0.7	–	0.5

Table 1 Counts for the 10 categories of Occupation, for a small area unit with 19 people, with three rule sets and the noise resulting from each; c = confidential.

Row 1 has the original counts. It can be reduced to its proportional frequency distribution: $P(0) = .6$, $P(1) = .3$, $P(16) = .1$. The mean cell size is 19/10 = 1.9.

Row 2 has random-rounded counts (R), and row 3 has the noise added by this process. The table passes M with parameter 1, and would have been published for 2001. The standard deviation of the noise, SD, measures information disturbance.

In row 4, we use rules RM with parameter 2 for M (as in our 2006 rules), and the table fails. A user could estimate the c's simply as the apparent mean: 18/10 = 1.8. The noise that results is in row 5, and has a higher value for SD.

Row 6 represents our current practice: rules RMT. The table fails M, the threshold of 5 is applied, and the one count over 5 (the 15) is rescued (rule T). Rule R is applied. A user could replace the c's with an improved guess: 3/9 = 0.33. The noise that results is in row 6, and has a lower value for SD.

1.5 Example 2: a two-way table, and the structures within it

Our analysis below uses a typical user request for customised output. For each area unit, we find the two-way table of counts for:
Travel: main means of travel to work: 11 categories: foot, cycle, bus, train, car etc
Age: age in 5-year groups: 11 categories: 15–19, 20–24, … to 65+.
There are 121 cells, and the limit for rule M is 2 x 121 = 242 people.

The count for any cell depends on some structural elements: the cell's values for Travel and Age, the interaction between these, and the adjacent values for Age. These structures result in the actual counts, and we can convert these counts into their frequency distribution. This is a rich source of information about utility and safety, and we use it to define measures of both.

If we assume nothing about a table apart from population N and number of cells K, then we could assume the counts are approximately Poisson, with parameter = N/K. Some of our graphs include curves for this model. They show that real tables have much more clustering than tables from the model would have.

2 Measuring the effect of our rules on utility and risk

We define one measure for safety and two for utility, view them against Area Unit Size for area units of size 1 to 3,000, and assess risks arising from the complexity of our RMT rule set.

2.1 Three measures defined

Let $P(x)$ = the proportion of cells with count x. $P(0)$, $P(1)$, $P(2)$, $P(3)$ and combinations of them are all useful as measures of safety (or its converse, risk). All these proportions, and the relationships among them, can be investigated using our set of area units. Our measure is: PS = $P(0) + P(1)$, the Proportion of Sparse cells. Both 0's and 1's are high-risk. The 2's are also risky, but are not used here. PS measures risk in the original table, before any rules are applied. Our rules reduce this risk: R replaces 0's, 1's and 2's with 0's or 3's, and M with T replaces them with c's for small area units.

Our first measure of utility is PR: the Proportion of people for whom (rounded) counts would be Released, assuming that they needed to be Rescued by the threshold (T) from suppression. PR measures utility only where T is used.

Our second measure of utility (or in fact a converse of it: information disturbance), SD, arose earlier. It is the standard deviation of the noise added by a set of rules. It compares the counts from any rule set with the original counts. We calculate and graph it for these four rule sets: R, RM, RT, RMT. It is related to variance (Wooton and Fraser, 2005) and mean absolute deviation (Schlomo and Young, 2005). SD has the same units as the counts that we are interested in, and is easily calculated.

2.2 Some data visualisation from our two-way table

The two-way table from Example 2 in Part 1.5, for Travel by Age, has 121 cells, and can be produced for each of our area units. The graphs below show the measures for those area units with subject population from 1 to 3,000. The areas above 3,000 have low risk and minimal loss from the rules, and so are omitted.

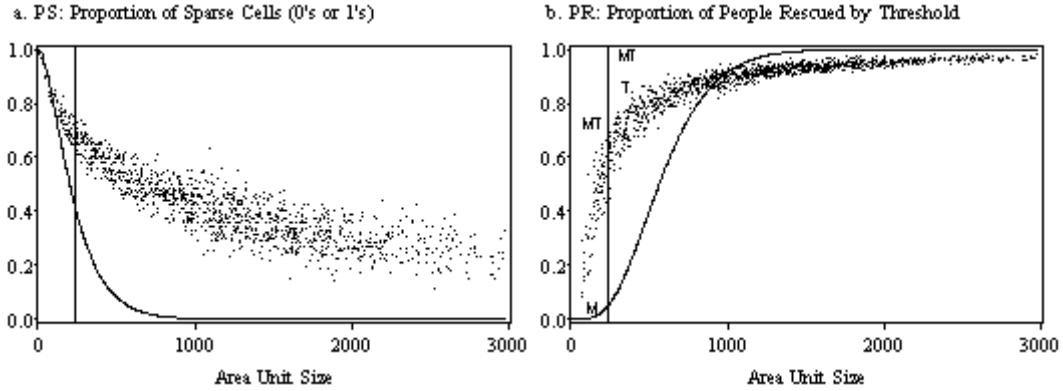


Fig 2: a: PS: Proportion of Sparse cells containing 0 or 1, by size of area unit. b: PR: Proportion of people Rescued from suppression by T, assuming that their area failed M and therefore needed rescue. Both graphs contain curves for the Poisson model.

Fig 2a shows risk, measured by PS = $P(0) + P(1)$. It is high for small areas, and drops off as area size increases. Small areas need protection, and the graph throws some light on what 'small' means here. For this table, our rule M (with parameter 2) protects areas up to 121 x 2 = 242. The graph suggests that the parameter 2 is not excessive. Protection could, in future, come from a new geographical classification. Graphs of further examples will suggest where the minimum size for this could be. (Much of the downward trend in this graph comes from $P(0)$; $P(1)$ is quite stable.)

Fig 2b shows utility, measured by PR. It helps to answer the question as to whether rule T, rescue by threshold, is worthwhile. The vertical line is at the limit for M: size = 242. To the left of the line are the areas that fail M. When we include rule T, the proportion of people whose data are released rises from zero (see M on the graph) to

the point cloud (see MT). Fig 2b also helps with the question as to whether we now need the mean cell size rule at all: we could suppress all counts below threshold, regardless of area size. Our rule set would be RT only. To the right of the line, under RMT, all areas pass M, and all counts are released (see MT). Under RT, the proportion of people whose data are released would fall to the point cloud (see T). The graph suggests that RMT is worth keeping.

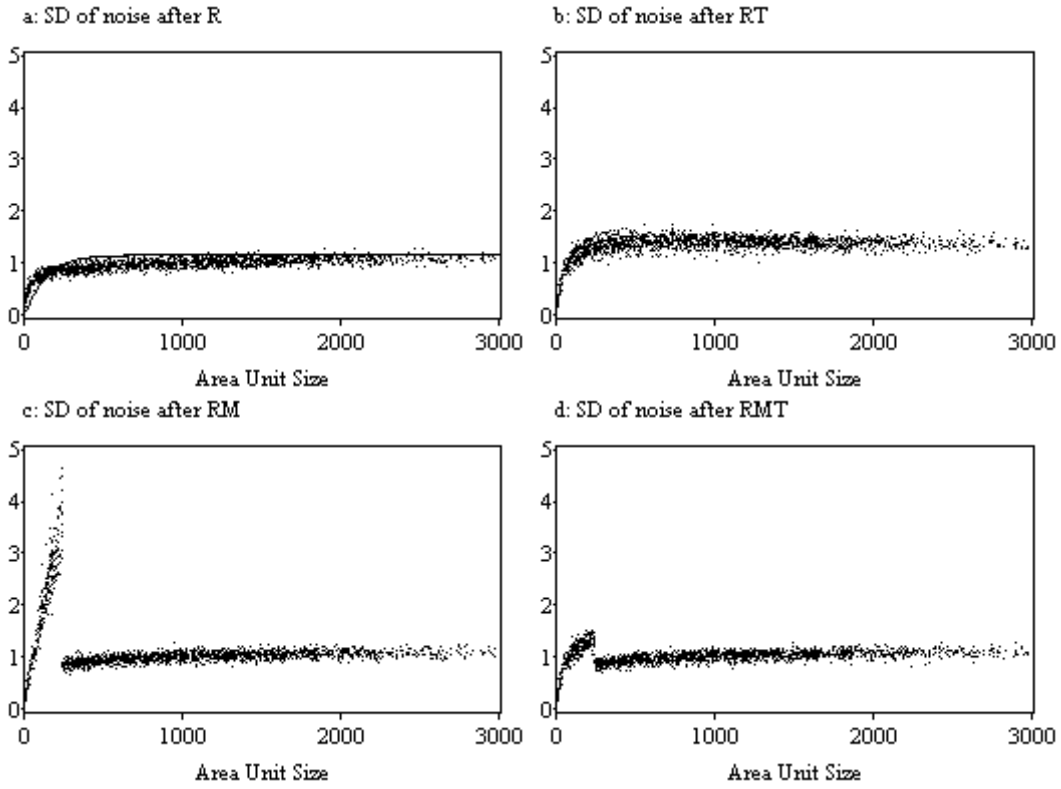


Fig 3 SD: Noise after rules R (with Poisson), RT, RM, RMT; by Area Unit Size.

Fig 3a shows that, under random rounding only (R), the noise rises up from 0 (zero cells stay zero and have no noise added) to a constant level. If a table has no 0's and counts spread evenly, we expect this level to be the noise for R: $2/\sqrt{3} = 1.155$, but it is lower. We have added the curve for Poisson tables, which does reach this level.

Fig 3b shows that, with rules RT, the noise is consistently higher. This helps further with the question as to whether we now need the mean cell size rule at all. Under RT, we would suppress cells below the threshold for all areas, large and small. We would eliminate all sparseness, but pay for it with the increased noise.

Fig 3c shows what happens with our earlier plan to use rules RM. The noise rises steeply as Area Unit Size approaches the limit for M.

Fig 3d shows the improvement in moving to our current plan, RMT. This helps with the question as to whether the rescue by threshold is worthwhile. The graph implies that it is: information disturbance is smaller and so utility is improved.

The parameters for R (base = 3), M (mean cell size = 2) and T (threshold = 5) could all be varied, and the effects assessed with these measures. Our threshold is set at 5 for several reasons. Numbers like 2, 5, 8 … as thresholds have a much tidier interaction with R than the others, with less bias and less disclosure risk. The threshold at 5 hides 0's, 1's and 2's as it must, and goes a little further. Users lose counts rounded to 3, but these have a high proportion of noise added by R. The proportion of people lost by T is quantifiable, and appears as 1 - PR in Fig 2b.

We have used these measures on other tables with area units, and on tables with the meshblock classification. The patterns and conclusions are similar. We need to extend the measures to test counts of households.

2.3 Complexity of the rule set

The use of the complex RMT combination of rules raises three concerns. The first is about complexity for users. Our discussions with expert users of tables suggest that they are not concerned about this, and that they see our path as a logical and useful one. The second is about complexity for automation of the rules. The rules have been programmed successfully into our software systems. The third is about interactions among the rules, and the effect on disclosure control. There are combinations of counts for which the rules provide a lower level of protection than usual. We balance this risk against the reduction in published 3's, and the increased utility.

3 Conclusions

3.1 Messages from the census data

The first message from the graphs above is about redesign of geographical classifications. Our example and others can suggest where the minimum size should be set, and they can indicate what the effect of this is on utility and safety. The upward pressure on this number needs to be balanced by downward pressure from user needs, and from geographical practicalities.

The second message is about classifications, and will become clearer with further examples. Some classifications are appropriate for certain tables, and some are not. The Age classification that we used (15–19, 20–24 to 65+) is good for the employed population. An alternative (0–9, 10–19 to 100+) would give tables where over half the cells held very few, if any, employed people. We need to design a variety of classifications for different output needs, and test them against the data with these graphical tools.

The third message is about safety and utility. For safety, our complex rule set, RMT, deals with sparseness to the left of the limit in Fig 2a, but not to the right. For utility, Fig 2b and all of Fig 3 show that RMT is worthwhile, both in releasing counts for high proportions of persons and in limiting the noise added to counts.

3.2 Further issues for investigation

For our 2011 Census, we are open to all possibilities for data access. We need to assess our RMT path against the quite different paths being developed and taken by other agencies. We need to deal with means of numerical variables, and other derivations, in consistent ways. We need to continue and enhance our microdata access: via licensed files, remote access, and our three data laboratories. Along with all this, we need a programme to inform users of these options, and enable them to strengthen their methods of accessing our data.

3.3 Summary

If we wish to optimise both utility and safety for tables of counts, we need to act at the design stage of the statistical process, as well as at the dissemination stage. The design of geographical and other classifications needs to combine evidence from within a census dataset with the needs of users. At the dissemination stage, it is appropriate to respond to a situation of complex risks and user needs with a complex set of rules that can filter out for release the information that is most useful and safe.

References

Duncan, G.T., Fienberg, S.E., Krishnan, R., Padman, R. & Roehrig, S.F. Disclosure Limitation Methods and Information Loss for Tabular Data. (2001). In *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*. Doyle, P., Lane, J.I., Theeuwes, J.M.M., & Zayatz, L.M. (Eds). Elsevier Science.

Jackson, L.F., Corscadden, L., & Zeng, I. (2005). *Small Area Disclosure Control: When Do Uniques Occur?* The 55th Session of the International Statistical Institute Sydney 2005 Proceedings.

Schlomo, N. (2005). *Assessment of Statistical Disclosure Methods for the 2001 UK Census*. Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality. Geneva.

Schlomo, N., & Young, C. (2005). *Information Loss Measures for Frequency Tables*. Joint UNECE/Eurostat work session on statistical data confidentiality. Geneva.

Wooton, J., & Fraser, B. (2005). *A Review of Confidentiality Protection for Statistical Tables with Special Reference to the Differencing Problem*. Australian Bureau of Statistics, Canberra.

Improving researcher access to USDA's Agricultural Resource Management Survey

Charles Towe and Mitch Morehart*

* Economic Research Service, U.S. Department of Agriculture, 1800 M Street NW, Washington, DC. (ctowe@ers.usda.gov)

Abstract. ERS and NASS annually collect a wealth of data that describe farming in America through the Agricultural Resource Management Survey (ARMS). While ARMS provides a rich set of economic data, customers outside the government, especially researchers, have gone largely under-served. ARMS data had been available only within USDA except when summary information was released publicly, when users requested ERS do special tabulations, or when academic researchers entered into special research agreements with ERS/NASS to access the data in USDA offices. In response, ERS developed easy to use web-based data delivery tools that expand access to farm survey data as a public good, while maintaining the security of the confidential data. Experience in the construct and maintenance of the tool since its inception, stakeholders use and satisfaction, and recent changes and enhancements are discussed.

1 Introduction

Economic research is highly data dependent. Often, the most interesting research problems are stimulated by policy questions on distributional issues that simply cannot be addressed without microdata on establishments and the firms that own them. The importance of microdata to economic research was the theme of Heckman's Nobel Lecture where he suggested that: "The availability of new forms of data has raised challenges and opportunities that have stimulated all of the important developments in the field and have changed the way economists think about economic reality (Heckman, 2001)." Establishment data are vital in understanding individual or firm behavior and are necessary to determine the marginal impacts of changes in policy or from other internal or external events. Microdata enable analysts to do multivariate regressions, whereby the marginal impact of key variables, controlling for other factors, can be isolated (Lane, 2003). Widely accessible microdata also have the additional benefit of allowing replication and verification of research results.

USDA's National Agricultural Statistics Service (NASS) and Economic Research Service (ERS) have been collecting annual, farm-level economic data for more than twenty years in what is now known as the Agricultural Resource Management Survey (ARMS.)[1] The ARMS is critical to the research and analysis mission of the Economic Research Service, and is a key input to estimates across the Department of Agriculture and in other agencies. It is a valued and unique resource, since it is the only national survey from which observations of field-level farm practices, the economics of the farm business, and the characteristics of the household operating the farm are all collected annually in a representative sample.

While ARMS provides a rich set of information, customers outside the government, especially researchers, have gone largely under-served (U.S. General Accounting Office, 1992). Data from ARMS had been available only within USDA except when summary information was released publicly, when users requested ERS do special tabulations, or when academic researchers entered into special research agreements with ERS and NASS to access the data in USDA offices to test specific hypotheses. The limited access to ARMS data frustrated those who saw the broad benefit of the information it contained. This friction between data access and data confidentiality is a common occurrence, particularly as it pertains to Government data. The Committee on National Statistics (CNSTAT) recently conducted a comprehensive review of the risks and opportunities of expanding access to confidential data. Their report highlights the main issues and documents current procedures and solutions employed by government agencies. (Panel on Data Access for Research Purposes, 2005). In essence, the trade-offs involve the desire to get the highest return possible for substantial data collection costs and respondent burden to gather information necessary to produce official statistics and support economic research on one hand and the requirement to uphold the pledge of confidentiality and ensure the future participation of respondents.

To expand external researcher access, ERS and NASS developed dynamic, technologically advanced, and easy to use web-based data delivery tools that are readily available through the ERS website (www.ers.usda.gov). Internet services have become part of a comprehensive suite of on-line services offered by the agency for its external customers. Traditional means of disseminating this vital but sensitive information did not meet the needs of users who we found wanted better access, transparent processes, and the ability to work with the data on demand; not limited to sets of pre-programmed tables. The wealth of data was dispersed across the website and hard to find. The new suite of tools that provide selective access to ARMS data not only expand access to farm survey data as a public good, but maintain the security of the confidential data. Researchers now have instant access to tailored information about agricultural production technology, farm business viability, and

[1]For a detailed perspective on the origin and use of ARMS as a principal USDA survey, see Johnson and Morehart (2006).

the structure of U.S. agriculture. This paper describes the framework that guided development of dissemination tools and explains procedures used to manage data access with adherence to confidentiality. We will also provide a perspective about some of the lessons learned in the construct of the tools and their acceptance by various stakeholders. The final section of the paper presents some recent improvements and future direction of the project.

2 Developing a two-tiered data dissemination system

One outcome of the initial planning stages for improving data access and dissemination was the recognition of two distinct audiences for ARMS information.[2] A large portion of the customer base was interested in having on-line access to summary tabulations of the data. Another group, primarily researchers, wanted the ability to access the raw data from their desktop and perform statistical analysis. With this in mind, two separate development tracks were initiated. One involved building a user friendly web tool that enabled users to select among survey data sets to build custom reports, refine queries with specific populations, group summary statistics for comparisons, and choose among several output options for results. The second track envisioned development of an experimental remote access for registered users (via a secure Extranet) to perform statistical analysis and economic modeling. Common to both of these products was the need to develop delivery methods and security protocols that ensured data confidentiality.

Considerable effort has gone into developing disclosure limitation methods for tabular data that effectively lower disclosure risk and provide useful products to data users. These techniques include cell suppression, local suppression, global recoding, rounding, and various forms of perturbation. Cell suppression techniques have been around for some time and were part of our internal data masking procedures prior to publishing survey results. A variety of automated and complex routines have been offered with varying degrees of success (see, for example, Kirkendall and Sande (1998), Fischetti and Salazar (2000), and Giessing (1999)). With guidance from these studies, we wrote a series of SAS IML macros to implement a sophisticated cell-suppression algorithm. Although table specific, it did recognize data relationships within rows of the predefined tables and across columns of output. The primary disclosure rules being implemented in the algorithm are know as the (n, k) rule. The n part of the rule identifies a threshold (3 observations in this case) for which sample size is small enough that the possibility of re-identification is too high. The k part of the rule establishes a threshold for dominance (60 percent in this case) where identity and attribute disclosure risk become too high when a single observation accounts for the threshold amount or more of the total estimate for the cell item considered.

[2]Several focus groups and usability studies were conducted during 2003.

The final, and perhaps most challenging, aspect of the cell-suppression routine, was defining heuristics for complementary cell suppression. The secondary cell suppression problem is to apply these complementary suppressions to the set of sensitive cells in such a way as to ensure that the complementary suppressions create the required uncertainty about the true values of the sensitive cells, while still preserving as much information in the table as possible. Using known equation relationships the equation checking functionality (ECF) was designed to recommend to the requested application an appropriate list of variables to obscure in order to prevent solving across row variables to determine the cell value that failed legal disclosure. The process is as follows. First, a cell check is completed to determine the row variable names that fail legal disclosure. Second the row variable name is passed to the ECF. The ECF routine utilizes the existing row variable list and the lookup table defining the implicit functional relationships to determine the optimal answer, consisting of variable names to obscure. The ECF procedure is designed to select the optimal variables, which is the minimum number of variables, to complete the obfuscation task. Dimensionality is a curse with this type of brute force program in terms of efficiency. However, with the functional structure of the ARMS data tables this was not of significant consequence.

An additional controlling feature of the tabular summaries was that they followed standard accounting guidelines, which prescribed the row content, but allowed some flexibility in the level of aggregation. So, for example, categories of expenses for which there were limited responses (having greater potential for disclosure risk) such as livestock leasing were combined into a more general category of other livestock-related expenses. Micro-aggregation is a data perturbation method, characterized by the publication of only small aggregates instead of the original data. Other aggregation approaches involve combining geographic areas or other attributes to minimize disclosure risk. One example of this approach is the Web-based query systems developed by the National Institute of Statistical Sciences (NISS) that disseminate NASS data on usage of agricultural chemicals (fertilizers, fungicides, herbicides and pesticides) on farms. This principal of aggregation was also applied to classification variables used in the tables created from the ARMS.

Once the disclosure routines were in place, we were able to develop prototype delivery systems for evaluation on internal servers. The first working system architecture placed a web interface in front of a dynamic SAS Intrnet data query structure. The user interface and menu utilized html coding and Microsoft .Net capabilities. The backbone of the system involved three servers, a web server, SAS server, and separate server to house the encrypted data (figure 1). User input was fed to the SAS broker, which decomposed the query string into SAS parameters. Data were then read into SAS based on the request. Requested statistics were then calculated along with measures of statistical reliability (relative standard error). The next step imposed the primary and complementary cell suppression algorithm. The final step

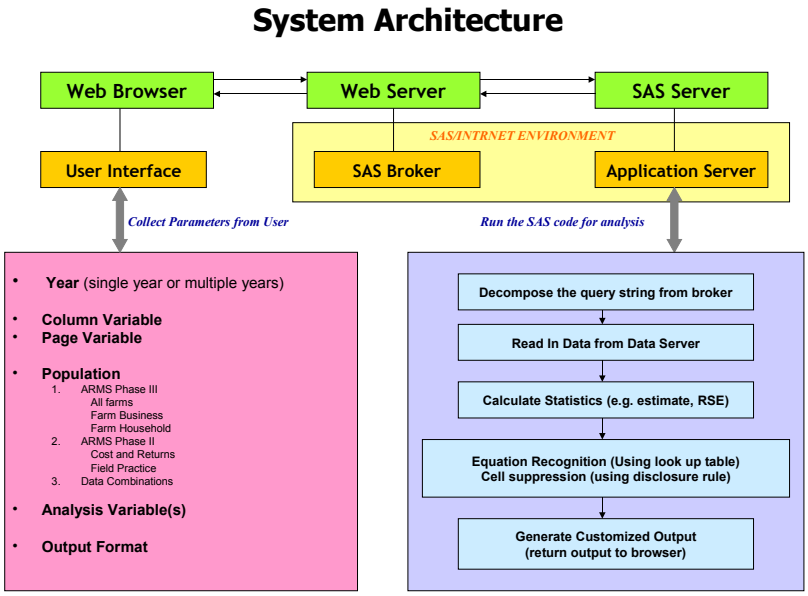


Figure 1: Initial data delivery system structure

generated a masked table for return to the web browser. This same backbone was used to service the advanced research user interface.

A peer review meeting that consisted of 32 attendees across two USDA Agencies (ERS, NASS), as well as individuals from other areas of the Federal Government was held on May 10, 2004. Its purpose was to provide a forum for open discussion of the ERS/NASS initiative to improve ARMS data dissemination via a controlled access web delivery tool, with specific regard to security and data confidentiality issues. The format consisted of a series of presentations by ERS and NASS staff that provided background on the Agricultural Resources Management Survey, current data dissemination methods, and the proposed solution to more broadly deliver ARMS data to the user community. The reviewers from five external agencies asked probing questions to which ERS and NASS staff responded. At the conclusion of the presentations, the reviewers were asked for their reactions to and concerns about any and all of the issues raised during any part of the presentations, to make suggestions for improvements, and to assist the project team by guiding its members to better solutions.

3 From prototype to final product

Peer review made an invaluable contribution that identified strengths and weakness of the working prototypes and established the major improvements necessary prior

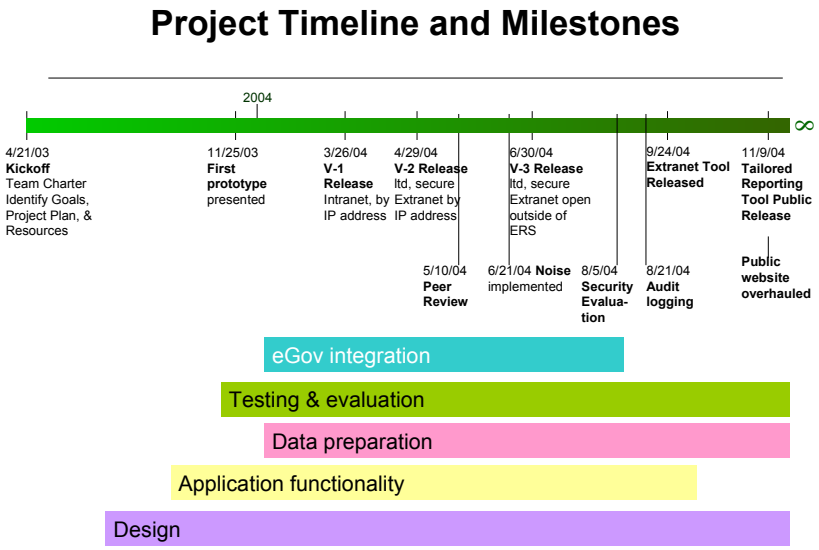


Figure 2: Project goals and timeline

to implementation. There were three primary areas of concern 1) strengthening data security and protecting confidentiality, 2) delivery speed and system load capabilities, and 3) access tracking capabilities. External contractors were consulted in addressing some of these issues. Advances in computer technologies and software also facilitated some improvements to the system architecture. Figure 2 show the project time line from first prototype to a public release of the tools and highlights major activities during this period.

An inherent layer of confidentiality protection for the ARMS is that it is a sample survey rather than a census. For sample surveys, estimates are made by multiplying an individual respondent's data by a sampling weight before they are aggregated. Since sampling weights are not published, this weighting helps to make an individual respondent's data less identifiable from published totals. Only providing weighted summary statistics also provided an opportunity to implement data masking directed toward the weights themselves. This involved adding zero mean noise to the determination of weights for each respondent such that the effect of the noise on items that were not at risk for disclosure was minimized (Evans et al. (1998), Duncan, George T. and Mukherjee, Sumitra (2000), and Camden et al. (2003)). In sample surveys, each respondent's data is generally weighted inversely proportional to the probability of being selected in the sample. Individuals with the lowest valued weights are those most at risk for disclosure. With noise added to the weights maintained in the survey data base, we retained the (n, k) cell primary suppression

rules and added additional rules that flagged cells that contained a large percentage of noise.[3] As noise was imposed in the raw data at NASS, these protections were one component of the security applied to the advanced statistical analysis Extranet. Additional precautions were added to the ARMS Extranet Online Tool to limit statistical analysis to no less than 30 samples, trim the upper and lower distribution tails so that minimum and maximum queries did not disclose individual values, and exclude any potential identifier variables including the weights (Allen (1992)).

The original prototype which used real-time SAS running in the background against web queries was too slow for Internet delivery, was problematic with high use loads, and had significant system maintenance costs. To address this we reformulated the system architecture for the dissemination of tailored reports to include two new additional steps. First, all table cells would be calculated and stored in an SQL database with the appropriate suppression algorithms applied. The web user interface was then reconfigured to query against this SQL data base. As a result, processing time was dramatically improved, and system maintenance reduced to semi-annual processing to create the SQL database. The new system architecture for tailored reports also provided greater flexibility in the type of tabular presentations (for example showing years as columns rather than just classification variables) and more easily accommodated adding graphing capabilities. The advanced statistical Extranet application continued to rely on a broker and SAS running, real-time, in the background of requests. There were some enhancements to this process that improved response speeds, but the majority of time was spent reconfiguring the interface based on usability test results.

The ability to track access and store the results of user submissions was advocated as a necessary component of the advanced statistical Extranet application. By signing the Confidentiality Agreement users agree that any data file provided to them "...will be used only for statistical reporting and analysis and will not be published or released in identifiable form." In this context, the term "statistical summary information" means the result(s) of statistical analysis in any of the following forms: record listings, frequency tabulations, magnitude tabulations, means, variances, regression coefficients, correlation coefficients, graphical displays and any other result of an analytic process. While the Extranet application has several built-in mechanisms to reduce the potential for disclosure, the ultimate measure of disclosure risk is visual inspection of output. To minimize the burden of this activity the web delivery system must have the capability to track access and store results in order to allow for output review.

[3]The specific procedure used to add noise is not known to data users and therefore not provided in this paper. This non-disclosure of the specific parameters is a necessary layer of confidentiality protection.

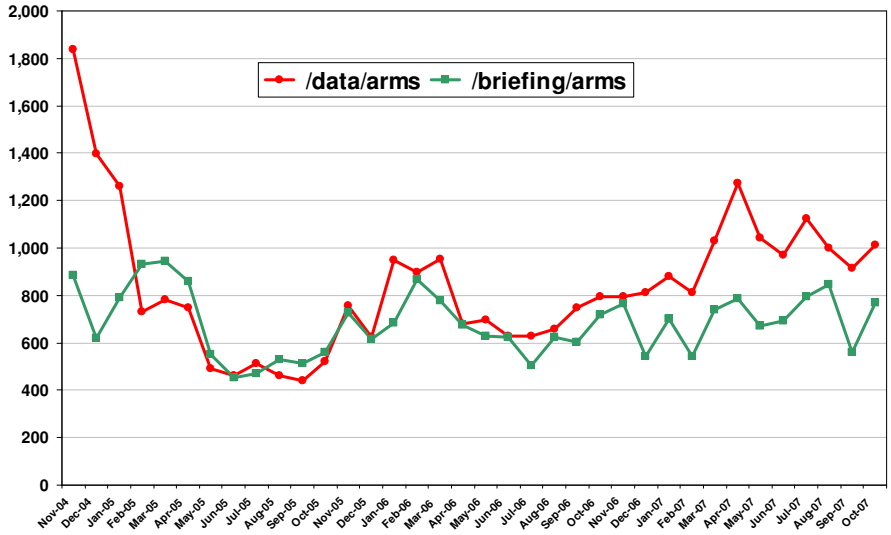


Figure 3: Unique monthly website visits

4 System performance

We are well into the third year of operating and maintaining the ARMS data delivery tools. Since its launch in November of 2004, the customized data summary tool has averaged about 800 unique visits per month (figure 3). It has become a main feature of the ERS website data page and has been widely accepted and used by our customer base. Online access makes analyzing natural resource, technology adoption, farm business, and farm household issues less costly and more efficient. One-stop shopping improves value and provides a usable, standardized method of obtaining ARMS data. Separately produced outputs, some with product specific formats and programming, dispersed throughout the website were replaced with a single robust product. The improved access to data has reduced demand on staff to help find information and in requests for special tabulations. Data consistency is better managed in the update process, the data is easier to find for users, and programming is centralized for improved sustainability. The centralized system also facilitates better access to survey procedures and documentation so that users know what they are getting with ARMS data.

The capability to interact with data users is an important feature of the customized data summary tool. We have received an average of 6 inquiries a month. Topics range from specific data questions to general comments about the web interface and usability of the tool. Several changes have been made as a result of user feedback. For example, the farm production specialty classification variable was modified in two of the featured states to accommodate separate analysis of fruits, vegetable, and nursery and greenhouse operations that are normally collapsed into one category called specialty crops.

The advanced statistical analysis component of the web delivery tools that is provided via secure Extranet has had limited use and is much more resource demanding to update and maintain. Since its launch, there have been fewer than 30 users that have accessed the system. Beyond the initial requirements of having a memorandum of understating that defines research goals and uses of the data and signing the confidentiality agreement, users are required to obtain a customer-level USDA eAuthentication ID. This involves some additional paper work and appearing in person at a local USDA designated Service Center. The advanced statistical tools that are currently available on the ARMS Extranet Online Tool are limited to exploratory variable summaries and linear regression analysis. This system has made access more convenient, but only provides data users the ability to do a preliminary evaluation of their research application.

5 Future Enhancements

The high level of user satisfaction and relatively low maintenance costs of the web tool that provides summary tabulations suggests that the primary focus for future development is better meeting the needs of researchers that want a convenient way to conduct statistical analysis. The capability provided by continued improvement of computer technologies has widened the scope of possibilities for access to restricted data (Wolf (2002) and King (2007)). Considerations for system costs and maintenance and delivering a high level of user services also are important, particularly for relatively small agencies such as the Economic Research Service. The wide acceptance and use of the tabular summaries provided on the ERS website had stimulated interest among data users, so there is no anticipation of a reduction in the demand for access. The other consideration, that we initially underestimated, is the sophistication of researchers in terms of computer software and data analysis. Their feedback confirms that most researchers prefer to have a more direct capability to write and submit code against the data rather than a complex web menu system that works as a front end for code processing.

With this in mind, we conducted a comprehensive review of existing systems that allow remote access to restricted data. Some of the systems reviewed included 1) the Luxembourg Income Study System (LIS), 2) Remote Data Access (RDA) of

Statistics Canada, 3) Remote Access Data Laboratory (RADL) of Australian Bureau of Statistics (ABS), and 4) Research Data Center (RDC) of National Center for Health Statistics (NCHS). There were many common features across these systems such as an email or web interface and allowance for a variety of statistical processing code (SAS, SPSS, STATA, etc.). In each of these systems, security protocols were put in place such that researchers did not have direct access to microdata, results were subjected to confidentiality review before being sent back to the user, and there were usage logs kept. While construction of a similar system at ERS is feasible, costs are prohibitive, particularly as the number of users increase.

ERS and NASS recently initiated a 2-year pilot project to examine the feasibility of remote access to ARMS data using the National Opinion Research Center (NORC) data enclave.[4] Researchers from eight universities across the United States with varying experience using the ARMS data were selected to participate in the project. The enclave is designed to provide a secure mechanism for producers of sensitive data to enable more convenient access to approved researchers. Researchers will be able to access ARMS data from their office desktop computers using a secure environment. The enclave is setup as a collaborative environment so that researchers can share documentation and code. Research output is reviewed for confidentiality disclosure by NORC and the sponsoring institutions.

Through this pilot project, ERS will obtain valuable feedback on the performance and necessary improvements to the data enclave, more comprehensive documentation of the data properties in the form of metadata, enhanced training for ARMS users, and determination of the feasibility of remote accesses to facilitate collaborative research using ARMS data. Cooperators will benefit through participation in training to enhance their ability and efficiency in working with ARMS data and in the completion of research that informs local farm issues. In addition to data warehousing and confidentiality protections, NORC will set up and manage an active outreach program to inform the national research community of the data and to foster the use of the data in research leading to conference presentations and journal publications. NORC also will establish an extensive education program to ensure appropriate use and disclosure of the data, including confidential aspects of the data.

References

Allen, R. (1992), "Statistical Defensibility as Used by U.S. Department of Agriculture, National Agricultural Statistics Service," *Journal of Official Statistics*, 8, 481–498.

Camden, M., Daish, K., and Krsinich, F. (2003), "The Noise Method for Tables - Re-

[4]For more information, see: http://dataenclave.norc.org/.

search and Applications at Statistics New Zealand," in *Joint ECE/Eurostat work session on statistical data confidentiality*, Luxembourg: United National Statistical Commissionand Economic Commission for Europe Conference of European Statisticians, no. Working Paper 28.

Duncan, George T. and Mukherjee, Sumitra (2000), "Optimal Disclosure Limitation Strategy in Statistical Databases: Deterring Tracker Attacks through Additive Noise," *Journal of the American Statistical Association*, 95, 720–729.

Evans, T., Zayatz, L., and Slanta, J. (1998), "Using Noise for Disclosure Limitation of Establishment Tabular Data," *Journal of Official Statistics*, 14, 537–551.

Fischetti, M. and Salazar, J. J. (2000), "Complementary Cell Suppression for Statistical Disclosure Control in Tabular Data with Linear Constraints," .

Giessing, S. (1999), "A Survey on Software Packages for Automated Secondary Cell Suppression," .

Heckman, J. J. (2001), "Micro Data, Heterogeneity, and the Evaluation of Public Policy: Nobel Lecture," *Journal of Political Economy*, 109, 673–748, available at http://ideas.repec.org/a/ucp/jpolec/v109y2001i4p673-748.html.

Johnson, J. and Morehart, M. (2006), *The Wye Group Handbook: Rural Households' Livelihood and Well-Being*, UNECE, Eurostat, FAO, OECD, World Bank, chap. Income and Wealth Statistics for Selected Countries: The Agricultural Resource Management Survey (ARMS), pp. 1–30, Chapter 14.1.1.

Karr, A., Lee, J., Sani, A., Hernandez, J., Karimiand, S., and Litwin., K. (2000), "Web-Based Systems that Disseminate Information from Data but Protect Confidentiality," Tech. rep., National Institute of Statistical Sciences.

King, G. (2007), "An Introduction to the Dataverse Network as an Infrastructure for Data Sharing," Tech. rep., Harvard University.

Kirkendall, N. and Sande, G. (1998), "Comparison of Systems Implementing Automated Cell Suppression for Economic Statistics," *Journal of Official Statistics*, 14, 513–535.

Lane, J. (2003), "Uses of Microdata: Keynote Speech," in *Statistical Confidentiality and Access to Microdata: Proceedings of the Seminar Session of the 2003 Conference of European Statisticians*, Geneva, pp. 11–20.

Nolte, M. A. and Keller, J. J. (2004), "Research Use of Restricted Data: The HRS Experience," Joint Statistical Meetings, Toronto.

Panel on Data Access for Research Purposes, N. R. C. (2005), *Expanding Access to Research Data: Reconciling Risks and Opportunities (2005)*, The National Academic Press.

U.S. General Accounting Office (1992), "Data Collection: Opportunities to Improve USDA's Farm Costs and Returns Survey," Publication GAO/RCED-92-175, Washington, DC.

Wolf, V. D. (2002), "Issues in accessing and sharing confidential survey and social science data," *Data Science Journal*, 2, 66–74.

Integer Rounding versus Continuous Adjustment for Tabular Data

Juan-José Salazar-González*

* DEIOC, University of La Laguna, Tenerife, Spain. (jjsalaza@ull.es)

Abstract. Controlled Rounding and Adjustment are two perturbation techniques in fashion as an alternative procedure for guaranteeing confidentiality during tabular data publication. To apply each technique on large data, automatic tools based on modern optimization procedures are necessary. In this paper we discuss the advantage and disadvantage inherent to mathematical models for both techniques.

1 Introduction

Statistical agencies collect data to make reliable information available to the public. This information is typically made available in the form of tabular data (i.e., a table), defined by cross-classification of a small number of variables. A fundamental characteristic of all tables is the existence of mathematical equations. Each equation says that a cell value (marginal cell) is identical to the sum of other values (internal cells). Depending on the size and structure of the table, the set of equations may create a complex linear system of equations, which may have a negative impact when applying some methodologies to protect private information. See Salazar [4] for a survey of articles concerning approaches for protecting tables.

Controlled Rounding consists of replacing each cell value by a multiple of a prespecified base number (e.g. 5). There are several variants of the methods, but the better accepted is the one where

1. original cell values which are multiple of the base number must remain unchanged;

2. other cell values must be replaced either by the minimum multiple which is larger or equal to the original value, or the maximum multiple which is smaller or equal to the original value;

3. the modified table must satisfy the same system of linear equations as the original table.

Figure 1 shows an example of unrounded table, which in Figure 2 has been rounded using base number 5. When the table structure satisfies some conditions (e.g., the

Unrounded data	total	male	female	young	adult	thin	fat
North East	60593	29225	31368	13856	46737	34565	26028
North West	174414	78129	96285	25673	148741	3432	170982
Yorkshire and Humberside	108769	46119	62650	2342	106427	32223	76546
East Midlands	93346	43201	50145	23443	69903	23434	69912
West Midlands	131817	61046	70771	23878	107939	432	131385
East	107060	47376	59684	24532	82528	34233	72827
London	110811	49053	61758	17635	93176	3423	107388
South East	123359	50949	72410	34223	89136	4567	118792
South West	119863	44718	75145	35980	83883	56356	63507
England	1030032	449816	580216	201562	828470	192665	837367
Wales	95388	49579	45809	34989	60399	6454	88934
Scotland	124678	61327	63351	36789	87889	5643	119035
Great Britain	1250098	560722	689376	273340	976758	204762	1045336

Figure 1: Original (unprotected) table.

Rounded data (base=5)	total	male	female	young	adult	thin	fat
North East	60595	29225	31370	13855	46740	34565	26030
North West	174415	78130	96285	25675	148740	3430	170985
Yorkshire and Humberside	108770	46120	62650	2340	106430	32225	76545
East Midlands	93345	43200	50145	23445	69900	23435	69910
West Midlands	131815	61045	70770	23875	107940	430	131385
East	107060	47375	59685	24530	82530	34235	72825
London	110810	49055	61755	17635	93175	3420	107390
South East	123360	50950	72410	34225	89135	4570	118790
South West	119860	44715	75145	35980	83880	56355	63505
England	1030030	449815	580215	201560	828470	192665	837365
Wales	95390	49580	45810	34990	60400	6455	88935
Scotland	124675	61325	63350	36790	87885	5640	119035
Great Britain	1250095	560720	689375	273340	976755	204760	1045335

Figure 2: Modified (protected) table.

cells can be represented by arcs in a network) a modified table exists, and it is known how to find a closest one to the original table with an efficient approach (e.g., a min-cost flow algorithm). However, for a general structure the problem of finding a modified table may be infeasible and some variants have been proposed in the literature (see, e.g., Salazar [3]). The better accepted variant relax the conditions (1) and (2), so a modified value is not necessary an adjacent multiple to the original value. Finding a solution to this extended model implies solving an Integer Linear Programming model, which is known to be be (in general) a complex mathematical problem (e.g., $\mathcal{NP}$-hard in Complexity Theory). This classification means that there are examples of tables where it is very difficult to find a modified table, and this has motivated the research of alternative methodologies.

Tabular Adjustment is an alternative approach to Controlled Rounding. It was originally proposed by Dandekar and Cox [1], and it consists of

- deciding whether each sensitive cell value should be rounded up or down;
- determining the continuous value for each non-sensitive cell value.

A mathematical formulation to find a solution also contains integer variables, but only for the sensitive cells. The non-sensitive cells are associated to continuous variables, which leads to a Mixed Integer Programming model. The problem of finding a Tabular Adjustment solution is again $\mathcal{NP}$-hard, but in practice it is much easier than a problem of finding a Controlled Rounding solution because the number of integer variables is smaller. Note that solving a mathematical problem with only continuous variables is easy ($\mathcal{P}$ in Complexity Theory). Other similar methods have also been proposed in the literature (see, e.g., Cell Perturbation in Salazar [3]), based on a different understanding of protection, but exploiting the advantage of simplifying the problem resolution by having continuous mathematical variables instead of integer mathematical variables.

Although replacing some integer variables by continuous variables may help to solve in practice a model, this paper points out some disadvantages that one should have in mind when replacing Controlled Rounding by a Continuous Adjustment.

2 Linear-programming relaxations

To illustrate some negative consequences of using continuous variables instead of integer variables, let us analyze the following mathematical problem:

$$\begin{array}{rcl} \min\ x_0 & & \\ 75000\,x_0 & = & 75001\,x_1 + 75002\,x_2 \\ & & x_0 \geq 1, x_1 \geq 0, x_2 \geq 0 \\ & & x_0 \in \mathbb{Z}, x_1 \in \mathbb{Z}, x_2 \in \mathbb{Z} \end{array}$$

This is an artificial simple example with three cells and one linear equation. It does not correspond to any table in practice, but the small size will help us to make clear the main observation of this paper.

Solving the integer mathematical model is difficult in practice. Indeed, using the best commercial solver (like Cplex or Xpress) will take more than one hour on a modern personal computer before finding an optimal solution. The difficulty of this problem is clearly not on the size, but on the integrability of the variables. If the integer variables are replaced by continuous variables then the problem becomes trivial:

$$x_0 = 1 \quad , \quad x_1 = \frac{75000}{75001} \quad , \quad x_2 = 0.$$

There is no need of a sophisticated solver for finding this trivial solution. However, when the variables must be integer, then a sophisticated solver is fundamental, and using this solver we will find:

$$x_0 = 37502 \quad , \quad x_1 = 2 \quad , \quad x_2 = 37499.$$

The immediate conclusion when comparing the two solutions is that both can be very far one from the other. Indeed, in theory, for any large number M, it is possible to design an instance where the integer and the continuous solutions are farther than M. The above example shows that this situation may happen in practice, even with tiny numbers of cells and equations.

3 Conclusion

The previous section has shown that the solution of the Linear Programming relaxation of an integer program may be very different from its integer solution. Then, using alternative methodologies where integer variables are replaced by continuous variables may create easier-to-solve models but wrong-to-use solutions.

In addition, it is also obvious that continuous variables contain decimal part. This is a serious drawback for protecting tables with frequency data, but also with magnitude values due to numerical errors during the computation and displaying. In fact, when using magnitude data one does not want to public cell values like 345.0000001, and the simple task of eliminating the decimal part of the numbers (i.e., rounding) may create non-additive tables. Also, if one wants to display the continuous solution of the above optimization problem with only four decimals, number 75000/75001 = 0.9999866668... would be replaced by 1.0000, thus leading to the non-additive solution

$$x_0 = 1.0000 \quad , \quad x_1 = 1.0000 \quad , \quad x_2 = 0.0000.$$

Hence, after applying a methodology (like Tabular Adjustment), the Controlled Rounding is mandatory unless we have been *lucky* with the original table and the

modified values (continuous numbers) are suitable to be published as they come out from the methodology.

The use of continuous variables in a methodology (as in Controlled Rounding) may reduce the computational complexity of finding a solution, but depending on the table itself the found solution may contains fractional values with a significant decimal part. The finite precision of computers and the necessary truncation of decimals during the publication phase require the use of Controlled Rounding to guarantee additivity. In other words, *only Controlled Rounding guarantees that the modified table satisfies the same linear equations as the original table.*

A final remark is that the above example belongs to a class of optimization problems (with one equation, no matter the number of variables) which can be solved in a very efficient way by using dynamic programming. For solving instances of this class a general-purpose commercial software (like Cplex or Xpress) is not convenient. Indeed, it takes less than one second to solve the above instances by dynamic programming on a computer. This positive result is the outcome of a research work done on the model, and remark also the importance of analyzing mathematical models instead of using a commercial software as a black-box solver. In other words, the fact of having a mathematical model for a new disclosure limitation methodology is not the end of a research line, as it may be of interest to study ad-hoc approaches to solve it.

References

[1] Dandekar, R.A., Cox, L.H. Synthetic Tabular Data: an alternative to complementary cell suppression for disclosure limitation of tabular data. Technical report (2002).

[2] Salazar, J.J. A Unified Mathematical Programming Framework for Different Statistical Disclosure Limitation Methods. *Operations Research* **53** (2005) 819–829. http://dx.doi.org/10.1287/opre.1040.0202

[3] Salazar, J.J. Controlled Rounding and Cell Perturbation: Statistical Disclosure Limitation Methods for Tabular Data. *Mathematical Programming* **105** (2006) 583–603. http://dx.doi.org/10.1007/s10107-005-0666-4

[4] Salazar, J.J. Statistical confidentiality: Optimization techniques to protect tables. *Computers & Operations Research* **35** (2008) 1638-1651. http://dx.doi.org/10.1016/j.cor.2005.09.009

III

Applications (SDC methods, issues within NSIs and software)

Rounding methods for protecting EU-aggregates

Sarah Giessing[1], Anco Hundepool[2] and Jordi Castro[3]

[1] Statistisches Bundesamt, 65180 Wiesbaden, Germany, Email: Sarah.Giessing@destatis.de
[2] Statistics Netherlands, Division of Methodology and Quality, P.O. Box 4000, 2270 JM Voorburg, The Netherlands, Email: ahnl@cbs.nl
[3] Universitat Politecnica de Catalunya, Department of Statistics and Operations Research, Jordi Girona 1–3, 08034 Barcelona, Catalonia, Spain, Email: jordi.castro@upc.edu

Abstract. In the European Statistical System the statistical information is collected by the National Statistical Institutes (NSIs). The NSIs produce aggregate tables at the national level. They are also responsible for proper protection of these tables and hence they have to keep certain cells confidential, suppressing them from publications. Eurostat produces statistical information at the EU-level. However, the national suppressions hamper very much the publication of EU-aggregates although it is often only a few smaller countries having to keep their contribution to the EU-total confidential.
This paper reports on a research-project that aims for making more EU aggregates available whilst at the same time guaranteeing the national suppressed figures to remain confidential.

Keywords. EU-aggregates, Controlled rounding, Controlled tabular adjustment, Cell-suppression, Interval protection.

1 Introduction

The NSIs in Europe collect a lot of statistical information and publish many statistical tables at the national level or below. They are also responsible to take care of the confidentiality aspects of their publications. In quantitative tables this implies often that several cells have to be suppressed due to confidentiality reasons. Cell suppression is the traditional way of protecting a statistical table. See for example the CENEX-SDC handbook (Hundepool et al, 2006).

The NSIs also deliver data to Eurostat. Eurostat aggregates the national data to tables at the European level. In this paper we study the tables from the production statistics (Prodcom) and SBS (Structural Business Statistics). These tables are broken down by geography (down to the member state level) and in the case of the SBS data by a hierarchical NACE classification.

Confidentiality charters have been agreed with the Member States for the data collected in the respective frameworks of Prodcom and SBS Regulation. These charters describe amongst other issues when an EU-aggregate can be published, given the national published and sometimes suppressed cells. In many cases these rules prevent publication of EU-aggregates. If, for instance, only one country is confidential, the EU-aggregate must not be published, because otherwise this confidential value could be computed by taking the difference between the EU-aggregate and the non-confidential member state figures which is a typical instance of "disclosure by differencing".

Using certain constraints on the cell values of the tables which are known independent from the publication (like f.i. non-negativity of cell values) it is possible to compute *feasibility intervals* (a minimum and a maximum bound for the set of feasible values) for each suppressed cell of a publication, for instance by solving two linear programming (LP)-problems per cell. Any user of the publication would in principle be

able to perform such an analysis. A table is protected properly, if all the feasibility intervals satisfy certain requirements, e.g. if they create a certain amount of uncertainty about the true cell value. For discussion of these requirements see (CENEX-SDC handbook, section 4.2.2). Technically these requirements can be expressed as *protection intervals* which must be covered by the feasibility intervals.

As the confidential national cell (often a cell of a smaller country) frequently makes only a marginal contribution to the EU-aggregate, the corresponding protection interval although perhaps rather large at the national level is often only marginal at the EU-level. So, small confidential contributions with relative small protection interval impede the publication of much larger EU-aggregates. It should therefore be possible to 'save' the EU-aggregate by introducing a relatively small amount of uncertainty into it. This can be achieved by replacing the true value of aggregates by approximations like for instance rounded versions of the true value, or by replacing them by intervals or by adding some random perturbation to the aggregates. Approximations have to be determined as to provide sufficient protection to confidential aggregates. This implies for instance that the bounds of rounding intervals must be at safe distance from the true value of a confidential aggregate.

Publication of approximations makes sense of course only, if users understand well the difference (in terms of reliability) between the true and the approximated values. For general purpose data such as the European SBS aggregates, rounding approaches seem to be appealing because rounded figures are easy to interpret even by a naïve user.
In the remainder of this paper we will describe the solutions proposed for the Prodcom and SBS tables. Section 2 proposes controlled rounding for the protection of Prodcom data, whereas section 3 suggests another rounding method for the SBS data.

2 Rounding method for the Prodcom tables

The Prodcom tables have a rather simple structure. European Prodcom aggregates are reported only at the lowest level of the NACE hierarchy and therefore there are no higher level NACE-aggregates. Technically this reduces the large Prodcom table to a large set of smaller tables at the lowest NACE classification. Only some hierarchy in the geography has to be taken into account. For the tables from before 2003 the EU25 is broken down by EU15 and EU10, while for the more recent years EU27 is broken down by EU25 and EU2 (=Romania + Bulgaria).

The member states do the confidentiality protection for their tables themselves and decide which cells have to be suppressed. Because the higher level NACE codes are not published there is no additive relation between aggregates. Therefore, only primary suppressions have to be assigned. When transmitting the data to Eurostat, the member states also provide Eurostat with the cell values of confidential cells, but flag them as confidential. Also they provide the nature of this confidentiality. This can be an unsafe cell due to too few contributions (frequency rule) or due to a violation of a dominance or p% rule threshold. In case of a frequency unsafe cell the member states also report the number of respondents while for a dominance unsafe cell the percentage of the contribution of the largest or largest 2 contributors is given.

Although Eurostat cannot publish these unsafe cells at the member state level, it can use this information to compute the EU-aggregates. And if no member state information is

confidential or a sufficient number of member states is confidential, the EU-aggregate can still be published according to the rules of the Prodcom Confidentiality charter.

For those situations where the EU aggregate cannot be published, we propose a rounding procedure. Recently a controlled rounding procedure (c.f. Salazar-Gonzalez et al., 2006) developed on behalf of ONS was included in the statistical disclosure control software τ-ARGUS. Unlike traditional deterministic or probabilistic rounding methods, this controlled rounding method is able to guarantee that the special protection requirements of tabulations of establishment data are satisfied. For a given table with sensitive cells, the method computes the closest rounded table that is additive subject to certain constraints. These constraints ensure that the rounding interval for any confidential cell covers the corresponding protection interval.

The special procedure implemented for the Prodcom data first decides on the minimal rounding base, given the protection intervals of the confidential member states. These protection intervals are computed on the basis of the additional information of the unsafe cells supplied by the member states. Then the τ-ARGUS rounding procedure is applied. Sometimes the initial rounding base may not provide enough protection and then the rounding base will be increased.

Initially we had in mind to restrict the procedure to rounding bases as a powers of 10 (10, 100, 1000, …); procedure 1. However sometimes this resulted in rather large rounding bases and larger information loss. So a more refined series was then adopted (10, 20,…,90, 100, 200, 300,…,900, 1000, 2000,…); procedure 2. This led to solutions with enough protection, but less information loss. Only in the publication it requires a bit more explanation.

Of course the rounding procedure cannot hide the already published national safe figures. Before applying the rounding procedure, these safe, published cells have been merged into one cell. The exact value of these cell combinations has been considered to be known (as information available to a possible intruder) when stating the protection of the table as controlled rounding problem.

Rounding base (% of EU-total)	Frequency	
	Proc.1	Proc.2
0 -< 1	23	70
1-< 5	190	277
5-<10	82	109
10-<20	92	58
20-<50	107	15
Over 50	39	4

Table 1: Distribution of the rounding bases used

As can be seen from table 1 in many cases a solution can be found with only limited information loss. In the majority of cases the rounding base is less than 10 % of the EU total. Cases where the rounding base is larger than 50 % of a EU-total are a rare exception. As we cannot modify the already published tables the result of this procedure is, that the required protection interval is wide enough, but sometimes a bit shifted. Nevertheless the size of the interval guarantees enough protection.

3 Rounding method for the SBS tables

Because of its more complex data structures, the rounding procedure proposed for the Prodcom case cannot be expected to work well in the SBS case. Unlike in the Prodcom case, there is a detailed hierarchical relation between SBS aggregates, because they are published on the EU-level at 5 different levels of the NACE classification. The τ-ARGUS controlled rounding method rounds all aggregates of a table to multiples of one rounding base. While for the protection of large confidential aggregates at high NACE levels a large rounding base would have to be chosen, this kind of rounding would lead to too much information loss on the lower NACE levels.

In the following we propose a rounding procedure, which – just like controlled rounding –is able to guarantee that the specific protection requirements of magnitude tables from business surveys are satisfied, i.e. it provides enough uncertainty round each primary unsafe cell. The procedure comprises several tasks which will be explained in 3.1. The following section 3.2 outlines an alternative methodology based on interval protection. Section 3.3 reports some test results. Finally, section 3.4 describes some ideas for future work.

3.1 Rounding Procedure based on Restricted CTA

We first compute protection intervals, and bounds on the cell values assumed to be general knowledge, taking care in particular of those EU-aggregates where the confidential cluster consists of one or two member states with only one single contributor (so called '*singletons*') in either of these two states. In those cases we must avoid the risk for instance that one of two singleton companies can use special knowledge (e.g. of its own contribution) to undo the protection provided (by the rounding) to the other singleton company.

We then apply Restricted Controlled Tabular Adjustment proposed in (Castro and Giessing, 2006) to compute an *adjusted table* that contains some true, original and some approximate ('adjusted') values with the following properties: The adjusted table is, according to some suitable measure of distance, the closest additive table to the original table satisfying the following constraints:

- the adjusted values of *all* confidential cells are safely (considering the protection levels assigned as explained above) away from their original values,
- the adjusted values are within a certain range, i.e. for a variable with only non-negative values adjusted values also have to be non-negative, and
- adjusted values for member state aggregates flagged as published must be identical to the original value.

Note, that the last constraint makes the procedure what we call a *restricted* CTA procedure.

The next step of the procedure is to compute rounded approximations for those cells on the EU-level that were subject to an adjustment in the RCTA step. We chose a suitable rounding base for each cell separately from the series (10, 20, …, 90, 100, 200, …, 900, 1000, 2000, …, 9000…). For each adjusted value we determine the rounding base b to be the smallest in the series which is larger than the distance between the true and the adjusted value. This property guarantees that it is possible to find a multiple $m*b$ of the

base, where the rounding interval [(m-1)*b+1; (m+1)*b-1] covers both, the true, and the adjusted value.

So far, our procedure now guarantees *sliding* protection but not in all cases *upper* protection for the confidential aggregates: As explained in the CENEX-SDC Handbook, 4.2.2 users of a table with suppressions can always compute a feasibility interval for any particular suppressed cell, i.e. they can derive upper and lower bounds for its true value. We assume now that for this kind of analysis users who attempt to compute feasibility intervals for confidential member state level cells take into account the rounding intervals for the rounded EU-level cells as *a priori* bounds. Our procedure so far guarantees that either the upper or the lower feasibility bound that can be computed in this way for a confidential member state level cell will be safely away from the true value. Assume now the member state level cell was declared confidential because of dominance, e.g. because the true cell value is an upper bound for the contribution of the dominant respondent that is considered too close. If now in the case of this cell the lower feasibility bound is safely away from the true cell value, but the upper feasibility bound is not, this means that - like the true cell value - the upper feasibility bound is an upper bound for this respondent contribution which is too close (this is our definition of 'not safely away'). Note that this is only a problem, if the reported variable takes only non-negative values, because otherwise it may happen that individual respondent contributions are larger than the cell value.

We can solve this disclosure risk problem for variables with non-negative values heuristically by extending the procedure in the following way: We first audit the rounding obtained so far by computing feasibility intervals considering the rounding intervals as just explained. If this audit establishes lack of upper protection we carry out an extra one-cell CTA procedure (one for each confidential cell lacking upper protection). One-cell CTA, targeted to one specific confidential cell, addresses upper protection of only this particular cell, i.e. it guarantees that the adjusted value of this cell will be larger than the true value, and safely away from it. The procedure is completed by another rounding step. This time we determine rounding bases b for the European level aggregates so that the corresponding rounding interval covers the smallest interval that contains the true value, the adjusted value from the RCTA procedure, and all the adjusted values from each of the one-cell CTA procedures. This extended procedure obviously guarantees sufficient upper protection of confidential member state level aggregates.

While the combination of restricted CTA and one-cell CTA guarantees sufficient upper protection of confidential member state cells, it does not guarantee that the set of intervals that is used to compute the rounding bases is the 'best' set of intervals satisfying our requirement. Optimal solutions for this problem could be achieved using a formulation as a large-scale linear optimization problem as outlined in the following section.

3.2 Outline of an interval protection methodology

We are given a table (i.e., a set of cells $a_i, i = 1,\ldots,n$, satisfying m linear relations $Aa = b, A \in \mathbb{R}^{m\times n}, b \in \mathbb{R}^m$). Any set of values x satisfying $Ax = b, l \le x \le u$, is a valid table, $l \in \mathbb{R}^n, u \in \mathbb{R}^n$ being known a priori lower and upper bounds for cell values. For positive tables we have $l_i = 0, a_i = +\infty, i = 1,\ldots,n$, but the procedure outlined is also valid for general tables.

Our purpose is to compute the set of smallest intervals $\left[lb_h, ub_h\right]$ for cells $h \in \mathrm{H}$ (in our instance, H is the set of EU-level cells) instead of the real value $a_h \in \left[lb_h, ub_h\right]$, such that, from these intervals, no attacker can determine that $a_s \in \left(a_s - lpl_s, a_s + upl_s\right)$ for all sensitive cells $s \in \mathrm{S}$. Introducing two auxiliary vectors $x^{l,s} \in \mathbb{R}^n$ and $x^{u,s} \in \mathbb{R}^n$ to impose, respectively, the lower and upper protection requirement, this problem can be stated as follows:

$$
\begin{aligned}
\min \quad & \sum_{i\in \mathrm{H}} w_i \left(ub_i - lb_i\right) & & & & \\
& & Ax^{l,s} &= b & & \\
s.t. \quad & l \le & x^{l,s} &\le u & & \\
& 0 \le lb_i \le & x_i^{l,s} &\le ub_i & i \in \mathrm{H} & \\
& & x_s^{l,s} &\le a_s - lpl_s & & \\
& & & & & \forall\ s \in \mathrm{S} \\
& & Ax^{u,s} &= b & & \\
& l \le & x^{u,s} &\le u & & \\
& 0 \le lb_i \le & x_i^{u,s} &\le ub_i & i \in \mathrm{H} & \\
& & x_s^{u,s} &\ge a_s + upl_s & &
\end{aligned}
$$

where w_i is a weight for the information loss associated with cell a_i.

Indeed this problem, in theory, it is simpler than optimal CTA, so it may be more efficient and provide a better solution than the procedure based on CTA, plus a post-process with one-cell CTA for unprotected cells.

3.3 Test results

The structure of the SBS test tables is 2-dimensional (by NACE and by country). Because Eurostat does not publish an overall cross-sectoral total, each table corresponds to only one particular NACE sector. Rounded approximations had to be computed for tabulations of two different variables. Variable 1 takes non-negative values only while variable 2 may also take negative values.

The procedure of 3.1 was applied to these data (so far only to the variable 1 tabulations). Only in the case of sector D we had to carry out one-cell CTA post-processing.

In the following, we present results obtained for tabulations of variable 1 for NACE sectors C, D and E. As an indicator for the loss of information caused by rounding a particular cell we use the percentage of the rounding base (in terms of the true value of the cell). The largest perturbations we observed were about 17 % in sectors C and D, and about 0.6 % in sector E. Table 1 presents the number of EU-level cells by range of these percentages.

Rounding base (in % of the cell value)	NACE-sector			
	C	D	E	C-E
0 %	16	259	1	276
(0%, 2%]	13	104	4	121
(2% , 5%]	0	4	0	4
(5%,10%]	3	0	0	3
> 10%	5	1	0	6

Table 1 No. of EU-level cells by rounding base percentage ranges for NACE sectors C, D and E

Overall, in the three sectors C, D and E 276 cells remained unperturbed (rounding base percentage 0%). Nearly half as much (121) were perturbed by less than 2 %. Only a few cells got larger perturbations.
Table 2 presents the results with respect to the hierarchical level of the cells in the table. It shows the distribution of cells (no of cells in %) by ranges of the perturbation percentages and by NACE level.

NACE-level	Rounding base (in % of the cell value)				
	0	(0%, 2%]	(2% , 5%]	(5%,10%]	> 10%
	Sector C				
4-digit	50	25.00	-	12.50	12.50
3-digit	46.15	30.77	-	7.69	15.38
2-digit	20	60	-	-	20
sub-sector	50	50	-	-	-
sector	-	100	-	-	-
	Sector D				
4-digit	80.18	18.94	0.44	-	0.44
3-digit	56.31	41.75	1.94	-	-
2-digit	30.43	65.22	4.35	-	-
sub-sector	78.57	21.43	-	-	-
sector	100	-	-	-	-
	Sector E				
4-digit	-	100	-	-	-
3-digit	-	100	-	-	-
2-digit	50	50	-	-	-
sector	-	100	-	-	-

Table 2 No. of EU-level cells (in %) by rounding base percentage range and by hierarchical level for NACE sectors C, D and E

Table 2 shows that all throughout the sector and sub-sector level cells remained unperturbed or were perturbed by less than 2 %. Stronger perturbations were observed only on the lower levels of NACE sectors D and C. While in sector D perturbations

beyond 5 % were very rare, and were observed only on the 4-digit level, larger perturbations (i.e. more than 2 % of the cell value) were observed more frequently in the C-sector. In this sector, 20 to 25 % of the cells below the sub-sector level were perturbed by more than 5 % . This means, on the other hand, that even on the lower NACE levels of the C-sector about 80 to 85 % of the cells were perturbed by less than 5 %, which is quite a positive result for that sector with serious dominance problems where 297 of the 925 Member State cells are flagged confidential.

For the purpose of comparison we have also computed a cell suppression pattern for the sector D table using the τ-ARGUS modular optimization method for secondary cell suppression. In principle – to avoid certain risks of underprotection – the method should be applied to the full table, including the member state level cells. In practice, however, this is not feasible. The fact that the suppression pattern for the member state cells must not be changed leads to infeasibility problems. Therefore the original 2-dimensional (by NACE and member states) cell suppression problem was relaxed and turned into a 1-dimensional problem, addressing only the selection of secondary suppressions on the European level. Primary suppressions on the European level and the corresponding protection levels were identified on the basis of the rules of the SBS confidentiality charter. As a result we got 27 secondary suppressions protecting 25 primary suppressions, e.g. 52 suppressed cells. Obviously, the rounding affected a lot more cells (109). On the other hand, the information loss resulting from rounding a cell is certainly less than from suppressing that cell.

Table 3 below compares the cell suppression result for the D-sector tabulation with the rounding result using three alternative measures of information loss for rounded cells, and two for suppressed cells. For suppressed cells, the first information loss measure is a simple count of the suppressed cells. The second, more sophisticated measure is based on a computation of the feasibility interval for each of the suppressed cells. It considers the size of this interval as measure for the information loss. For the computation of the feasibility intervals we have taken into account as a lower *a priori* bound (i.e. a bound known to data users) for each suppressed EU-level cell the sum over the corresponding published (e.g. non-confidential) member state cells.

For rounded cells, the first measure is a simple count of the number of rounded cells, the second measure is a count of rounded cells where the rounding base exceeds 2 % of the cell value, and the third one considers the size of the rounding interval as information loss for a rounded cell.

NACE-level	Cell Suppression		Rounding		
	# sup-pressed	∑(size of feasibility intervals) (in tsd.)	# rounded	# rounded by more than 2%	2 ∑(size of rounding bases) (in tsd.)
4-digit	19	1920	45	1	581
3-digit	20	1961	45	2	579
2-digit	11	1431	16	1	405
sub-sector	2	18	3	0	12
total	52	5330	109	4	1578

Table 3 Information loss of the cell suppression result for NACE sector D tabulation compared to information loss of rounding result

Table 3 shows that much less cells were rounded by more than 2 % than suppressed (4 vs. 52, over all NACE levels). The impression that rounding outperforms cell suppression in this instance is also confirmed by the more sophisticated evaluation of feasibility interval sizes for suppressed cells (5330 tsd. in total), vs. rounding intervals for rounded cells (1578 tsd. in total).

3.4 Future work

There are some methodological aspects which have not yet been considered closely so far, but will need some attention in the future.

Linked tables: Eurostat also publishes tabulations of variables 1 and 2 by NACE and size class. These 3-dimensional tables have of course cells in common with the 2-dimensional tables we studied so far creating a linked-tables problem. Consequently, the current approach would have to be extended as to guarantee the use of identical rounded approximations for identical aggregates between tables.

One option to solve this problem could be joining linked tables into a single big 'table', and to solve the resulting large optimization problem. This is the only way to guarantee a feasible and good (or optimal) solution. It is not yet clear, however, how 'expensive' (in terms of computer resource requirements) and how efficient this solution would be.

Alternatively one could try an iterative so called 'coordinate descent' approach. In such an attempt we would first compute a solution for the first table, and then, considering these results, compute a solution for the second table. This will have to be repeated until some stage of convergence has been reached.

Related tables and time series: There are pairs of variables, for which Eurostat also intends to publish the ratio of the two. Computation of approximations of these ratios as ratio of the rounded approximations obtained by our procedure is of course straightforward, but has not yet been done for the test data sets. Afterwards, intervals for these ratios (given the rounding intervals of the corresponding enumerator and denominator indicator) will have to be computed and examined. If it turns out that those intervals are too large, i.e. the quality of the approximation for the ratio is too low, it could be considered to develop a more advanced procedure. The objective of the advanced procedure could for instance be increasing the likelihood, that if the rounded approximation of one indicator is smaller than its true value, the rounded approximation of the other indicator will also be smaller than its true value, e.g. that both approximations perturb the true value in the same direction. A similar approach could be attempted in order to improve the behaviour of the rounding method when applied to a time series of tabulations of a variable.

Interval protection methodology: Because of its advantages in terms of efficiency as mentioned above, it might be worth to fully develop the alternative interval protection methodology outlined in section 3.2.

4 Conclusions

We have proposed and tested rounding methodology for disclosure control of European aggregates of the ProdCom statistics and Structural Business Statistics (SBS). The procedure suggested for ProdCom data is based on the τ-ARGUS rounding procedure. Because of its more complex data structures, this procedure cannot be expected to work well in the SBS case. For the SBS data, we have therefore developed a rounding procedure based on restricted controlled tabular adjustment. For the special case of strictly non-negative tables we also outlined an alternative method based on interval protection instead of controlled tabular adjustment.

Both rounding procedures gave promising results. In the Prodcom case the majority of EU-level cells were perturbed by at most 10 % of the EU total. In the SBS case about 95 % of the cells remained unperturbed or were perturbed by at most 2 %. We also provided some evidence that the rounding in this case outperforms cell suppression.

With respect to the SBS data set, some aspects need further attention. Before the method can be used for production, the methodology has to be extended to be applicable to sets of linked tables. Another issue that should be addressed in future research is how to improve the behaviour of the method in a situation where we want to preserve to some extent the correlation between tabulations of different variables, or of tabulations of a time series of a variable. Finally, it might also be interesting to implement the alternative method based on interval protection and compare its behaviour to the current procedure.

Acknowledgements

This research was financed by the European Commission under two specific contracts (No. 22100.2006.002-2006-796 and No. 22100.2006.002-2006-795).

References

Castro, J., Giessing S. (2006). Testing variants of minimum distance controlled tabular adjustment, in *Monographs of Official Statistics*. Work session on Statistical Data Confidentiality, Eurostat-Office for Official Publications of the European Communities, Luxembourg, 2006, pp. 333-343

Hundepool., Anco,. Josep Domingo-Ferrer, Luisa Franconi, Sarah Giessing, Rainer Lenz, Jane Longhurst, Eric Schulte Nordholt, Giovanni Seri, Peter-Paul De Wolf (2006), *CENEX handbook on Statistical Disclosure Control,* CENEX-SDC project, http://neon.vb.cbs.nl/cenex/CENEX-SDC_Handbook.pdf

Hundepool, Anco, et al (2006), *τ-ARGUS manual Version 3.2*, Voorburg. The Netherlands, http://neon.vb.cbs.nl/casc/Software/TauManualV3.2.pdf

Salazar-Gonzales, J.J., Bycroft, C., Staggemeier, A.T. (2006). The controlled rounding implementation, in *Monographs of Official Statistics*. Work session on Statistical Data Confidentiality, Eurostat-Office for Official Publications of the European Communities, Luxembourg, 2006, pp. 303-308

The availability of Dutch census microdata

Eric Schulte Nordholt [1]

[1] Statistics Netherlands, P.O. Box 4000, 2270 JM Voorburg, The Netherlands, ESLE@CBS.NL[*]

Abstract. Data from many different sources were combined to produce the census 2001 tables for the Netherlands. Statistics Netherlands conducted a virtual census, using registers and surveys already available. The virtual census is cheaper, comparable to earlier Dutch censuses and more socially acceptable. In this paper the availability of Dutch census microdata from 1960, 1971 and 2001 is described in more detail. In the last section of this paper some background information is given about the applied Statistical Disclosure Control (SDC) methods. The aim is to release as much census information as possible. However, the privacy of individual respondents should be respected. Therefore, SDC techniques have been developed to protect sensitive information that can be attributed to individual respondents. SDC is thus relevant to be able to decide properly what kind of census tables and microdata can be released.

Keywords. census; micro datasets; micro linking; Statistical Disclosure Control (SDC)

1 Introduction

In 2003 data were combined to produce the Dutch 2001 census tables. In the Netherlands this was not done by interviewing inhabitants in a complete enumeration, but by using data that Statistics Netherlands already had available. This way, the Dutch tax payer got a much lower census bill. The costs for a traditional census would be about three hundred million euros, while the costs made now are 'only' about three million. The estimate includes the costs for all preparatory work such as developing a new methodology and accompanying software. The costs of the registers are not included, but the analyses of the results are. Registers are not kept up-to-date for censuses but for other purposes. Saving money on census costs is only possible in countries that have sufficient register information.

The 2001 census relates to forty extensive frequency tables. Twenty-eight are about the Netherlands as a whole, nine are at the regional level (NUTS 3) and three at municipal level (NUTS 5). The forty tables fall into a number of groups. Eight tables concern housing, two tables concern commuting and the other thirty tables are demographic tables, relating to occupation, level of education and economic activity. Additionally, demographic, housing and labour figures are compiled at sub-city district level for ten large cities that participate in Urban Audit II (Statistics

[*] The views expressed in this paper are those of the author and do not necessarily reflect the policies of Statistics Netherlands.

Netherlands, 2003). These ten large cities are Amsterdam, Rotterdam, The Hague, Utrecht, Eindhoven, Tilburg, Groningen, Enschede, Arnhem and Heerlen.

The virtual census in the Netherlands was off to a later start than in other countries where a traditional census was conducted. It did not make sense to really start the 2001 Census Project until all sources were available; some registers were available relatively late. Nevertheless, the Netherlands was quicker with the compilation of the forty census tables than most of the other countries that participated in the 2001 Census Round. In fact, the Netherlands was one of the first to send the complete set of forty tables to Eurostat, which co-ordinated the contributions of all European Union (EU) member states, accession countries and European Free Trade Association (EFTA) member states. The Netherlands had the advantage that the incoming census forms did not need to be checked and corrected. However, one must realise that for some variables only sample information is available, which implies that it was impossible to meet the level of detail required in some Dutch tables.

The reason why Statistics Netherlands has compiled the set of tables is a gentlemen's agreement. In 1991 the Census Act was rescinded, officially cancelling Statistics Netherlands obligation to hold a census once every ten years (Corbey, 1994). There was no European obligation to supply 2001 Census data, but it is almost inconceivable that the Netherlands would not compile census data for the international organisations just like all other European countries do. Eurostat has a co-ordinating role in collecting harmonised data on the EU and a duty to make international comparisons of the outcome.

It was the third time that the Netherlands conducted a virtual census. However, the Dutch data that have been compiled for 1981 and 1991 were of a much more limited character than the set of tables of the 2001 census. Moreover, they were largely based on a register count of the population in combination with the then existing surveys on the labour force and housing conditions.

In sections 2, 3 and 4 the method of compiling the 2001 Census is explained, the micro linkage aspects are illuminated and information is given about recent publicity about censuses in the Netherlands. In section 5 some more detailed information can be found how microdata of the 1960, 1971 and 2001 Censuses were released. The applied rules for these public use microdata files can be found in section 6. In that section also other methods that allow use of census data are discussed.

2 Method of compiling

The current virtual census relates to 2001. The backbone of the census is the central Population Register (PR), which is the combination of all municipal population registers. The Population Register (PR) contains demographic information on every inhabitant of the Netherlands (Prins, 2000).

A number of integrated surveys and registers were linked to the PR. For this linking process only exact matches were used based on the unique so-called Social security and Fiscal (SoFi) number. The integrated system is called the Social Statistical Database (SSD) system. It is developed originally to conduct virtual censuses, but now it is also used for many social statistics.

The Nordic countries (Denmark, Finland, Iceland, Norway and Sweden) have more variables available in registers than the Netherlands (United Nations, 2007). Moreover, some Nordic countries conduct a (limited) enumeration for variables missing in the registers. Most of the other countries are in a similar position as the Netherlands where some variables relevant for the census can be found in registers, while other variables are available on a sample basis only. That's why much interest exists in the Dutch approach to combine registers and surveys and to use modern statistical techniques and accompanying software to compile the tables.

To be able to estimate every table as accurately as possible, each estimate is based on the largest possible number of records. Tables that contain register variables only are counted from the registers. Tables that contain at least one variable from a survey are estimated from the largest possible combination of registers and surveys.

We guaranteed consistency among the tables by using the technique of repeated weighting. It generates a new set of weights for each estimated table and is based on the repeated application of the regression estimator. When using repeated weighting, the weights of the records in the microdata are adapted in such a way that a new table estimate is consistent with all earlier table estimates.

It is possible to use the technique of repeated weighting in other countries as well. However, first it should be possible to use registers for statistical purposes. In most countries, not all census variables can be derived from register information. Additional surveying then remains a necessity, but a consistent set of census tables can be produced using the technique of repeated weighting.

3 Micro linkage

Most of the present administrative registers are provided with a unique linkage key. This SoFi-number is a personal identifier for every (registered) Dutch inhabitant and those abroad who receive an income from the Netherlands and have to pay tax over it to the Dutch fiscal authorities.

To prevent misuse of the SoFi-number, Statistics Netherlands recodes it for statistical processing into a so-called Record Identification Number (RIN-person). Personal identifiers, such as date of birth and address, are replaced by age at the reference date and RIN-address. This is all done in accordance with regulations of the Dutch Data Protection Authority to protect the privacy of the citizens.

All social statistics data files can be linked exactly to the PR. In practice this means that these data files are all indirectly linked to each other via the PR. Therefore the PR can be considered the backbone in the set of social data sources. When linking the PR and the jobs register, or the PR and a register of social benefits, it is a linkage between different statistical units (persons, jobs, benefits). In that case multiple linkage relationships can exist because someone can have more than one job or can benefit from several social benefits.

In household sample surveys, like the Labour Force Survey (LFS), records do not have a SoFi-number. For those surveys an alternative linkage key is used, which is often built up by a combination of the following personal identifiers:

- sex;
- date of birth;
- address[†].

This sort of linkage key will usually be successful in distinguishing people. However, it is not a 100 percent unique combination of identifiers. Linking may result in a mismatch in the case of twins of the same sex. False matches may also occur when part of the date of birth or the postal code and house number is unknown or wrong. Another drawback is that the linkage key is not person but address related, which may cause linkage problems if someone has recently moved. When linking the PR and the LFS with this alternative key, and tolerating a variation between sources in a maximum of one of the variables sex, year of birth, month of birth or day of birth, the result is that close to 100 percent of the LFS records will be linked.

In its linkage strategy, Statistics Netherlands tries to maximize the number of matches and to minimize the number of mismatches. So, in order to achieve a higher linkage rate, more efforts are made to link the remaining unlinked records by means of different variants of the linkage key. For example, leaving out the house number and tolerating variations in the numeric characters of the postal code. To keep the probability of a mismatch as small as possible, some 'safety' devices are built in the linkage process. This last linking attempt accomplishes an extra one percent matches.

4 Publicity about censuses in the Netherlands

At the end of 2003 the complete set of forty census tables for the Netherlands was sent to Eurostat. The book ‘The Dutch Virtual Census of 2001, Analysis and Methodology’ was written afterwards (Schulte Nordholt et al., 2004). This book provides a wide-ranging description of the socio-demographic and socio-economic

[†] In fact, the combination of a postal code (mostly related to the street) and house number is used as substitute for the address. The postal code in the Netherlands consists of four figures, followed by two letters.

state of the Netherlands based on the 2001 census results. It discusses differences in size and composition among households, economic activity of households, individual activity status by region, age, education level and branch of economic activity. There are separate chapters on the economic activities of young people and people of retirement age. The economic activities, levels of education and occupation of foreigners from various countries of origin are compared with each other and with the native Dutch population. Regional aspects are also examined, including commuting. The results of the 2001 census are compared with the census results of some other European countries and with earlier Dutch censuses. Lastly, the virtual census methodology used is described in some detail.

The PDF version of the book can be found at the Statistics Netherlands website, at page http://www.cbs.nl/en-GB/menu/themas/dossiers/volkstellingen/publicaties/default.htm. An extra Chapter (number 15) is available at page http://www.cbs.nl/en-GB/menu/themas/dossiers/volkstellingen/methoden/default.htm with an overview of the used data sources, methods and definitions. Hard copies of the book were sent to all authors of the book, to the management of Statistics Netherlands and to several libraries. The book was also offered to the Prime Minister, the Minister of Economic Affairs and the Minister of Education, Cultural Affairs and Science of the Netherlands and to Director-Generals of statistical offices in several countries. In August 2004, the book was publicly released at an official presentation in the Statistics Netherlands' office in Voorburg. The research process and the main findings were then presented to an audience of academics, press representatives, government officials, as well as Statistics Netherlands' employees. Several articles were written in national and regional newspapers about the Dutch virtual census of 2001 and its results. Announcements, book reviews of Schulte Nordholt et al. (2004) and interviews appeared in several journals, mailing lists and newsletters. The methodology and key results of the virtual census of 2001 were also published as Schulte Nordholt (2005).

The set of forty standard tables for the Netherlands (in Excel format) can be found at page http://www.cbs.nl/en-GB/menu/themas/dossiers/volkstellingen/publicaties/artikelen/archief/2005/default.htm

5 Producing microdata

5.1 Introduction

Protected 1 percent samples of the microdata of the Dutch censuses of 1960, 1971 and 2001 were in 2005 disseminated via the IPUMS (Integrated Public Use Microdata Series) project, see http://www.ipums.org/international. These micro

datasets contain a number of demographic and economic variables and can also be analysed via the institute DANS (Data Archiving and Networked Services), see http://www.dans.knaw.nl/en/. Bona fide researchers who want to make more detailed studies on these three censuses can work on-site at the premises of Statistics Netherlands. More information about this last option can be obtained via Statistics Netherlands' Centre for Policy Related Statistics (http://www.cbs.nl/nl-NL/menu/informatie/beleid/centrum-voor-beleidsstatistiek/diensten/default.htm).

The Dutch censuses of 1960, 1971 and 2001 have been selected to be part of the IPUMS project. The censuses of 1960 and 1971 are traditional censuses, of which most of the micro data records have been recovered. The 2001 census is a virtual census, which means that it is composed of available register data and existing surveys. Unfortunately, this results in not having all variables available for all individual records. As a consequence we have not released the complete set of micro data, but an anonymised balanced sample of the individual personal records for which we have all demographic and economic variables. The sample fraction is a little bit over 1 percent of the total population.

The first stage in our cooperation in the IPUMS project has been the release of the 2001 census micro data. The selection of variables of the 2001 census has been leading in the selection of the variables of the censuses of 1960 and 1971. Due to differences in variable definitions, classifications and variable availability over time, differences among the three micro data sets remain. Also for 1960 and 1971 we release anonymised balanced 1 percent samples of the total population.

More information about the variables selected of the three censuses can be found in the next subsections. Links to more information about the 2001 census can be found in the previous section. For some more background and documentation of the 1960 and 1971 censuses we refer to the following web site: http://www.volkstellingen.nl/en/documentatie/.

5.2 The variable selection of the 2001 census

5.2.1 The sample

The sample is composed in three stages.

Of the persons 0-14 years a 1 percent sample, stratified to age (in years) and sex is drawn from the combined register data.

Out of the records from persons 15-74 years all complete records are selected. (Remember that we only had existing surveys and register data at our disposal). These records sum up to about 1 percent of the population in this age group. Record weights have been provided.

Of the records of persons 75 years or older all complete records are selected. Record weights have been provided.

Persons in institutional households are not included in the sample because they are not included in the surveys used for the virtual census of 2001.

5.2.2 The variables and their categories

Because of disclosure reasons we have to limit the detail. This results in the impossibility of providing regional detail. Moreover, some rare combinations of identifying variables will lead to a limited number of suppressions of variable scores. The list of variables of the 2001 census sample is as follows.

1. Sex
2. Age
3. Household position
4. Household size
5. Place of residence one year prior to the census
6. Country of citizenship
7. Country of birth
8. Level of educational attainment (ISCED level 4)
9. Economic status
10. Occupation (ISCO-COM 1 digit)
11. Branch of current economic activity (NACE, 1 letter)
12. Marital status

5.3 The variable selection of the 1971 census

5.3.1 The sample

For the census year 1971 the gross sample of 1.25 % of the total population is randomly drawn, stratified to sex, 17 age groups (16 5-year groups and 80+) and 12 regions (11 provinces and 1 region consisting of newly made land (polders) and the centrally registered population). After removing incomplete and other problematic records a net sample of over 1 percent remained. The records have been weighted to the published census combined totals of sex times age in years, 11 provinces (the newly made land and the centrally registered travelling population are added to one of the provinces) times sex times age in 5-year classes and simple totals of most of the published variables.

The total population includes persons in institutional households, this in contrast to the IPUMS dataset of the virtual census of 2001.

5.3.2 The variables and their categories

Because of disclosure reasons we have to limit the detail. This results in the impossibility of providing regional detail. Moreover, some rare combinations of identifying variables will lead to a limited number of suppressions of variable scores. The list of variables of the 1971 census sample is as follows.

1. Sex
2. Age
3. Country of citizenship
4. Marital status
5. Household position
6. Religious denomination
7. *Country of birth*
8. Household size
9. Economic status
10. Level of educational attainment
11. Occupation (ISCO 1 digit)
12. Branch of current economic activity (SBI 1970)

5.4 The variable selection of the 1960 census

5.4.1 The sample

For the census year 1960 the gross sample of 1.25 percent of the total population has been drawn randomly, stratified to sex, 17 age groups (16 5-year groups and 80+) and 12 regions (11 provinces and 1 region consisting of newly made land (polders) and the centrally registered population). After removing incomplete and other problematic records a net sample of over 1 percent remained. The records have been weighted to the published census combined totals of sex times age in years and region times sex times age in 5-year classes.

The total population includes persons in institutional households, this in contrast to the virtual census of 2001.

The source material is the over 11 million original punch cards, which have been reread and digitised from 1973 onwards. A report in English on the reconstruction of the dataset is available.

5.4.2 The variables and their categories

Because of disclosure reasons we have to limit the detail. This results in the impossibility of providing regional detail. Moreover, some rare combinations of identifying variables will lead to a limited number of suppressions of variable scores. The list of variables of the 1960 census sample is as follows.

1. Sex
2. Age
3. Marital status
4. Household position
5. Religious denomination
6. Country of birth
7. Economic status
8. Level of educational attainment
9. Occupation
10. Branch of current economic activity (SITC)

6 Applied Statistical Disclosure Control methods

6.1 Introduction

The information from statistics becomes available for the public in tabular and microdata form. Historically, only tabular data were available and National Statistical Institutes (NSIs) had a monopoly on the microdata. Since the eighties the PC revolution leaded to the end of this monopoly. Now also other users of statistics have the possibility of using microdata. These microdata can be conveyed with floppies, CD-ROMs, USB sticks and other means. Recently also other possibilities of getting statistical information have become more popular as remote access and remote execution. With these techniques researchers can get access to data that remain in a statistical office or can execute set-ups without having the data on their own PC. For very sensitive information some NSIs have the possibility to let bona fide researchers work on-site within the premises of the NSI.

The task of statistical offices is to produce and publish statistical information about society. The data collected are ultimately released in a suitable form to policy makers, researchers and the general public for statistical purposes. The release of such information may have the undesirable effect that information on individual entities instead of on sufficiently large groups of individuals is disclosed. The question then arises how the information available can be modified in such a way that the data released can be considered statistically useful and do not jeopardize the privacy of the entities concerned. The Statistical Disclosure Control theory is used to solve the problem of how to publish and release as much detail in these data as possible without disclosing individual information (Willenborg and De Waal, 1996 and 2001).

6.2 The release of public use census microdata files

Many users of statistics are satisfied with the safe tables released by statistical offices. However, some users require more information. For many surveys microdata for researchers are released. In the case of census data it was preferred to release public use microdata files. Public use microdata files contain less detailed information than microdata for research. However, the audience for these files is much larger. The software package μ-ARGUS (Hundepool et al, 2007a) is of help in producing all kinds of protected microdata files. For the public use microdata files Statistics Netherlands uses the following set of rules:

1. The microdata must be at least one year old before they may be released.
2. Direct identifiers should not be released. Also direct regional variables, nationality, country of birth and ethnicity should not be released.
3. Only one kind of indirect regional variables (e.g. the size class of the place of residence) may be released. The combinations of values of the indirect regional variables should be sufficiently scattered, i.e. each area that can be distinguished should contain at least 200 000 persons in the target population and, moreover, should consist of municipalities from at least six of the twelve current provinces in the Netherlands. The number of inhabitants of a municipality in an area that can be distinguished should be less than 50 % of the total number of inhabitants in that area.
4. The number of identifying variables in the microdata is at most 15.
5. Sensitive variables should not be released.
6. It should be impossible to derive additional identifying information from the sampling weights.
7. At least 200 000 persons in the population should score on each value of an identifying variable.

8. At least 1 000 persons in the population should score on each value of the crossing of two identifying variables.
9. For each household from which more than one person participated in the survey we demand that the total number of households that correspond to any particular combination of values of household variables is at least five in the microdata.
10. The records of the microdata should be released in random order.

According to this set of rules the public use microdata files are protected more severely than the microdata for research. For public use files it is not allowed to release direct regional variables. This is not considered to be a big problem for the census microdata files as the aim is to make international comparisons.

The software package μ-ARGUS is of help to identify and protect the unsafe combinations in the desired microdata file. Thus the rules 7 and 8 can be checked with μ-ARGUS. Global recoding and local suppression are two data protection techniques used to produce safe microdata files. In the case of global recoding several categories of an identifying variable are collapsed into a single one. This technique is applied to the entire data set, not only to the unsafe part of the set, so that a uniform categorisation of each identifying variable is obtained.

Both global recoding and local suppression lead to information loss, because either less detailed information is provided or some information is not given at all. A balance between global recoding and local suppression should always be found in order to make the information loss due to the Statistical Disclosure Control measures as low as possible. It is recommended to start by recoding some variables globally until the number of unsafe combinations that has to be protected is sufficiently low. Then the remaining unsafe combinations have to be protected by local suppressions.

6.3 Other methods that allow use of census data

All Statistical Disclosure Control techniques necessarily involve data manipulation or suppression and are likely to reduce the quality of estimates to be produced from the data. As a result, NSIs have begun to investigate other methods that allow use of data while protecting confidentiality of sensitive information given by respondents. These methods allow the data to be used in an environment controlled by the NSI and require that its use be subject to the same legal and ethical protections placed on the NSI itself.

Probably the most important access modality developed in the past decade is that of restricted access sites. These sites permit NSIs to respond to the microdata needs of researchers. Some researchers need namely more information than is available in the released public use census microdata files. As the releasing of richer data is not allowed, it is then possible for individual researchers to perform their research on

richer microdata on the premises of the NSIs. Statistics Netherlands is one of the NSIs that has such a facility. Bona fide researchers have the opportunity to work on-site in a secure area within Statistics Netherlands. Researchers can choose at will between the two locations of Statistics Netherlands: Voorburg in the west of the Netherlands and Heerlen in the south of the Netherlands. The possibility to export any information is however only possible with the permission of the responsible statistical officer. They can apply standard statistical software packages and also bring their own programmes. Like all employees of Statistics Netherlands, these people who work on-site have to swear an oath to the effect that they will not disclose the individual information of respondents (Kooiman, Nobel and Willenborg, 1999).

Finally, an option is to allow remote access. This access modality combines the advantage that researchers can stay in their own institute and the advantage that the data stay in the NSI. Normally, researchers get access through an intermediary controlled by the NSI that guarantees that all use conforms to the law. One step further goes the option of remote execution. An intermediary is then no longer placed between the researcher and the NSI. With remote execution researchers can execute set-ups without having the data on their own PC. Although remote execution is a more efficient option than remote access the question is whether the security systems are strong enough to let this technique become an often used modality. Currently, Statistics Netherlands has a Centre for Policy Related Statistics that is running the on-site and remote execution and access facilities for researchers.

6.4 Discussion and conclusions

In this paper methods have been described that have been developed to protect confidentiality, while at the same time providing access to data, through various means that either alter the data or restrict access to them. The balance between data confidentiality and data access is a delicate one. Hopefully, the new research methods and software for Statistical Disclosure Control can help in keeping the right balance for the Dutch census microdata.

The software packages μ-ARGUS and τ-ARGUS have emerged from the Statistical Disclosure Control (SDC) project that was carried out under the Fourth Framework Programme of the European Union. The Computational Aspects of Statistical Confidentiality (CASC) project in the Fifth Framework Programme of the European Union can be seen as a follow-up of the SDC project.

New manuals for μ-ARGUS and τ-ARGUS become regularly available (Hundepool et al, 2007a and b). The ARGUS packages have moved towards interfaces with several state of the art engines produced by statisticians from many different countries. The most recent information is published at the CASC website: http://neon.vb.cbs.nl/casc.

References

Corbey, P., 1994, "Exit the population Census", *Netherlands Official Statistics*, 9, summer 1994, pp. 41-44.

Hundepool, A., A. van de Wetering, R. Ramaswamy, L. Franconi, S. Polettini, A. Capobianchi, P.P. de Wolf, J. Domingo, V. Torra, R. Brand and S. Giessing, 2007a, *μ-ARGUS, user's manual, version 4.1*, Voorburg, The Netherlands: Statistics Netherlands.

Hundepool, A., A. van de Wetering, R. Ramaswamy, P.P. de Wolf, S. Giessing, M. Fischetti, J.J. Salazar, J. Castro and P. Lowthian, 2007b, *τ-ARGUS, user's manual, version 3.2*, Voorburg, The Netherlands: Statistics Netherlands.

Kooiman, P., J.R. Nobel and L.C.R.J. Willenborg, 1999, "Statistical data protection at Statistics Netherlands", *Netherlands Official Statistics*, 14, pp. 21-25.

Prins, C.J.M., 2000, "Dutch population statistics based on population register data", *Monthly Bulletin of Population Statistics*. Vol. 2000/02 (February 2000), pp. 9-15.

Schulte Nordholt, E., 2005, "The Dutch virtual Census 2001: A new approach by combining different sources", *Statistical Journal of the United Nations Economic Commission for Europe*, 22, 2005, pp. 25-37.

Schulte Nordholt, E., M. Hartgers and R. Gircour (Eds.), 2004, *The Dutch Virtual Census of 2001, Analysis and Methodology*, Statistics Netherlands, Voorburg / Heerlen, July, 2004. http://www.cbs.nl/NR/rdonlyres/D1716A60-0D13-4281-BED6-3607514888AD/0/b572001.pdf

Statistics Netherlands, 2003, "Urban Audit II, the implementation in the Netherlands", *Report, BPA no. 2192-03-SAV/II*, Statistics Netherlands, Voorburg. http://www.cbs.nl/nr/rdonlyres/8c6e4c9d-4338-4e32-848b-8d43b9b3242d/0/urbanauditiinetherlands.pdf

United Nations, 2007, "Register-based statistics in the Nordic countries – Review of best practices with focus on population and social statistics". *Report,* United Nations Economic Commission for Europe (UNECE) Statistical Division, Geneva.

Willenborg, L.C.R.J. and T. de Waal, 1996, *Statistical Disclosure Control in practice, Lecture Notes in Statistics 111*, New York: Springer-Verlag.

Willenborg, L.C.R.J. and T. de Waal, 2001, *Elements of Statistical Disclosure Control, Lecture Notes in Statistics 155*, New York: Springer-Verlag.

The Application of the Concept of Uniqueness for Creating Public Use Microdata Files

Jay J. Kim [1] and Dong M. Jeong[2]

[1] National Center for Health Statistics, Hyattsville, MD 20872, USA
[2] Korea National Statistical Office, Daejeon, Republic of Korea

Abstract. In general, agencies use non-systematic ad hoc approaches for protecting against disclosure in microdata. This paper develops probability models quantifying disclosure risk for a microdata file. This is a modification of the Marsh, et al (1991) procedure. The model can use population and sample uniques only, or it can also include population and sample twins or triplets. For identifying population uniques, twins or triplets, we need to determine what type of information intruders have which is also available on the microdata file. This common information is called "key variables." Using the models, disclosure risk can be computed for the original microdata, and we can determine whether the risk is too high or not. If the risk is too high, grouping categories, post randomization methods or randomized response techniques, etc., can be used for masking the categorical variables. If the variable is continuous, grouping or noise inoculation, etc., can be used for masking the variable. The probability model using population and sample uniques only was applied for creating disclosure-limited microdata files using the 2005 Korean demographic census data (8). In this paper, an attempt is made to develop a theoretically defensible and systematic approach for protecting against disclosure in microdata.

Keywords. disclosure risk, population and sample uniques, key variables, grouping

1 Introduction

Government agencies release microdata files from their survey data or administrative records data. Large amounts of information on individuals is available to many organizations and data users. If a public use microdata file (PUMF) is released, intruders could try to match their records with the ones from the PUMF and gain access to new information. Note that in linking the records on two files, common information is used, which is called "key variables." Data disseminating agencies must protect the confidentiality of individuals on their files. In the U.S., laws such as Title 13 stipulate protection of the confidentiality of many types of data such as survey data and income tax return data. On the other hand, agencies should not ignore the data users' need, i.e., the utility of the data files. Therefore, in creating a

[3] The findings and conclusions in this paper are those of the author and do not necessarily represent the views of the National Center for Health Statistics, Centers for Disease Control and Prevention.

PUMF, agencies attempt to eliminate disclosure risk of the file while maintaining maximum utility of the data.

PUMF does not carry names and addresses of individuals on the file. Thus intruders have to rely on other variables. However, even if records on two files match exactly, there is no guarantee that they represent the same individual(s). If they are population and sample uniques, no errors are in the data (key variables), and the data are collected almost at the same time (the reference periods are similar), the match would be correct, but if an individual is not a population unique, there is still a chance of identification, but the probability would be lower.

Marsh, et al (1991) developed a probability model for measuring disclosure risk for creating a PUMF from the United Kingdom's census data. Their model depends on population uniques. In this paper we will show modified versions of the model and apply them for creating a PUMF using the 2005 Korean census data.

2 Intruders and Disclosure

Potential intruders may attempt to match the records on the PUMF to their databases which contain identifiers and glean new information from the PUMF. Examples of potential intruders are credit card companies, mortgage departments of banks, insurance companies, credit bureaus, trade associations, central government agencies such as Internal Revenue Service and local government agencies which have access to organizational data base. However, intruders do not need to be organizations holding data sets. There is a lot of information in the public domain and because of readily available high computing power, individual intruders can assemble data files themselves.

There are two types of disclosure. The first is an identity disclosure. Lambert (1993) calls this identification. If the intruder is a journalist and tries to embarrass the data disseminating agencies, his claim that he has been successful in identifying someone on their PUMF would be sufficient. If the intruder publicizes the findings in the news media, it could have a devastating effect on the agencies' data collection efforts. The other is gaining new information after identification (2, 9). Lambert calls this an attribute disclosure. Identity disclosure is a prerequisite for disclosure of additional information. Variables other than key variables are available in the PUMF, and hence once the identity is disclosed, there is no doubt that new information can be disclosed. Thus for defining a measure of disclosure risk, identity disclosure and attribute disclosure will be considered equivalent in this paper.

3 Measures of Disclosure Risk

Let

P(a) = the probability of key variables being recorded identically in both PUMF and intruder's file;

P(b|a) = the probability that an individual appears in a PUMF is the same as the sampling fraction for that individual in the PUMF;

P(c|a,b) = the probability of population unique;

and

P(d|a,b,c) = the probability of verifying population uniqueness.

Marsh, et al (1991) defined the probability of correct identification of an individual as

$$P(\text{correct identification of an individual}) = P(a)P(b \mid a)P(c \mid a,b)P(d \mid a,b,c) \quad (1)$$

We modify the above model in the following manner.

Disclosure can occur whether a person in a population is unique or not based on the values of key variables. That is, even if there are two or three persons in a population who are the same on key variables, one of their identities could be disclosed, if additional information is accessed, because more information than that on key variables is available on the PUMF and intruder's file. However, most researchers have been paying attention to the population unique cases only in defining the disclosure risk. Thus, the disclosure risk can be defined narrowly or broadly depending on whether or not it is restricted to the unique cases in the population. Here, we will develop formulas for both narrow and broad definitions of the disclosure risk. In deriving the formulas, we assume that there are no recording or classification errors for the values of the key variables. In other words, *P(a)* = 1 in Marsh, et al's formula. It is also assumed that *P(d|a,b,c)* = 1. Our model will be restricted to the sample unique.

Disclosure can occur when the following conditions are met:

1. An individual is unique in a population based on key variables.
   If the intruder's file is a 100 percent population file, he can establish uniqueness of a certain individual by using his file.
2. The individual is on a PUMF from a survey or an administrative records file.
3. The individual is on another file which an intruder is working with.
   An intruder can have information on key variables for a specific person and try to examine whether that person appears in the above microdata. In this case, the second file has a single record.
4. The individual on the files in conditions 2 and 3 above is unique.

In general, agencies use non-systematic ad hoc approaches for protecting against disclosure in microdata. It is well known that the disclosure risk is affected by the inclusion probability in the file(s) (3). In this paper we develop comprehensive probability models for the disclosure risk which incorporate the inclusion probability in the PUMF and intruder's file.

Let

A = an individual of interest;

F_1 = PUMF in condition 2 above

F_2 = a file held by an intruder (in condition 3 above);

P_1 = unique class in the population;

S_{1F_1} = unique class in F_1;

and

S_{1F_2} = unique class in F_2.

3.1 A Narrow Definition of Disclosure Risk

The narrow definition of disclosure risk is based on the population and sample uniques only.

3.1.1 Assuming an Intruder Does a Phishing (Fishing) Expedition

Here we assume that the intruder does not know anyone on PUMF, but tries to link the records on his file to those on the PUMF. When all four conditions mentioned above and the assumption of no measurement error are met, then the probability of correct identification is

$$P\left[\left(A \in F_1\right) \cap \left(A \in F_2\right) \cap \left(A \in S_{1F_1}\right) \cap \left(A \in S_{1F_2}\right) \cap \left(A \in P_1\right)\right] \tag{2}$$

If an individual is a population unique, it would also be a sample unique. In other words,

$$\begin{aligned} &P\left[\left(A \in S_{1F_1}\right) \cap \left(A \in S_{1F_2}\right) \cap \left(A \in P_1\right)\right] \\ &\quad = P\left[\left(A \in S_{1F_1}\right) \cap \left(A \in S_{1F_2}\right) \mid \left(A \in P_1\right)\right] P\left(A \in P_1\right) \\ &\quad = P\left(A \in P_1\right) \end{aligned}$$

Because of the above, the disclosure risk for individual A can be reduced to the following.

$$P\left[\left(A \in F_1\right) \cap \left(A \in F_2\right) \cap \left(A \in P_1\right)\right] \tag{3}$$

which can be further re-expressed as follows:

$$P\left[\left(A \in F_1\right) \cap \left(A \in F_2\right) \mid \left(A \in P_1\right)\right] P\left(A \in P_1\right) \tag{4}$$

Note that the event that A is unique in population is independent of whether A is selected in sample or not. Thus, equation (4) reduces to

$$P\left[\left(A \in F_1\right) \cap \left(A \in F_2\right)\right] P\left(A \in P_1\right) \tag{5}$$

The event that A is in the PUMF is usually independent of the event that A is in the intruder's file. For example, the event that an individual is selected in a labor survey is independent of the event that it is on the income tax return file. In this case, equation (5) can be simplified as

$$P\left(A \in F_1\right) P\left(A \in F_2\right) P\left(A \in P_1\right) \tag{6}$$

However, a survey can be a subset of another survey. Thus if F_2 is a larger survey and F_1 is a subset of F_2, then

$$P\left[\left(A \in F_1\right) \cap \left(A \in F_2\right)\right] = P\left(A \in F_2\right) P\left[\left(A \in F_1\right) | \left(A \in F_2\right)\right]$$

Note that $P\left[\left(A \in F_1\right) | \left(A \in F_2\right)\right]$ is a subsampling rate of F_1 from F_2. Thus if there is dependence between the two files, equation (6) becomes

$$P\left(A \in F_2\right) P\left(A \in P_1\right) \square \text{Subsampling Rate} \tag{7}$$

If F_2 is a 100 percent file and $P\left(A \in P_1\right)$ is calculated using F_2, S_{1F_2} in equation (2) should be ignored. S_{1F_2} is needed in equation (2) when F_2 is not a 100 percent census file.

3.1.2 Assuming an Intruder Already Knows That A is in F_1

If an intruder already knows that A is on the PUMF, that is, if the intruder has response knowledge (1), then $P\left(A \in F_1\right) = 1$. Thus, from equation (6), the disclosure risk will be

$$P\left(A \in F_2\right) P\left(A \in P_1\right) \tag{8}$$

Note that if the intruder's file is a 100 percent population file, then $P\left(A \in F_2\right) = 1$. The disclosure risk is $P\left(A \in P_1\right)$ (9)

3.2 Broader Definition of Disclosure Risk

Even if an individual is not unique in the population, he/she still can be identified with additional information included in the PUMF and intruder's file. Thus, the definition of disclosure risk can be broadened. Suppose C individuals in the population have the same values of the key variables and matching to any one of them is equally likely. Define

$P_C =$ Equivalence class of size C in the population.

Then, the probability of correct identification is

$$\frac{1}{C} P\left[\left(A \in F_1\right) \cap \left(A \in F_2\right) \cap \left(A \in S_{1F_1}\right) \cap \left(A \in S_{1F_2}\right) \cap \left(A \in P_C\right)\right] \tag{10}$$

In the above, since a sample unique individual can be matched to either one of the C individuals in population, a multiplier of 1/C is needed.

4 Evaluation of Disclosure Risk

In Korea, there exists a full national population register and every member of the nation excluding foreigners is assigned a resident registration number similar to Social Security Numbers in the U.S. On the register, a person's name, gender, birth year, month and date, place of birth and the place of registration are available. The Korean government does not make this information available to the public. However, commercial enterprises make this information readily available on the web. Because of this situation, one of the missions of the Korean National Statistical Office (KNSO) is to make their PUMF disclosure protected.

For this study, we selected the 2005 census data from a Korean province, Choongchung (CC) Province. The population size, and the number of households and housing units are shown in Table 1. Note that a family or household can be disclosed first, then its members can be further identified. In this paper, however, we will direct our attention to the direct disclosure. The 2 percent microdata file is created by taking a 20 percent subsample of the 10 percent census sample. Thus, the 2 percent microdata corresponds to F_1 and the census sample to F_2. The probability of a population unique is calculated using the 100 percent census.

Table 1. Population Size, and Number of Households and Housing Units - CC Province

	Population	Households	Housing Units
Census	1,798,397	660,526	586,757
Census Sample (10%)	189,505	71,091	65,398
2% Microdata	38,027	14,218	13,038

The following are matching keys we used: gender (2); age (111); marital status (4); relationship to householder (14); household type (5); tenure (6); building type of residence (12); and type of housing and number of floors of the building (12). The number in the parentheses is the number of categories for each of the variable. All the key variables are discrete.

Using the census file (N = 1,798,397) with all eight variables, 9,664 persons were unique, which is 0.54 percent of total population. This is much higher than Bethlehem, et al's threshold of 0.1 percent. If we assume that the intruder has a 10 percent census sample file, the disclosure risk is $0.1 \times 0.2 \times 0.0054 \approx 0.00011$, according to equation (7). However, whole blocks are selected in the 10 percent census sample, thus residents in the sample blocks know that their neighbors are also in the sample, i.e., many sample persons have response knowledge. To those who have response knowledge, the disclosure risk is $0.2 \times 0.0055 \approx 0.0011$, according to equation (8). If we assume that the intruder has the 100 percent population file, the

risk is 0.54 percent. The risks are too high, thus we decided to coarse categories of selected variables. This approach is often used to reduce the risk (1, 5, 9).

Table 2 below shows how the grouping affects the probability of uniques by using up to 4 variables. In the table, x in the columns 2 – 5 indicates which variable(s) is (are) used for identifying uniques.

Table 2. Number of Unique Persons before Grouping Categories –Population Data

# of Vars	Gender	Age	Relationship	Marital Status	# of Uniques
1	x				0
1		x			2
1			x		0
1				x	0
2	x	x			5
2	x		x		0
2	x			x	0
2		x	x		65
2		x		x	11
2			x	x	0
3	x	x	x		167
3	x	x		x	30
3	x		x	x	2
3		x	x	x	349
4	x	x	x	x	713

Table 2 above shows that as we use more variables for identification, the number of uniques increases. Note that the number of uniques with more variables also includes the number of uniques with fewer variables. Note also that the variable age is in single years. When only one key variable is used, age alone provides 2 unique ones. When gender and age are used, the total number of uniques increases to 5. These 5 uniques includes the 2 uniques due to age alone. That is, the unique counts are cumulative. When age and relationship are used, the number of uniques becomes 65. When age, relationship and marital status are used, the total number of unique persons is 349. When marital status is included along with gender, age and relationship, the number of uniques increases to 713.

Table 3 below shows the number of uniques before and after age groupings. Before the grouping, there were 111 age categories, but after the grouping, the number of age categories was reduced to 20.

Table 3. Number of Uniques with 5 Year Intervals for Age – Population Data

# of Vars	Gender	Grouped Age	Relationship	Marital Status	# of Uniques

1		x			2 → 0
2	x	x			5 → 2
2		x	x		65 → 6
2		x		x	11 → 1
3	x	x	x		167 → 18
3	x	x		x	30 → 3
3		x	x	x	349 → 53
4	x	x	x	x	713 → 106

In Table 3 above, the last column shows how much the number of uniques in Table 2 gets reduced due to age groupings. The first number in the column is the number of uniques before grouping and the latter is the number after grouping. Note that by grouping age alone into 5 year intervals, the number of uniques was lowered, sometimes by the ratio of 11:1. If 10 years of age are grouped, with 4 variables, the number of uniques becomes 51, less than half of the 106 uniques when age was grouped into 5 year intervals.

Comparing Table 4 below with Table 3, in Table 4, the variable relationship is also grouped.

Table 4. Number of Uniques with Grouped Age and Relationship Categories – Population Data

# of Vars	Gender	Grouped Age	G. Relationship	Marital Status	# of Uniques
2	x	x			2 → 2
2		x	x		6 → 2
2		x		x	1 → 1
3	x	x	x		18 → 4
3	x	x		x	3 → 3
3		x	x	x	53 → 3
4	x	x	x	x	106 → 8

Table 4 above shows that by using grouped relationship, the number of uniques gets further reduced most of the time.

In comparison to Table 4 above, Table 5 below has grouped marital status.

Table 5. Number of Uniques with Grouped Age, Relationship and Marital Status Categories – Population Data

# of Vars	Gender	G. Age	G. Relationship	G. Marital Status	# of Uniques
2	x	x			2 → 2
2		x	x		2 → 2
3	x	x	x		4 → 4

3	x	x		x	3 → 1
3		x	x	x	3 → 1
4	x	x	x	x	8 → 4

Table 5 shows that grouping marital status sometimes reduces the number of uniques, but not as much as grouping by age or relationship.

We tried two different groupings in terms of the number of categories for: (i) relationship, (ii) building type, and (iii) type of housing and the number of floors of the building. The first grouping has 9, 6, and 6 categories in order (as previously mentioned) and the other grouping has 3, 4 and 4 categories. The first grouping provides 501 uniques and the second results in 495 uniques. The difference is minor. Both represent around 0.028 percent of the total population. This is much lower than Bethlehem, et al's threshold of 0.1 percent. If we assume the intruder has the 10 percent census sample file, the disclosure risk is 0.0000056. This translates into one person per 100,000 people. This is a very low risk. Some agencies provide geographical information if an area has at least 100,000 people. If we assume response knowledge, the disclosure risk goes up to 0.000028. However, if we assume that the intruder has the 100 percent population data, then the risk further goes up to 0.028 percent. This means 28 persons per 100,000 people. In addition to grouping categories of key variables, other variables such as occupation, lot size, and the size of living space need to be investigated.

5 Concluding Remarks

Comprehensive probability models quantifying disclosure risk have been developed for microdata files. The measure of disclosure risk can depend on the number of population uniques, population twins or triplets. We developed two probability models, one for the population unique case and the other for population twins or triplets, etc. The models assume that only unique persons in the sample files face disclosure risk.

We measured the probability of the population uniques using the original census data of KNSO which was 0.54 percent. Assuming that an intruder performs a fishing expedition using the sample data from the census, disclosure risk will be 0.011 percent. If the intruder has response knowledge, the disclosure risk goes up to 0.11 percent. If the intruder has the 100 percent census data, the risk goes up further to 0.54 percent. In this case, the threshold of Bethlehem, et al (1990) is exceeded.

To lower the probability of identifying uniques, we grouped categories by 5 year age intervals. Two different numbers of categories were used for three variables (relationship, building type, and type of housing and the number of floors of the

building), one grouping has fewer categories in all three than the other grouping. However, both provided similar probability of identifying uniques, 0.028 percent, which is much lower than Bethlehem, et al's threshold. In this case, if we assume the intruder has the 10 percent census sample file, the disclosure risk is 0.0000056. This means one person per 100,000 people. Even if we assume response knowledge, the disclosure risk is 0.000028. Thus, the above grouping seems to provide sufficient disclosure protection.

Using this approach, the second author created PUMFs for KNSO which were released to the public. It should be noted that, in addition to the above, the second author used the broadest occupation codes and masked some other variables such as the size of living space and lot size.

References

1. Bethlehem, J., Keller, W., and Pannekoer, J. (1990). *Disclosure Control for Microdata, Journal of the American Statistical Association*, 85, 38-45.
2. Cox, L.H. and Sande, G. (1979). *Techniques for Preserving Statistical Confidentiality, Bulletin of the International Statistical Institute*, 42.3, 409-512.
3. Elliot, M. (2001), *Disclosure Risk Assessment, in: Confidentiality Disclosure and Data Access, North-Holland* 75-90.
4. Feinberg, S. and Markov, U. (1998), *Confidentiality, Uniqueness, and Disclosure Limitation for Categorical Data, Journal of Official Statistics,* 14, 385-397.
5. Khare, M., Battaglia, M.P. and Hoaglin, D.C. (2003). *Procedures to Reduce the Risk of Respondent Disclosure in a Public-Use Data File: The National Immunization Survey, Federal Committee Statistical Methodology Research Conference,* 5-12.
6. Lambert, D. (1993), *Measures of Disclosure Risk and Harm, Journal of Official Statistics,* 9, 313-331.
7. Marsh, C., Skinner, C., Arber, S., Penhale, B., Openshaw, S., Hobcraft, J., Lievesley, D. and Walford, N. (1991), *The Case for Samples of Anonymized Records from the 1991 Census, Journal of the Royal Statistical Society A, 154, Part 2*, 305-340.
8. Korean population census, www.nso.go.kr/eng2006/emain/index.html
9. Skinner, C., Marsh, C., Openshaw, S., and Wymer, C. (1994), *Disclosure Control for Census Microdata, Journal of Official Statistics, 10,* 31-51.

Statistical Disclosure Control for the 2011 UK Census

Jane Longhurst*, Nicola Tromans*, Caroline Young*, and Caroline Miller*

* The Office for National Statistics, Segensworth Road, Titchfield, Fareham, PO15 5RR, UK
Jane.Longhurst@ons.gov.uk, Nicola.Tromans@ons.gov.uk, Caroline.Young@ons.gov.uk, Caroline.Miller@ons.gov.uk

1. Introduction

This paper describes the strategy that is being employed to develop a Statistical Disclosure Control (SDC) solution for the 2011 Census. The key aim is to ensure a harmonised UK SDC strategy for all outputs (pre-defined tabular outputs, microdata samples and possibly flexible user defined tabular outputs) which ensures that the public interest in the figures is met while managing data confidentiality risks. The most desirable qualities for the SDC strategy are;

- Maximum data utility
- Minimum disclosure risk
- Acceptable to users
- Simple to understand and transparent
- Easy to implement

The next section of the paper provides a brief background to the Census and disclosure control, and Section 3 gives a high level description of the project. Section 4 provides a high level overview of possible SDC methods that could be used to protect census tables and focuses on two methods as examples; record swapping and cell perturbation. Section 5 details the short-listed SDC methods and the criteria used to reach this short-list, and Section 6 provides an evaluation of two example methods demonstrating the approach that has been adopted to decide on the SDC method(s) that will be used to protect 2011 Census outputs.

2. Background

Every 10 years since 1801, the UK has set aside one day for the census, whereby information is obtained on every member of the population. It is the most complete source of information about the population that we have with details of family composition, health, employment and other socio-economic characteristics. The information provided allows central and local Government, health authorities and many other organisations to target their resources more effectively and to plan housing, education, health and transport services for years to come. The next census is due to take place in 2011.

Census data is released in a number of different formats; standard pre-planned tables, commissioned tables requested by users and Census sample microdata. In addition in

2011 the aim is to release user defined tables via flexible table generating web-based software. Publishing aggregate or individual data carries the risk that individuals or entities could be identified and confidential information about them could be released. The UK Census Offices need to protect the confidentiality of census respondents for a number of reasons. The production and use of official statistics depends on the cooperation and trust of citizens. Such trust cannot be maintained unless the privacy of individuals' information is protected. There are also legal and policy obligations that must be respected. The Census Act 1920 as amended by the Census (Confidentiality) Act 1991 and the Census Act (Northern Ireland) 1969 as amended by the Census (Confidentiality) (Northern Ireland) Order 1991, make it an offence for the Registrars General (or any person under their control or a supplier of any services to them) to disclose any personal census information to another person without lawful authority. The National Statistics Code of Practice sets out principles for protecting confidentiality. These include the principle that 'The National Statistician will set standards for protecting confidentiality, including a guarantee that no statistics will be produced that are likely to identify an individual unless specifically agreed with them.'

The aim of SDC is to ensure that statistical outputs provide as much value to the users while protecting the confidentiality of information concerning individuals or entities. SDC methods modify, summarise or perturb the data and a range of different methods can be used to protect different census outputs. SDC methods can be pre-tabular (applied to the underlying microdata) or post-tabular (applied to tables).

A pre-tabular method of disclosure control, random record swapping, was initially planned for the 2001 UK census tables. This method of disclosure control was followed up by applying population thresholds to the tables. The General Register Office for Scotland (GROS) adopted smaller thresholds than the Office for National Statistics (ONS) and the Northern Ireland Research Agency (NISRA). Prior to releasing tabular outputs from the 2001 Census concerns were raised that the public would perceive that no disclosure control method had been applied. ONS decided that the additional method of small cell adjustments was required for tabular outputs. The small cell adjustments added more uncertainty and removed small cells from tabular outputs. NISRA also applied the additional method of small cell adjustment but GROS did not. This late change in SDC methodology and lack of UK harmonisation caused a number of problems for users. A different SDC technique was used to protect the microdata samples or Sample of Anonymised Records (SARs) from the 2001 Census. The disclosure risk was reduced by recoding variables and applying PRAM (Post-Randomisation Method), a perturbative microdata disclosure control technique for categorical variables.

3. Approach

3.1. Development and agreement of UK SDC Policy for 2011 Census Outputs

In November 2006 the UK SDC Policy position for the 2011 Census was agreed by the Registrars General of Scotland, England and Wales and Northern Ireland. The

Registrars General have agreed to aim for a common UK SDC methodology for 2011 Census outputs to achieve harmonisation. The SDC Policy position is based on the principle of protecting confidentiality set out in the National Statistics Code of Practice. The Registrars General concluded that the Code of Practice statement can be met in relation to census outputs if no statistics are produced that allow the identification of an individual (or information about an individual) with a high degree of confidence. The Registrars General consider that, as long as there has been systematic perturbation of the data, the guarantee in the Code of Practice would be met. It has therefore been agreed that small counts (0's, 1's, and 2's) could be included in publicly disseminated Census tables provided that (a) uncertainty as to whether the small cell is a true value has been systematically created; and (b) creating that uncertainty does not significantly affect data quality.

The exact threshold of uncertainty required has not been decided. The Registrars General will make this judgement at a later stage in the context of results from methodological research into the balance of protection afforded, and changes to data quality caused by various SDC methods. The decision to allow small cells in publicly disseminated tables means that both pre-tabular methods and post-tabular methods or combinations of the two can be considered for 2011. The Registrars General have expressed a preference for pre-tabular methods, provided there is not undue change to the quality of the data.

The UK SDC policy also highlighted the following points;

- Aim is to make as much of the census tabular outputs as possible publicly accessible. However, if certain tabular outputs are seriously compromised by SDC then these could be released under other access arrangements (e.g. licence or safe setting) where data access restrictions allow less stringent levels of SDC to apply in order to increase data utility.
- It is considered that attribute disclosure is the key disclosure risk, because identification reveals no new information to the user. Attribute disclosure involves a user discovering something new from the census data that was not previously known to them.
- Consistency and additivity across tabular output is a priority for users and these will be given a high priority when assessing the utility of SDC methods.
- Methods will be chosen which afford an acceptable level of protection and preserve the highest level of utility of outputs.
- Clear explanations will be given to users and expert audiences on the protection afforded by the SDC strategy and other steps applied which protect confidentiality.
- SDC methods for all types of census output will be assessed concurrently because of their interdependencies.
- Users will be updated and consulted during the research period.
- An Independent review will be conducted by the UK Census Design and Methodology Advisory Committee (UKCDMAC).

3.2. Quality Assurance for the 2011 Census SDC Strategy

A UK SDC working group has been formally set up to steer work, provide advice and quality review work associated with developing the SDC methodology for the 2011 Census. The working group consists of representatives from all three UK Census Offices to ensure a harmonised approach to the development of the 2011 Census UK SDC Strategy that is in line with the agreed policy.

A Disclosure Control Subgroup of UKCDMAC has also been set up This subgroup has been responsible for providing advice on methodological issues and has acted as a formal quality review panel for the SDC workpackage prior to seeking methodological agreement from UKCDMAC, and formal sign off from the UK Census Committee (UKCC).

3.3. 2011 Census UK Statistical SDC Work plan

A work plan for the methodological research phase of the 2011 Census UK SDC strategy has been developed. The research addresses (pre-defined) tabular outputs, microdata samples and flexible user defined tabular outputs whilst taking into account the impact of interactions between these types of output. An outline of the agreed approach for developing the SDC strategy follows.

The initial stage of methodological research involved conducting a review of SDC in a census context. This review has facilitated the development of the SDC strategy for the 2011 Census by drawing together;

i) research conducted prior to implementing SDC in the 2001 Census
ii) reasoning behind SDC decisions for the 2001 Census
iii) evaluations of SDC methods used in the 2001 Census
iv) lessons learnt from 2001
v) international approaches to SDC
vi) work already conducted for the 2011 Census

Following this a high level review has been conducted to address the advantages and disadvantages of a wide range of SDC techniques for protecting all types of 2011 Census outputs, and the issues concerned with implementation and the interactions between the outputs. Examples are provided in Section 4. Using this high level review, a preliminary list of SDC techniques which should be explored further has been drawn up (see Section 5). SDC methods not on this short-list have been discounted from further research.

The short-listed disclosure control methods are to be evaluated using a disclosure risk - data utility framework (Shlomo and Young 2006). This quantitative evaluation will follow the approach used by Shlomo (2006) and will be used to identify the recommended SDC method(s) for the 2011 Census for all types of outputs although the focus will be on tabular outputs. Examples of this evaluation are provided in Section 6.

An additional stage of research has been timetabled to further develop methods for safeguarding microdata to ensure sufficient protection from disclosure.

It is vital that the development of the 2011 Census UK SDC strategy takes account of the interdependencies which exist with the 2011 Downstream Processing Schedule and work to design 2011 Census Geography and Outputs. Findings from user consultations will be incorporated into the evaluation process and users will be consulted and updated with research findings and decisions as appropriate. At this early stage, the final sign-off of the UK SDC strategy has been timetabled for July/August 2009.

4. SDC methods

This paper focuses on SDC methods for protecting census tabular outputs rather than microdata samples although the dependencies between the methods used to protect different outputs will be recognised in the evaluation stage. SDC methods for census tables implemented at National Statistical Institutes include both pre-tabular and post-tabular methods or combinations of both. Pre-tabular methods are implemented on the microdata prior to the tabulation and typically include forms of record swapping between a pair of households matching on some control variables (Willenborg and de Waal, 2001). This method has been used for protecting census tables at the US Census Bureau (in 1991 random record swapping was used whereas targeted record swapping was used in 2001) and for the 2001 UK Census. Record swapping can be generalized into a pre-tabular method called PRAM (the Post-Randomization Method) (Gouweleeuw, Kooiman, Willenborg and De Wolf, 1998). This method adds 'noise' to categorical variables by changing values of categories for a small number of records according to a prescribed probability matrix and a stochastic process based on the outcome of a random multinomial draw. Another pre-tabular SDC method is over-imputation. This involves randomly deleting variables in existing records and imputing the variables using the Edit and Imputation System already in use during census processing.

Post-tabular methods are implemented on the entries of the tables after they are computed and typically take the form of random rounding, either on the small cells of the tables or on all entries of the tables. Small cell adjustments (rounding) have been carried out on the census tables by the Australian Bureau of Statistics (ABS) in 2001 and the ONS for the 2001 Census in England and Wales, and full random rounding has been carried out by Statistics Canada and Statistics New Zealand. Controlled rounding is a procedure that uses linear programming techniques to round entries up or down and ensures that all rounded entries add up to the rounded totals. It is available in the SDC software package, Tau Argus, (Hundepool, 2002), however, at present the controlled rounding option is not able to cope with the size, scope and magnitude of census tabular outputs. Other post-tabular methods include cell suppression or some form of random perturbation on the internal cells of the census tables. Cell suppression is not primarily used in the census context because of the large number of cells that need to be consistently suppressed. The ABS have developed a cell perturbation method for their 2006 Census that is designed to potentially alter every cell in every table by a small

amount, remove all small cells, always randomise the same table in exactly the same way and ensure additivity.

In addition to the methods described above disclosure risk can also be managed by sampling from the population database or by restricting the design/complexity of the tables, setting geographical thresholds or implementing rules that determine the sparsity of tables, e.g. minimum average cell size.

This paper focuses on record swapping and the ABS cell perturbation method in order to demonstrate the evaluation that will be undertaken for SDC methods for 2011 UK Census.

4.1. Record Swapping

Record swapping involves exchanging geographical variables between randomly selected pairs of households within the Census data. In order to minimise bias pairs of households are determined which match on some control variables, such as a large geographical area and age-sex distribution of the households. Record swapping can be targeted to high-risk households ensuring that households most at risk of disclosure are likely to be swapped. Record swapping can also be modified to take into account imputation rates, i.e. by only swapping those records with no imputation. In a census context, geography variables are often swapped between households because this results in less edit failures due to the assumption that other census variables are independent of geography. Swapping geographical variables also means that at higher geographical levels and within control strata marginal distributions are preserved.

For this analysis, random record swapping was carried out for a 10% swapping rate. The control variables used to determine the pairs of households were the number of persons in the household according to sex and three broad age groups and a "hard-to-count" index of the household based on the 1991 UK Census enumeration. The record swapping was carried out within a large geographical area (Local Authority (LA)) and households were swapped in and out of small geographical areas (Output Areas (OA)). A targeted record swapping was also carried out by defining an additional control variable based on a "flag" for the household that had at least one person in a small cell in one of the census tables under evaluation. Note that on average, about 0.15% of the households selected for swapping were not swapped because no paired record was found. Those records would have to be swapped outside the large geographical area (LA) but this was not carried out in this analysis.

Table 1 illustrates the advantages and disadvantages of record swapping.

Advantages	Disadvantages
o Swapping rates are flexible. o Marginal distributions preserved at a higher geographical level and within control strata. o Given household characteristics (used as control variables), other census variables are likely to be independent of geography therefore less bias will occur. o Maintains internal consistencies within households. o Control variables (variables upon which swapped records must match) can be determined according to requirements. o Swapping geographies will not necessarily result in inconsistent and illogical records i.e. less edit failures. o Provides some protection against differencing (geographical and variable differencing). User defined flexible table generation possible provided record swapping provides enough protection or used in addition to other suitable SDC methods. o Previous experience of using record swapping in 2001 Census.	o Users cannot be provided with the swap rate hence no measure of whether a value in a table is the true value - Difficult for analysts to properly account for its impacts at levels below Local Authority. o Rare combinations of variable characteristics are still disclosive (i.e. special uniques). o Using current methods all geographic fields such as workplace are swapped hence work place tables not protected. o Generally introduces bias into the results at geographical levels below Local Authority and causes lower level characteristics to become more homogeneous.

Table 1: The Main Advantages and Disadvantages of Record Swapping.

There are also additional advantages and disadvantages associated with both random and targeted record swapping. Random record swapping maintains a higher data utility compared with targeted record swapping at the same swap rate, however, targeted record swapping provides a greater level of protection against disclosure since it targets the risky records. Targeted record swapping results in a greater distortion to tabular distributions (particularly the joint distributions) compared to random record swapping since perturbation is carried out on uniques and outliers rather than at random.

4.2. ABS Cell Perturbation

For the protection of their 2006 Census outputs, the ABS has conducted research into a new cell perturbation algorithm (Fraser and Wooton 2006). In the past they have released static tables of data however flexible table generation will be used for 2006. This will enable users to design and populate their own tables. The new perturbation algorithm is designed to protect these tables by potentially altering every cell in every table by a small amount. In doing so it adds sufficient 'noise' to each cell so that by differencing, users would end up with more noise than real data. The algorithm always randomises the same table in exactly the same way. It also preserves higher level totals between tables with common geographies. The SDC algorithm involves two stages; the

first adds the perturbations to the cell values and the second stage restores additivity to the table.

Perturbation Stage

1 A value m is predetermined defining the range of the perturbation distribution.
2 Each record in the microdata is assigned an rkey or record key. The rkey is a value drawn at random from the discrete uniform distribution [0, m-1].
3 Each table is then considered independently. The rkeys relating to the records making up each cell in the table are combined to give a cell key or ckey as follows:
ckey = $\mathrm{mod}\left|\sum(rkey), m\right|$
The use of the mod (remainder) function means that the distribution of the ckeys is also a discrete uniform on [0, m-1].
4 A look-up table is defined with original cell values on the rows and ckeys on the columns. Thus the lookup table will have m-1 columns and the maximum cell value in the original table will correspond to the number of rows.
5 The look-up table provides the perturbation value relating to each cell determined by the original cell value (row) and the ckey (column).
6 This perturbation is then added to the original cell value.

ABS have designed look-up tables which minimise bias and preserve certain variances but in theory the look-up table can be specified according to the needs of the statistical agency. For example the first row of the look-up table could be specified with all zeros which means all original cell values of zero have zero perturbation added, moreover, it would also be possible to design the look-up table such that all ones and twos are removed from tabular output (as they plan to do for the ABS 2006 Census). In fact, the look-up table could also be designed to mimic the effects of other SDC procedures such as random rounding.

Additivity Stage

After the perturbation stage, the same cell in different tables is consistent (has the same perturbation added). However the tables do not add up. Additivity is restored using an iterative algorithm which visits single and pairs of cells adding -1, 0, +1 at each iteration stopping when all rows and columns add up. It does this at the same time as minimising the overall difference between the additive and original table. For this analysis only the perturbation stage has been implemented since the code for the additivity stage is not currently available. The following look-up table was used:

Original Cell value	**Perturbation to be drawn from the following distribution (using the cell key)**
0	Remain as zeros
1	Normal distribution with mean 0 and variance 2 truncated at -1 and +5
2	Normal distribution with mean 0 and variance 2 truncated at -2 and +5
3	Normal distribution with mean 0 and variance 2 truncated at -3 and +5
4	Normal distribution with mean 0 and variance 2 truncated at -4 and +5
5+	Normal distribution with mean 0 and variance 2 truncated at -5 and +5

The ABS cell perturbation method is a more informed post-tabular method of disclosure control since it utilises microdata information during the perturbation stage. Table 2 summarises the advantages and disadvantages of this method.

Advantages	Disadvantages
o Tables are consistent o Provides protection for flexible tables o Depending on the look-up table design, the method can perturb distributions that are approximately unbiased with small variances o Efficient - allegedly has a quick run time o Able to produce perturbations for large high dimensional hierarchical or cross classified tables o Protects against differencing o Method is flexible; look-up table can be specifically designed to suit needs and different look-up tables could potentially be used for different tables. Moreover the look-up table could be designed to mimic random rounding for example.	o Tables not additive (additivity module is not applied here) o Once additivity is applied, consistency is lost o Needs to be applied to each table separately o Public perception that no disclosure control has been applied (unless incorporated into look-up table) o Method less transparent than others e.g. rounding o Depends on the appropriate choice of look-up table which may not be suitable for all tables (i.e. sparse) o Statistical effects are highly dependent on the choice of look-up table

Table 2: The Main Advantages and Disadvantages of ABS Cell Perturbation

Both SDC methods are feasible for the 2011 Census but have their limitations. Both provide protection against disclosure by differencing which will be important if flexible user defined tabular outputs are to be made available in 2011. Record swapping ensures that marginal distributions are preserved at a higher geographical level and within control strata and results in additive and consistent tables. However, random record swapping can result in a high proportion of risky cells left unaltered. Targeting the risky cells reduces the risk of disclosure at the same record swapping level but causes greater distortion to tabular distributions. The ABS method (when applying both stages) results in additive tables however tables representing the same population subgroups may not end up with consistent totals.

5. Short-listing SDC methods for quantitative evaluation

The following pre-tabular and post-tabular methods were considered for short-listing:

1. Record Swapping
2. Over-Imputation
3. Data Switching
4. Post Randomisation Method (PRAM)
5. Sampling
6. Conventional Rounding
7. Random Rounding
8. Controlled Rounding
9. Semi-Controlled Rounding
10. Suppression
11. Barnardisation
12. ABS Cell Perturbation Method

The quantitative risk - utility framework being used to evaluate the SDC methods (see Section 6) for the 2011 Census is not sufficient on its own. Many SDC methods have qualities which cannot be accounted for quantitatively and thus qualitative advantages and disadvantages of SDC methods were addressed to enable a short-list of methods to be produced prior to commencing the in-depth quantitative evaluation. Each of the methods was assessed by representatives from the UK Census Offices on whether they met, partly met or did not meet the seven qualitative criteria listed below. The criteria were split into primary and secondary criteria and an additional requirement was that any method that did not meet one of the primary criteria would not be considered for short-listing:

Primary Criteria

i) Will the method provide additive and consistent tables which are a priority for users?
ii) Overall, will users accept the method?
iii) Does the method protect against (geographical and categorical) differencing?
iv) Is the method practical, feasible to implement and has it been used for protecting similar outputs to date?

Secondary Criteria

i) Method should not restrict microdata releases
ii) Method should be simple to understand
iii) Method should be easy to account for in analyses

Following this assessment four methods were chosen for short-listing:

- Record Swapping
- Over-Imputation
- ABS Cell Perturbation Method
- Small Cell adjustment with record swapping (included to provide a comparison with 2001)

All these methods passed the primary criteria although it should be noted that small cell adjustment on its own would have failed as it did not protect against differencing. Small cell adjustment is a form of rounding which only applies to the small cells in a table. It was agreed to include it (with record swapping) in the final short-list to ensure that the first three methods could be compared against the 2001 method. Discounted methods will be excluded from further consideration and an assessment of the short-list of SDC methods using the risk - utility framework will shortly begin. Combinations of the short-listed methods will also be considered for the quantitative evaluation. The short-list will allow an SDC strategy to be developed for the 2011 Census that will meet the needs of the UK SDC policy.

Example results from some preliminary work evaluating the risk and utility of record swapping and the ABS cell perturbation method on tabular outputs are presented in Section 6.

6. Quantitative Analysis of Proposed SDC methods

As described above the short-list of SDC methods will be evaluated quantitatively focusing on an assessment of the impact of the method on data utility and disclosure risk. A software package (Shlomo and Young, 2006) developed to calculate a variety of information loss metrics (by comparing the protected data with the original pre-disclosure controlled data) will be used for this analysis. Here we present a selection of the information loss measures and one risk measure described in Shlomo and Young, 2006 and use them to compare record swapping and ABS cell perturbation for two example tables. It should be noted that these are included as an illustration of the analysis that will be undertaken. A more thorough analysis investigating further methods using a wide range of tables, varying parameters (e.g. swapping rates, look-up table), and further disclosure risk and information loss measures will be required for the final analysis of the short list.

6.1. Data

The effects of the SDC methods will be considered for two tables at two different levels of geography, Output Area (OA) and ward level for an Estimation Area in England relating to Southampton, Eastleigh and Test Valley. These two tables were selected to study whether the methods have varying effects over different levels of geography. Geography is represented as rows in the table and the other variables span the columns. Table 3 describes the structure of the two tables.

	Variables and Number of Categories	Number of Persons in the Table	Number of Internal Cells	Average Cell Size	Number of Zeros	Number of Small Cells
Table A	Religion (9) Age-Sex (6) OA (1,487)	437,744	80,298	5.45	47,433 (59.1%)	10,137 (12.6%)
Table B	Economic Activity (9) Sex (2) Long term illness (2) Ward (70)	317,064	2,520	125.82	427 (16.94%)	226 (8.97%)

Table 3: Example tables

6.2. Risk and Utility Measures

6.2.1. Disclosure Risk

Let R_i represent the record *i*, I the indicator function having a value of 1 if true and 0 if false, C_1 the set of cells with a value of 1, C_2 the set of cells with a value of 2, $|C_1 \cup C_2|$ the number of cells with a value of 1 or 2. The disclosure risk measure can be interpreted as the percentage of records in small cells not perturbed:

$$DR = \frac{\sum_{i \in C_1 \cup C_2} I(R_i \ not \quad perturbed \ or \quad imputed)}{|C_1 \cup C_2|}$$

6.2.2. *Distance Metrics on Internal Cells of the Tables*

Distance metrics are used to measure distortion to distributions. A distance metric is calculated for each row in the table and then the overall average across all of the rows is taken as the information loss measure. This format is used since the rows in census tables generally represent a geographical area whereas the columns define the categories of a specific table, such as sex×age group×economic activity. When comparing the average distance metric across rows, we need to take into account the level of dispersion as expressed by the standard error (confidence interval).

Let D^k represent a row k of table D and let $D^k(c)$ be the cell frequency c in the table. Let n_r be the number of rows in the table. *Pert* refers to the disclosure-protected table and *orig* to the original table. The distance metrics are:

Hellinger's Distance:

$$HD(D_{pert}, D_{orig}) = \frac{1}{n_r} \sum_{k=1}^{n_r} \sqrt{\sum_{c \in k} \frac{1}{2} (\sqrt{D^k_{pert}(c)} - \sqrt{D^k_{orig}(c)})^2}$$

Relative Absolute Distance:

$$RAD(D_{pert}, D_{orig}) = \frac{1}{n_r} \sum_{k=1}^{n_r} \sum_{c \in k} \frac{| D^k_{pert}(c) - D^k_{orig}(c) |}{D^k_{orig}(c)}$$

Average Absolute Distance per Cell:

$$AAD(D_{pert}, D_{orig}) = \frac{1}{n_r} \sum_{k=1}^{n_r} \frac{\sum_{c \in k} | D^k_{pert}(c) - D^k_{orig}(c) |}{n_k} \quad \text{where } n_k = \sum_c I(c \in k)$$

the number of cells in the k^{th} row.

The standard errors are calculated as follows (for example, the *AAD* metric):

$$\frac{1}{n_r - 1} \sum_{k=1}^{n_r} (AAD(D^k_{pert}, D^k_{orig}) - AAD(D_{pert}, D_{orig}))^2$$

where $AAD(D^k_{pert}, D^k_{orig}) = \frac{\sum_{c \in k} | D^k_{pert}(c) - D^k_{orig}(c) |}{n_k}$

These distance metrics can also be calculated for sub-totals and totals of the tables.

6.2.3. *Variance of Cell Counts*

An information loss measure can be calculated to measure the impact on the variance of the estimates. The variance of the counts is examined across the rows before and after the SDC methods as follows:

For each row *k*, we calculate: $V(D^k_{orig}) = \frac{1}{n_k - 1}\sum_{c \in k}(D^k_{orig}(c) - \overline{D}^k_{orig})^2$ where $\overline{D}^k_{orig} = \frac{\sum_{c \in k} D^k_{orig}(c)}{n_k}$ and $n_k = \sum_c I(c \in k)$ the number of cells in the k^{th} row. Next we calculate the ratio for each row:

$$VR(D^k_{pert}, D^k_{orig}) = \frac{V(D^k_{pert})}{V(D^k_{orig})}$$

6.2.4. *Change to Rank Orderings*

Changes to the underlying ordering of cell counts (impact on rank correlation) within the table can be studied. The original counts are sorted according to their size and deciles (10 equal groupings) $v^{orig}(c)$ are defined. This is repeated for the perturbed cell counts which are sorted according to both their size and the original order in order to maintain consistency for the tied variables. Deciles $v^{pert}(c)$ are then defined for the perturbed variable after the sort. The information loss measure is the percent of cells that have changed deciles. The measure is calculated across different categories in the table e.g. table columns; an overall average is the final measure:

$RC = \frac{100 \times \sum_{c \in k} I(v^{orig}_k \neq v^{pert}_k)}{n_k}$ where I is the indicator function and is 1 if the statement is true and 0 otherwise, k is a column in the table and n_k is the number of cells in that column.

6.3. Results

Table 4 displays the risk measure for each SDC methods for the two tables.

Probability that a record in a small cell has not been perturbed	10% Random Swap	10% Targeted Swap	ABS Cell Perturbation
Risk (ward)	0.619	0.509	0.142
Risk (OA)	0.651	0.506	0.188

Table 4: Disclosure risk measures

The risk is far smaller for the ABS method in comparison to record swapping because there is a higher probability that a small cell would receive a non-zero perturbation. The targeted swap focuses on perturbing small cells and hence the risk is less than for the random swapping method.

Tables 5 and 6 display distance metrics at OA and ward level respectively.

OA – Distance Metrics	10% Random Swap	10% Targeted Swap	ABS Cell Perturbation
Hellingers' Distance	1.2875 (0.0249)	1.6027 (0.0265)	1.7388 (0.0228)
Relative Absolute Distance	4.2542 (0.1149)	5.2674 (0.1289)	11.2215 (0.2767)
Absolute Average Distance	0.4870 (0.0100)	0.5275 (0.0093)	0.6217 (0.0088)

Table 5: Distance metrics, OA level

Ward – Distance Metrics	10% Random Swap	10% Targeted Swap	ABS Cell Perturbation
Hellingers' Distance	1.2389 (0.1211)	1.3811 (0.1177)	1.2167 (0.1113)
Relative Absolute Distance	3.4715 (0.4775)	4.0725 (0.5093)	6.1881 (1.0078)
Absolute Average Distance	3.6897 (0.6579)	3.7048 (0.5732)	1.300 (0.1235)

Table 6: Distance metrics, ward level

At the OA level the ABS method performs the worst for all three distance metrics because there is a high probability that small cells are perturbed using our specified look-up table (see section 4.2) and at OA level the table is particularly sparse (see Table 3). In all cases the targeted swap distorts the distributions in the table more than the random swap as expected. The best method (in terms of distortions to distributions) is the random swap in this case, but the results in general would depend on the table and the distance metric considered. The standard errors for each measure are displayed in brackets.

Table 7 shows the impact of the SDC methods on marginal totals.

Change in marginal totals (relative difference) BY SUBGROUP	10% Random Swap	10% Targeted Swap	ABS Cell Perturbation
Sex by age-group (over all OAs) (table A)	0	0	1.078
Religion (over all OAs) (table A)	0	0	3.468
Sex by long term illness (over all wards) (table B)	0	0	0.007
Economic activity (over all wards) (table B)	0	0	0.024

Table 7: Distance metrics, marginal totals

The random and targeted record swapping result in no change to the marginal distributions of the tables. This result occurs because by definition record swapping maintains the marginal distributions at levels above Local Authority District and the marginals here represent subgroups of the Estimation Area. The marginal totals representing Estimation Area by religion are affected by the greatest change in relative difference when performing the ABS method. This result is likely to be caused by the uneven distribution of marginal counts across religions resulting in a greater number of small cells which are affected to a greater extent by this method. The other marginal totals considered are affected to a lesser degree because the marginal counts will be more evenly distributed across all variable categories and hence the perturbations applied to the marginal cells are small relative to the marginal count.

Table 8 shows the impact of the SDC methods on the variability of cell counts.

Average variance ratio over all rows	10% Random Swap	10% Targeted Swap	ABS Cell Perturbation
OA	0.9944	0.9985	1.0326
Ward	1.0263	1.0079	1.3389

Table 8: Variance of cell counts

The impact of the SDC methods on the variability of cell counts by row is not significant since no firm patterns can be seen. At ward level, the effect of the SDC method on the variance appears to be more noticeable with the ABS method increasing the variance. The results from the ABS cell perturbation method are dependent on the look-up table and can vary if the perturbation distributions are changed.

Table 9 displays the impact of the SDC methods on the rank ordering of cells.

Cells moved into different percentile (groups of 10)	10% Random Swap	10% Targeted Swap	ABS Cell Perturbation
OA	26%	34%	20%
Ward	2%	3%	2%

Table 9: Change to Rank Orderings

This test shows how swapping and to some extent cell perturbation distorts the underlying patterns in the data by changing the rank order of cells. At OA level there is a lot of distortion because more than 70% of cells have values less than 3 whereas at ward level there is greater variation in the cell counts so the SDC methods have less of an impact.

7. Summary

This paper has described the approach that will be adopted to develop the SDC strategy for all 2011 Census outputs. A review of past work (particularly undertaken for 2001) has been conducted and is being used to inform further stages of the project. A high level review of SDC methods has been conducted and a summary of the approach used

to develop the shortlist of methods for further evaluation has been described. Examples from the high level review and a quantitative evaluation (measuring risk and information loss) have been presented for two SDC methods; record swapping and a cell perturbation method. These preliminary results are included as an illustration of the final more detailed evaluation that will be undertaken. It is recognised that developing a 2011 UK SDC strategy which satisfies all user requirements whilst maintaining a high level of data utility is likely to be an unachievable task hence compromises will need to be made. The final recommended approach to SDC for 2011 Census will be informed by both quantitative and qualitative evaluation and the trade-offs between the different methods will need to be communicated to users.

References

Fraser, B. and Wooton, J. (2006) A proposed method for confidentialising tabular output to protect against differencing, Internal report, Data Access and Confidentiality Methodology Unit, ABS.

Gouweleeuw, J., P. Kooiman, L.C.R.J. Willenborg and P.P. De Wolf (1998). Post Randomisation for Statistical Disclosure Control: Theory and Implementation. *Journal of Official Statistics*, 14, pp. 463-478.

Hundepool, A. (2002). The CASC Project. In Domingo-Ferrer, J. (eds.): *Inference Control in Statistical Databases: From Theory to Practice*. *Lecture Notes in Computer Science*, Vol. 2316. Springer-Verlag.

Shlomo, N. and Young, C. (2006) Statistical Disclosure Control Methods through a Risk - Utility Framework: Proceedings of the Privacy in Statistical Databases CENEX-SDC Project International Conference, Rome, 13-15 Dec 2006.

Shlomo, N. (2006) Review of Statistical Disclosure Control Methods for Census Frequency Tables. *Survey Methodology Bulletin, 57, Office for National Statistics.*

Willenborg, L. and de Waal, T. (2001). *Elements of Statistical Disclosure Control, Lecture Notes in Statistics*, 155 , Springer Verlag, New York.

Applying Tau-Argus to SuperCROSS tables: A practical example using the UK Business Register Unit data

Andrea Toniolo Staggemeier[1], Philip Lowthian[2], and Grant Lee[3]

[1] Information Management (Strategies), Office for National Statistics, Newport, United Kingdom, andrea.staggemeier@ons.gsi.gov.uk
[2] Methodology Directorate (Statistical Disclosure Control), Office for National Statistics, London, United Kingdom, philip.lowthian@ons.gsi.gov.uk
[3] Space-Time Research Pty Ltd, Melbourne, Australia, grant.lee@str.com.au

Abstract. The Business Register Unit (BRU) at the Office for National Statistics (ONS) in the UK produces a wide range of tabular outputs, many of which are published on the ONS website. SuperCROSS is a tabulation tool used by the Office for National Statistics (ONS) and the idea of linking this program with the disclosure control package Tau-Argus has been developed over the last 3 years, culminating with the implementation of the SuperCROSS/Tau-Argus link live in production of BRU outputs from August 2007. A table created in SuperCROSS can be confidentialised by the user selecting the required rules from a drop down menu. This action opens Tau-Argus and uses a batch file to either round or suppress cells in the table. The safe table is then returned to SuperCROSS without the need for the user to interact with Tau-Argus.

This paper describes the interface between the two programs, typical rules that can be applied, how those rules are set, and the different output formats available. Also discussed in this paper are the reasons behind why a link between Tau-Argus and SuperWEB (thin client tool from SuperSTAR product suite) is not recommended as an alternative to SuperCROSS (thick client, i.e. desktop installation tool) within ONS. Finally benefits of this approach to both individual business areas and National Statistics Institutes (NSIs) in general are given.

Keywords. SuperCROSS, Tau-Argus, Statistical Disclosure Control, Controlled Rounding

1 Introduction

The idea to develop a link between a tabulation tool and a statistical disclosure tool was initiated by the Office for National Statistics (ONS) in 2004. One business area within ONS, the Business Register Unit (BRU), who have been long term users of SuperCROSS[1] for tabulation, had a business need for a more robust and efficient means of Statistical Disclosure Control (SDC) in order to protect their outputs. It was highly desirable to retain SuperCROSS for tabulation, and the tool of choice for SDC within ONS was Tau-Argus[2]. An interface has since been developed in SuperCROSS to allow tables to be passed seamlessly to Tau-Argus, and back to SuperCROSS once disclosure has been applied; all without any manual interaction from the user.

The process that BRU followed prior to this integration of tools was described as time consuming, labour intensive, and with some level of risk to publishing disclosive tables. As a consequence of this risk, less than optimal information was published in each table in order to minimise any event of disclosure.

1.1 Business Problem

This paper examines outputs produced by the BRU from the Inter-Departmental Business Register (IDBR[3]). This is the comprehensive list of UK businesses used for statistical purposes. IDBR also provides a sampling frame for surveys of businesses carried out by the ONS and by other government departments. It is therefore a key data source for analyses of business activity in the UK.

The IDBR covers businesses in all parts of the economy, other than some very small businesses (self-employed and those without employees and low turnover) and some non-profit organisations. With 2.1 million businesses listed, it provides nearly 99% coverage of UK economic activity. It holds a wide range of information on business units including: name; address; standard industrial classification; employment and employees; and turnover.

The amount of outputs that BRU produce require a well thought out process, including a procedure to apply the required disclosure control rules using Excel spread sheets. Historically, both rounding (for frequency tables) and suppression (for magnitude tables) have been carried out. Since this was not automated, the process was very time consuming and there was a potential risk that unsafe outputs might be produced.

[1] SuperCROSS is the client tabulation tool, and part of the SuperSTAR suite from Space-Time Research (http://www.spacetimeresearch.com)
[2] Tau-Argus (http://neon.vb.cbs.nl/casc)
[3] IDBR website (http://www.statistics.gov.uk/CCI/nugget.asp?ID=195)

The tabulation tool, SuperCROSS, already in use by BRU, was enhanced to provide an interface to support the disclosure control process in Tau-Argus. The following sections will briefly discuss Tau-Argus and SuperCROSS, the approach to integrating the two technologies, and a more detailed description on how the link and the application of safety rules were implemented.

1.2 Tau-Argus

Tau-Argus is a software tool which enables statistical disclosure control to be carried out to protect tabular output. It can be run in either interactive or batch mode and can import tables or microdata, allowing the user to create tables. Tau-Argus supports either frequency or magnitude data types and once imported along with a metadata file, the user can apply a number of confidentiality rules.

Typically for magnitude tables, safety rules such as threshold and dominance rules are set by the user, and cells failing these rules are highlighted, allowing the user to select them for suppression. In order to avoid disclosure by differencing, secondary suppression can be applied using a variety of techniques. For frequency tables, controlled rounding is commonly applied. This method rounds cell values to the nearest multiple of a user specified base, whilst maintaining the table additivity.

Tau-Argus was initiated as result of the CASC (Computational Aspects of Statistical Confidentiality)[4] project, which was European Union (EU) funded with additional support from many statistical organisations, including ONS, and EU Universities. Tau-Argus is currently being used by BRU and by a number of other Government Departments and Agencies who supply data for the ONS Neighbourhood Statistics website

1.3 SuperCROSS

SuperCROSS is part of the SuperSTAR Suite of products, developed by Space-Time Research. As a desktop tabulation tool, it allows a user to create tables via a drag and drop interface. It has features which include the ability for the user to derive new variables, add statistical calculations to tables, transform data dynamically, and to drill down on selected cells in a table to view the contributing unit records. These features give the user great flexibility for creating tables without the need for specialist skills such as SQL. SuperCROSS has been used in the ONS for many years, and is the current tabulation tool for BRU, Labour Force Survey, and Census.

[4] http://neon.vb.cbs.nl/casc

1.4 Proposed Solution

The proposed solution was based on the use of Tau-Argus in batch mode, integrated with the familiar SuperCROSS tabulation environment. SuperCROSS users can choose a pre-defined rule to be applied to a table, and Tau-Argus is invoked automatically. Once processed, the results would be passed back to the SuperCROSS environment. The key to this approach is that only a small number of SuperCROSS users would need to have some level of understanding of SDC. The SDC rules to be applied are stored in template files, which are specified and quality assured by a statistical disclosure control specialist.

Figure 1 shows the high level workflow of the Tau-Argus and SuperCROSS integration.

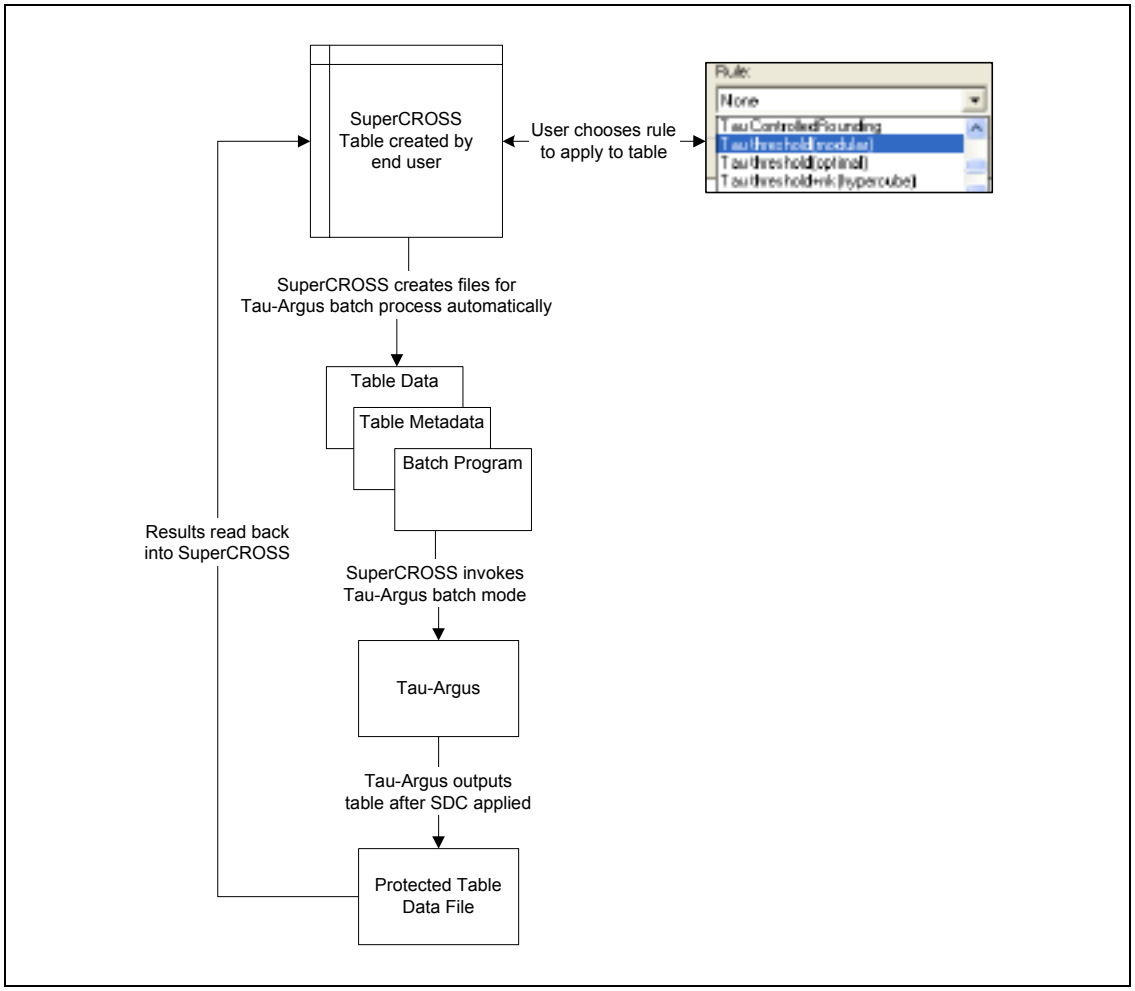


Figure 1. Tau-Argus and SuperCROSS workflow

2 Detail of Solution

2.1 How does Tau-Argus work in Batch Mode

Tau-Argus can be used through its own Graphical User Interface (GUI), or through a batch mode. Batches are composed of a set of instructions which allows Tau-Argus to identify the data source, metadata, disclosure rules to apply, and any transformation to the data that is required.

Through the batch file it is also possible to set parameters to create output files after disclosure rules are applied, as well as an HTML report file with summary information relating to the impact of the SDC method on the outputs produced. Figure 2 shows an example of a batch file created for Tau-Argus. It specifies the location of the table data and metadata, the specification of the table, the disclosure rules to apply, and the location of the output.

```
<OPENTABLEDATA>   "D:\Tau-Argus\temp_tauinput.tab"
<OPENMETADATA>    "D:\Tau-Argus\temp_taumetadata.rda"
<SPECIFYTABLE>    "var1""var2""var3"|"resp_var1"|"resp_var1"|"resp_var1"
<SAFETYRULE>      FREQ(3,30)|
<READTABLE>
<SUPPRESS>        RND(1,5,0,0,20,0,0,2)
<WRITETABLE>      (1,1,1,"D:\Tau-Argus\temp.csv")
```

Figure 2. Example of Tau-Argus batch file

For further information on each of the elements in the batch file, refer to the Tau-Argus manual available on the CASC website.

2.2 Integration of Tau-Argus with SuperCROSS

Figure 1 shows the workflow of how a user interacts with SuperCROSS.

a. Firstly, a rule template is defined by a SDC specialist. The user need not understand the rules in detail, nor how they are specified.

b. Next, the user creates a table in SuperCROSS, and chooses which rule they want to apply from a drop down menu.

c. SuperCROSS tabulates the table, and then generates the batch file, and table data and metadata files required for Tau-Argus. SuperCROSS then invokes Tau-Argus in batch mode, and the table is processed.

d. Once the Tau-Argus batch is finished, the results are read back into SuperCROSS and displayed to the user.

Figure 3 shows how a rule template file is defined for SuperCROSS. It contains the detail of the rule to apply when chosen by the user.

```
<SAFETYRULE>      FREQ(3,15)|
<READTABLE> 1
<SUPPRESS>        RND(1,10,0,0,20,0,0,3)
<WRITETABLE> (1,3,1,"C:\TEMP\report_tauoutput.csv")
```

Figure 3. Example Rule Template file for Controlled Rounding

This file is then referred to in a SuperCROSS configuration file (confid.ini), along with the location of the Tau-Argus executable. An example of the SuperCROSS configuration file is shown in Figure 4.

```
[Tau-Argus]
Tau-Argus Location=\\TauServer\TauArgus
Tau-Argus Exe=TauArgus.exe
Tau-Argus Log=C:\Program Files\TauArgus\sxtaulog.txt

[Rules]
Tau ControlledRounding=Tau:ControlledRounding

[ControlledRounding]
ARBFile=\\TauServer\RuleTemplates\TauControlledRounding.arb
TopN=0
```

Figure 4. SuperCROSS Configuration File

When the user tabulates a table in SuperCROSS with a chosen rule, the table data and metadata is saved to appropriately formatted text files, and the rule template file is copied and the location of the saved data is inserted. All of these operations are hidden from the SuperCROSS user and occur automatically.

2.3 Controlled Rounding Example

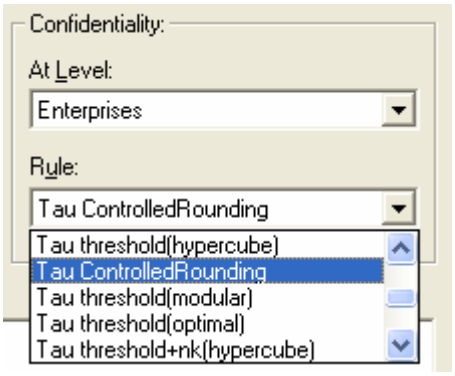
This section gives an example, not using standard ONS rules, of applying controlled rounding to a table within SuperCROSS. Figure 5 shows the steps required to apply the rule to a table.

Example Table One
Country/GOR by Status Type
for Number of Enterprises

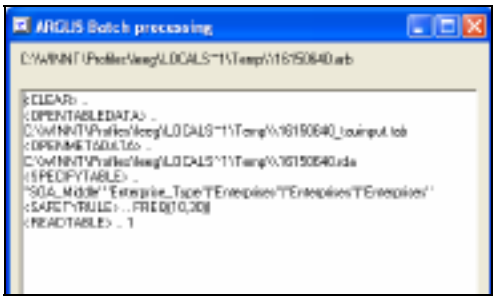
	Type 1	Type 2	Type 3	Type 4	Total
England	**36649**	**5042**	**52**	**12168**	**53911**
South West	1279	204	5	655	2143
West Midlands	4835	642	2	1755	7234
Yorkshire	3816	474	1	1234	5525
East Midlands	3177	417	1	966	4561
North West	4383	535	2	1222	6142
East of England	3990	613	7	1223	5833
South East	6161	650	17	1725	8553
London	5917	874	9	1915	8715
North East	3091	633	8	1473	5205
Wales	**1378**	**299**	**8**	**811**	**2496**
Total	***38027***	***5341***	***60***	***12979***	***56407***

Sample data only

a. Raw Table



b. Choose Rule in SuperCROSS



c. Tau-Argus Batch Process Runs

Example Table One
Country/GOR by Status Type
for Number of Enterprises

	Type 1	Type 2	Type 3	Type 4	Total
England	**36650**	**5040**	**50**	**12170**	**53910**
South West	1280	200	0	660	2140
West Midlands	4830	640	0	1760	7230
Yorkshire	3820	480	0	1230	5530
East Midlands	3180	420	0	960	4560
North West	4380	540	0	1220	6140
East of England	3990	610	10	1220	5830
South East	6160	650	20	1720	8550
London	5920	870	10	1920	8720
North East	3090	630	10	1480	5210
Wales	**1380**	**300**	**10**	**810**	**2500**
Total	***38030***	***5340***	***60***	***12980***	***56410***

Sample data only

d. Protected Table

Figure 5. Example of Controlled Rounding

a. The table is created by the user within SuperCROSS.

b. To apply disclosure control, the user can choose a pre-defined rule from a drop down menu in SuperCROSS. In this case, the rule is called "Tau ControlledRounding". The user does not need to know, or understand, the specifics of how the actual rule works, although basic knowledge of the method is required. For reference, the rule used in this example was as follows:

<SAFETYRULE> FREQ(10,30)|
<READTABLE> 1
<SUPPRESS> RND(1,10,0,0,20,0,0,3)

In this example, a threshold rule is initially applied to the table. All cells with fewer than 10 contributors are defined as disclosive. After the table is checked for

additivity and any undefined hierarchical levels added, controlled rounding is applied to base 10.

c. After the SuperCROSS tabulation process finishes, the table data is passed to Tau-Argus along with the batch file and other required information. Tau-Argus is invoked and the disclosure control rule selected is then applied.

d. The protected table is now displayed in the SuperCROSS interface. The table can then be saved to the desired format for dissemination.

2.4 Example of Suppression

The original table in Figure 5 shows the count of the number of enterprises. If you change the definition of the table to sum turnover instead, it may be more appropriate to apply a different rule to the table. In this example, a suppression rule, not the ONS standard, is applied, which hides table cells rather than rounding them. In this case, the rule applied is as follows:

```
<SAFETYRULE>    P(15,100,1)| FREQ(3,30)
<READTABLE>
<SUPPRESS>      MOD(1)
```

This is a multi stage process. First, a threshold rule of 3 at a safety range of 30% is applied. In addition to this, the p% rule states that if the sum of all the contributors to a cell, excluding the top 2, is greater than 15% of the total value of the cell then the cell is not disclosive. This is applied to all cells in the table, with those failing either rule being suppressed. Finally, secondary suppression is applied using the modular method. Figure 6 shows the table results as the user would see in SuperCROSS, after the disclosure rules have been applied.

Example Table Two
Country/GOR by Status Type
for Turnover

	Type 1	Type 2	Type 3	Type 4	Total
England	**1395845**	**30924**	**13014**	**33152**	**1472934**
South West	11104	6355	284	808	18552
West Midlands	67683	3004	..C	..C	76009
Yorkshire	128084	1752	..C	..C	132151
East Midlands	41731	2606	..C	..C	46221
North West	72975	4692	..C	..C	79644
East of England	125157	3451	..C	..C	131367
South East	719698	..C	..C	11996	740686
London	177595	4808	578	6001	188982
North East	51818	..C	..C	2914	59322
Wales	**13713**	**121**	**55**	**953**	**14843**
Total	***1409558***	***31045***	***13069***	***34105***	***1487777***

Sample data only

Figure 6. Example of suppression rule

3 Business Benefits

The integration of the two tools, SuperCROSS and Tau-Argus, has advantages for both the producer of tabular statistics, and the user of the outputs. Tau-Argus has been shown to be a powerful disclosure control tool, enabling both suppression and controlled rounding to be carried out. The majority of producers of tables in a business area will not require any knowledge of Statistical Disclosure Control principles beyond the basics, as the rules required are pre-configured and quality assured by an SDC specialist. This means that a powerful disclosure control tool can be made available to users with minimum disruption to the office.

In general terms, any process which joins together important operations in the production of tables should be beneficial. A smoother output delivery process can be put in place and the users do not have to switch between different tools.

As the disclosure control procedure is now fully automated, this will result in savings in terms of time, particularly as users do not have to manually process data via Excel. There should also be quality improvements in the outputs (the tables have greater utility), with less likelihood of errors. This will assist in maintaining a positive reputation of the data supplier. Moreover, the business area could process more ad-hoc requests in the same amount of time.

The current implementation could easily be adapted for other business areas, with only minimal changes required to the rule templates for SuperCROSS, depending on the nature of the data.

4 Issues

The benefits to the business in terms of robustness and reduction in risk are obvious. However, the solution is not without its limitations, and these are predominately caused by the underlying methodology. There are also some disadvantages that can be found when comparing perceived and actual complexity of the tools.

The problem of the user creating the best design of a table that is statistically meaningful to data consumers, and to both of the tools involved, represents a major challenge to be overcome. SuperCROSS users have the flexibility to create tables in whichever fashion desired. All of these tables, prior to disclosure control, are valid. However, the tables do not necessarily have meaning when it comes to statistical disclosure control.

SuperCROSS users are able to define tables which span across multiple statistical unities (for example households and persons). This is an example of linking two or more tables from the same microdata, and there is currently no rule in Tau-Argus

which solves this problem. Linked tables at the macrodata level are also not currently supported; tables which share one or more common variables cannot be protected simultaneously.

It is possible in SuperCROSS to create tables based on hierarchical variables, for example geography or industrial classification. SuperCROSS allows users to include any combination of values from any of the levels in the hierarchy, and it does not have to be complete. Although in certain instances Tau-Argus can calculate the missing hierarchical values, this must be used with great caution and it is recommended to supply a complete hierarchical structure.

The size of a table created in SuperCROSS can sometimes pose a problem for Tau-Argus in terms of the time required to solve the problem. It is often possible to overcome this by fine-tuning the rule applied asking Tau-Argus to find a feasible rather than an optimal solution, and then using table partitioning rules where available.

Many of the other features in SuperCROSS, such as grouping items within recodes, which are very useful in general for tabulation and table design purposes, also cause issues for the current methodology of Tau-Argus. Workarounds have been provided to BRU which do result in the desired table output format. Future development could overcome some of these limitations.

SuperWEB, also from Space-Time Research, is a browser based interface for self-service tabulation, accessed by the data consumer. This differs from SuperCROSS, which is typically used by the business area to generate pre-defined tables for publication. Given the nature of the SDC rules in Tau-Argus, it would not be recommended to apply this solution to a web-based, dynamic interface, either for data suppliers, or users of that data. Users of SuperWEB could not be expected to understand, nor appreciate, the limitations involved with attempting to apply SDC on demand through the web.

Aside from the methodological issues, any implementation of the SuperCROSS and Tau-Argus solution should not overlook the resources and effort required for installation and configuration of the tools. Tau-Argus requires mathematical solver software to be installed for some of the SDC rules, which adds a further layer of administration.

5 Conclusion

The integration of SuperCROSS and Tau-Argus will be beneficial in many ways to the ONS. The main benefit is that a business area can publish more, with better quality (i.e. data utility), and faster response times.

It is strongly advised that with any new implementation of this solution, the organisation invests in understanding the technology of the tools involved, and the administration that is required for all components. This should not be underestimated as Tau-Argus, SuperCROSS, and the mathematical solver of choice are from three different vendors, and all have their own method of dealing with issues.

Good user knowledge of SuperCROSS is required, but with little extra to learn. Once both products are installed the statistical disclosure business processes should flow more smoothly than previously experienced.

6 Acknowledgements

This paper was written with the cooperation and support of the Business Register Unit at the ONS.

The anonymisation of the CVTS2 and income tax dataset

An approach using R-Package *"sdcMicro"*

Bernhard Meindl*, Matthias Templ**

* Department of Methodology, Statistics Austria, Guglgasse 13, 1110 Vienna, Austria. (bernhard.meindl@statistik.gv.at) and

** Department of Statistics and Probability Theory, Vienna University of Technology, Wiedner Hauptstr. 8-10, 1040 Vienna, Austria. (templ@statistik.tuwien.ac.at)

Abstract. The demand for microdata for research and teaching purposes gets higher and higher. To overcome this fact we not only provide secure workstations where researchers can deal with "original" data but also provide more and more anonymised microdata. When using flexible software tools for anonymisation we can provide high quality anonymised data which can be generated in less time. In this paper we show such anonymisation processes on two different data set, the continuing vocational training survey (CVTS2) and the income tax data from 2005.

1 Introduction

Since the demand to publish micro data for researchers grows, it is necessary to take actions that published data will not lead to correct re-identification of either an individual or an enterprise. To assure the anonymity of micro data, different anonymisation methods such as global recoding, local suppression or microaggregation are used. Special emphasis should be placed on trying to keep the (multivariate) structure of the data and changing the original data as little as possible while guaranteeing a low individual risk of re-identification. With new and modern software it is possible to use anonymisation methods very effective.

1.1 Terms of Use

In Statistics Austria we provide two different variants of anonymised data sets. An SDS (standardised data set) is basically an anonymised micro data set for research and teaching purposes whereas an ADS (task-related dataset) is generated for special research purposes for only one research project or research collaboration with other organisations. However, users have to agree on the terms of use for both variants of anonymised data. From the anonymisation point of view, a SDS is somewhat between a public use file and a scientific use file. The complete terms of

use can be found at: http://www.statistik.at/web_de/services/mikrodaten_fuer_forschung_und_lehre/datenangebot/standardisierte_datensaetze_sds/index.html.

2 Software used

To protect the CVTS data and the Austrian income tax data we used the R-package sdcMicro [17]. R [12] is an open-source, free-available, high-level programming language for statistical computing and graphics. The main advantages of package sdcMicro are the reproducibility of any results, the flexible usage of the package, the import/export facilities, the richness of methods implemented, the graphical power for comparison of original data and perturbed data, the easy usage of the package and the fast calculation of results. Furthermore, one can easily use the whole power of R since the package runs in R.

Since all the results from anonymisation can be reproduced by running a script or parts of the script with the code, a user can do the anonymisation approach very flexible and in an explorative manner and can interactively communicate with all the objects in the workspace of R. It's a bit like playing with the data instead of writing a "batch file" which then must be processed.

3 CVTS2 Data

It was the objective to generate a SDS for the CVTS2 dataset (continuing vocational training survey). The goal of this survey is to gain information on internal measures enterprises have taken and (partly) payed for on advanced vocational training for employees. The raw data consisted of 2613 enterprises for which a total of 197 variables has been recorded. It has to be noted that the sample of the CVTS survey is chosen in a way that only enterprises with at least 9 employees may be drawn into the sample. Due to imputation and plausibility checks we observed some abnormalities such as ratios being greater than 100%, most of which can be explained though. Further information on the CVTS2 data can be found at Statistics Austria's webpage at: http://www.statistik.at/web_de/statistiken/bildung_und_kultur/erwachsenenbildung_weiterbildung_lebenslanges_lernen/index.html.

The difficulty in generating a SDS for this data was the large number of categorical variables and the fact that any combination of these variables might be used by an attacker to correctly identify an enterprise. Thus it was necessary to assess different scenarios of statistical disclosure control by considering different subsets of the available categorical variables as key variables in order to receive an impression of the disclosure and re-identification risk.

Another important point is that we decided to calculate and publish ratios for

most of the numeric variables. Doing so, absolute values can not be recycled anymore. This approach was described as well by Brandt and Hafner [2]. However, most multivariate statistical methods are invariant against transformations and the same results can be obtained as for the original values. But, of course, certain aggregations of the data are not suitable for the transformed subset of the data.

3.1 Anonymisation of the CVTS2 data

Before actually starting to apply anonymisation methods, 29 variables which were either direct identifiers or were including non relevant information were deleted from the data set. Then ratios for most numeric variables in the dataset were calculated as already explained above. Furthermore, we recoded several variables such as the *economic classification of the enterprise* or the *number of employees* into broader categories. All operations have been done using `R` and package `sdcMicro` which make the entire anonymisation procedure flexible, fast and reproducible.

We started by comparing different scenarios and several combinations of possible key variables by having a look at the corresponding individual risks for re-identification [8] as well as the number of unique combinations of the characteristics in the key variables. It is in fact very convenient to compare different scenarios using `sdcMicro` because the user only has to specify the desired key-variables and re-run the code. After comparing several possibilities we decided to use the following key variables which are listed below:

- **economic classification of the enterprise:** 10 categories
- **number of employees:** 4 categories
- **generated revenues for vocational training:** 2 categories
- **expenses for vocational training:** 2 categories

It should be noted that all of these key variables have been already changed in an explorative manner by global recoding techniques or have been generated from other variables.

Then we looked at the number of unique combinations of the key variables, the number of observations with a given combination of the key variables that occurs only twice as well as the individual risk for re-identification. We observed 58 unique observations and 50 observations whose combination of values of the key variables occurred exactly two times.

`sdcMicro` provides a method to plot the individual risk and to interactively change the threshold value similar to the μ-Argus [10] plot method. This is helpful

to determine suitable threshold values for the local suppression methods that need to be applied to the key variables. Figure 1 shows the individual risk for re-identification along with its empirical distribution function for the original, non perturbated data.

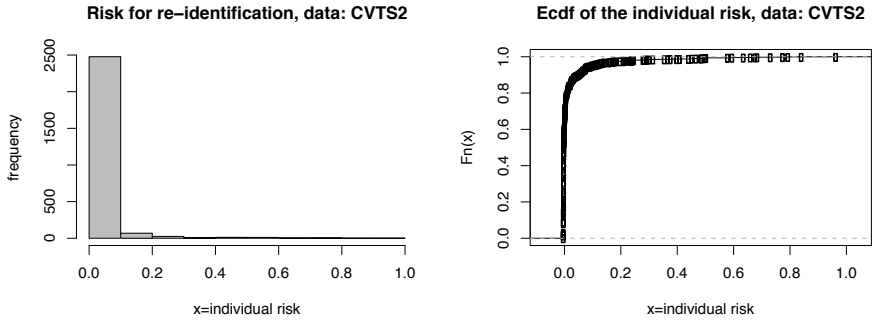


Figure 1: individual risk (left) and empirical distribution (right) in original data.

We want to provide k-anonymity ([15, 13, 14]) for this dataset. This means that for any combination of key variables at least k observations must exist in the data set sharing that combination. The `sdcMicro` function `localSupp()` can be used to suppress values in the key variables. We find 3-anonymity in combination with the other anonymisation methods applied to be a sufficient for publishing for this dataset since the CVTS2 data are not as interesting from an attacker's point of view as for example the income tax data described later.

We started with the variable *beiträge* and set the threshold value to 0.25. This resulted in suppressing 19 values in this key variable. Afterwards, the risk of re-identification of enterprises is plotted again and a new threshold values is determined. Using the threshold value of 0.167, `localSupp()` was applied to the key variable *einnahmen*. This led to a suppression of 25 values in this variable. The same procedure was used to suppress values in the key variable *a299tot*. After choosing a suitable threshold value (0.104) and applying `localSupp()` we note that 12 values were suppressed in this key variable.

We then observed that there were still enterprises left that had unique combination of the key variables or a combination which only occurred two times in the dataset. Thus, we manually set the values of the variable *beiträge* for those enterprises that had a unique combination of key variables to missing. As a result, 14 suppressions had to be done. Additionally, the variable *einnahmen* was set to missing for all enterprises that had a combination of the key variables that occurred only twice. This resulted in suppressing one additional value. Summarizing this process, a total of 33 values had to be suppressed in variable *beiträge*, 26 variables had to be

suppressed in variable *einnahmen* and 12 variables had to be suppressed in variable *a299tot*. After these suppressions, each combination of the values in the key variables occurs at least three times and we note that the goal of 3-anonymity is reached.

In Figure 2 the individual risk for re-identification is plotted along with its empirical distribution function after using local suppression for anonymisation. It is obvious, that we achieved a clear reduction in the individual risk of re-identification compared to Figure 1 as the different scaling of the x-axis indicates.

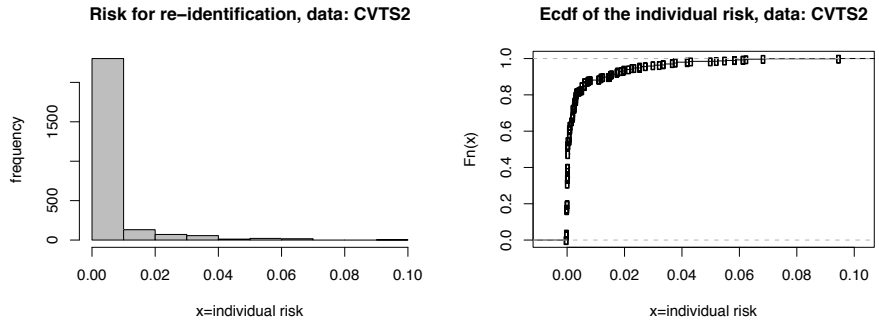


Figure 2: individual risk (left) and empirical distribution (right) in anonymised data.

After dealing with categorical variables and indirect identifiers, we took additional precautions by microaggregating (references on microaggregation can be found in [1], [7], [4], [3], [5]) the available numeric variables. This effectively means that for each numeric variable the values are grouped by a proximity measure into small groups consisting of 4 values. Then the values within each group are averaged and the aggregate are finally released in the anonymised dataset. This assures that each numeric value occurs at least 4 times in the anonymised dataset. In this case we have used a version of individual ranking [4] which can also be applied on data with missing values.

4 Austrian Income Tax Data

The Austrian income dataset for 2005 consists of 5.919.739 rows. The data have already been aggregated from pay slip level to person-level. Thus, exactly one row exists in the raw data for each person that has been liable to pay taxes in 2005. The dataset consists of a total of 74 variables, however, only 17 variables have been included in the published micro data SDS file.

The categorical variables give information about the social status, sex, the federal state and the age of the individual as well as of the number of pay slips considered,

the number of weeks employed, whether the person was employed part- or fulltime and the economic classification of the enterprise the individual was predominantly employed at. The quantitative variables included give information for example about the gross wages, social security contributions or information about the amount of other payments a given person has generated.

Further information on the income tax data can be found at the Statistics Austria web page located at: `http://www.statistik.at/web_de/statistiken/oeffentliche_finanzen_und_steuern/_steuerstatistiken/lohnsteuerstatistik/index.html`

The final anonymised dataset is available for download at the following web page: `http://www.statistik.at/web_de/services/mikrodaten_fuer_forschung_und_lehre/datenangebot/standardisierte_datensaetze_sds/index.html#index8`

4.1 Anonymisation of the Austrian Income Tax Data

We will now describe the anonymisation methods applied to the raw data set. Our anonymisation approach is quite different than the one used to generate the german public and scientific use file, respectively of income tax data. More on the german contribution in this field can be found on [11].

First, many quasi direct identifiers and variables that should not be included in the final anonymised data set have been deleted from the raw data. Among the variables deleted are the tax-id which is unique for each person as well as regional information on the individuals such as the exact address. Then, the anonymisation methods described later are applied to the resulting dataset which consisted of all in all 17 variables. Eight variables are scaled categorically while eight variables are quantitative. Furthermore, the sampling weight resulting from drawing a subset of the original data is attached to the dataset as an additional variable.

In a second step, a 1% random sample stratified by age, sex and federal states was drawn from the raw dataset. This resulted in a dataset including 59.279 individuals. This should be seen as a quite effective anonymisation method because even if an attacker manages generate a one to one match from a reference file to an individual from the sample using key variables, he cannot be sure if the possibly identified individual has even been drawn to the sample.

After generating a subset from the raw data it was necessary to define key variables. As already mentioned before, a total of 8 categorical variables was included in the data set. We used `sdcMicro` to compare different scenarios and several combinations of key variables by having a look at individual risks for re-identification and the number of unique combinations of the characteristics in the key variables.

After comparing several possibilities, we considered the following five key variables:

- **social status:** 7 categories
- **federal state:** 10 categories
- **sex:** 2 categories
- **age classes:** 8 categories
- **economic classification of the enterprise:** 12 categories

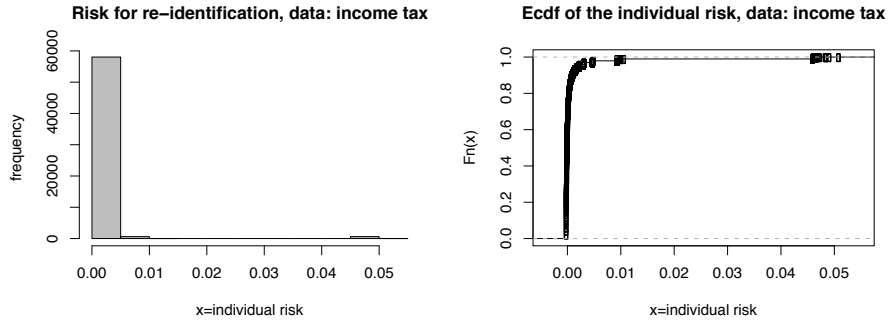


Figure 3: individual risk (left) and empirical distribution (right) in original data.

After deciding on the key variables, the individual risk for re-identification was calculated for all the key variables defined above. It turned out in this explorative approach that we should recode some variables in order to reduce the re-identification risk. Therefore, we recoded variable *social status* as well as the the employment state into less categories. It turned out that 623 observations had a unique combination of characteristics of the key variables and 604 individuals had a combination of values of the key variables that only existed twice. In Figure 3 the individual risk of re-identification is plotted for the data using the five key variables discussed above.

In the next step, values of certain key variables are set to missing for individuals with high risk of re-identification. To assess the individual risk and to find a suitable threshold value which determines the number of suppressions we used again the interactive plot method of `sdcMicro` to assess the individual risk of re-identification and to interactively change the threshold value and look at the resulting re-identification rates conditional on the chosen threshold value.

After choosing a suitable threshold we used the function `localSupp()` to suppress data in the variable *economic classification of the enterprise* the individuals

were employed at with a quite low threshold of 0.01. As a result, a total of 631 values (=≈ 0.01%) for this key variable was set to missing. As a result we obtain that the number of individuals that are unique in the dataset drops to 85 and the number of individuals with a combination of values in the key variables that occurs twice drops down to 256.

After this step, the relative risk is plotted again interactively in order to find a suitable threshold value for local suppression of values in the key variable *social status*. Applying the local suppression method with a threshold value of 0.01 results in no observation that has a unique combination of values in the key variables after setting 93 values in the variable *social status* to missing. However, there are still 154 observation left that have a combination of key values that occurs only twice.

We find that 3-anonymity in addition to microaggregation of numeric variables and the fact that the released data set itself is just a 1% random sample from the population, is adequate for this critical dataset. In order to guarantee 3-anonymity we had to set additional values in the key variables to missing. As already described we decide on a threshold value (0.01) and apply the function `localSupp()` to the key variable *"federal state"*. By setting 154 values of this variable to missing, 3-anonymity for this dataset is obtained.

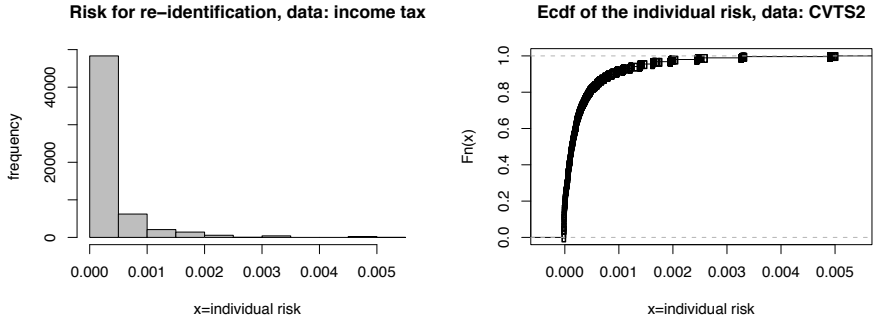


Figure 4: individual risk (left) and empirical distribution (right) in anonymised data.

In Figure 4 the individual risk after locally suppressing values in the key variables is plotted. The graph looks similar to Figure 3, however one should note the different scaling on the x-axis.

After dealing with categorical variables and indirect identifiers, we additionally microaggregated all numeric variables available in the dataset. As for the CVTS2 data we used a version of individual ranking for the microaggregation procedure

which guarantees that each numeric value exists at least 4 times in the SDS.

5 Conclusion

In both anonymised data sets the re-identification of individuals is very hard or even impossible. For example, each combination of extremely identifiers [9] such as the regional variable occurs more than 1926 times for the income tax data set which itself is only a 1% sample from the original data. Most of the proposed anonymisation rules of the handbook on SDC [9] are not considered since our data are too small to fit this criteria. Nevertheless, global recording, local suppression and microaggregation were applied to achieve sufficient anonymisation of the data, i.e. to provide both 3-anonymity and low re-identification risk. In the other hand, the perturbed data are of high quality and have nearly the same (multivariate) structure as the original data since only few selective recodings and local suppressions on categorical variables were made. It is also well known, that microaggregation does not destroy the (multivatiate) structure of the data (see e.g. in [16] or [6]). The software used had allowed the anonymisation in less time and in an explorative way.

References

[1] N. Anwar. Micro-aggregation - the small aggregates method. In *Internal report*. Luxembourg: Eurostat, 1993.

[2] M. Brandt and H. Hafner. Leitfaden zur Anonymisierung für die Erstellung eines Campus-Files aus den Einzeldaten der zweiten europäischen Erhebung zur beruflichen Weiterbildung. Technical report, Statistisches Bundesamt, Hessisches Statistisches Landesamt., 2007. Nr. 5, 10/2005.

[3] D. Defays and Anwar M.N. Masking microdata using micro-aggregation. *Journal of Official Statistics*, 14(4):449–461, 1998.

[4] D. Defays and P. Nanopoulos. Panels of enterprises and confidentiality: the small aggregates method. In *Proceedings of the 1992 Symposium on Design and Analysis of Longitudinal Surveys*, pages 195–204. Statistics Canada, Ottawa, 1993.

[5] J. Domingo-Ferrer and J.M. Mateo-Sanz. Practical data-oriented microaggregation for statistical disclosure control. *IEEE Trans. on Knowledge and Data Engineering*, 14(1):189–201, 2002.

[6] J. Domingo-Ferrer, J.M. Mateo-Sanz, A. Oganian, and A. Torres. On the security of microaggregation with individual ranking: analytical attacks. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):477–492, 2002.

[7] M. Elliot, A. Hundepool, E.S. Nordholt, J-L. Tambay, and T. Wende. Glossary on statistical disclosure control, 2005.

[8] L. Franconi and S. Polettini. Individual risk estimation in μ-ARGUS: a review. In *Privacy in Statistical Databases. Lecture Notes in Computer Science. Springer*, pages 262–272, 2004.

[9] A. Hundepool, J. Domingo-Ferrer, L. Franconi, S. Giessing, R. Lenz, J. Longhurst, E. Schulte Nordholt, G. Seri, and P. De Wolf. *Handbook on Statistical Disclosure Control*, 2007.

[10] A. Hundepool, A. Van deWetering, Ramaswamy R., L. Franconi, A. Capobianchi, P-P. DeWolf, J. Domingo-Ferrer, V. Torra, R. Brand, and S. Giessing. μ-argus version 4.1 software and users manual, 2006.

[11] J. Merz, D. Vorgrimler, and M. Zwick. De facto anonymised microdata file on income tax statistics 1998. Technical report, SRI Intl. Tech. Rep., 2005. Nr. 5, 10/2005.

[12] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2007. ISBN 3-900051-07-0.

[13] P. Samarati. Protecting respondents' identities in microdata release. *IEEE Transactions on Knowledge and Data Engineering*, 13(6):1010–1027, 2001.

[14] P. Samarati. Achieving k-anonymity privacy protection using generalization and suppression. *Intl. Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):571–588, 2002.

[15] P. Samarati and L. Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical report, SRI Intl. Tech. Rep., 1998.

[16] M. Templ. Software development for SDC in R. In *Privacy in Statistical Databases. Lecture Notes in Computer Science. Springer*, pages 347–359, 2006.

[17] M. Templ. *sdcMicro. Manual and Package.* Statistics Austria and Vienna University of Technology, Vienna, Austria, 2007. *http://cran.r-project.org/src/contrib/Descriptions/sdcMicro.html.*

sdcMicro: a new flexible R-package for the generation of anonymised microdata: Design issues and new methods

Matthias Templ*

* Department of Methodology, Statistics Austria, Guglgasse 13, 1110 Vienna, Austria. (matthias.templ@statistik.gv.at) and

* Department of Statistics and Probability Theory, Vienna University of Technology, Wiedner Hauptstr. 8-10, 1040 Vienna, Austria. (templ@tuwien.ac.at)

Abstract. Data protection specialists need flexible software tools for the exploratory use of protection methods to generate high quality confidential data. Microdata protection is widely used and is often the only possible way to provide data to both researchers and users. In this paper we present a methodological and computational framework for the generation of anonymised microdata and give insights to the developed **R**-package `sdcMicro`. This package may become the standard software for microdata protection since it is very flexible, easy to use and contains all popular methods, plus some new ones. The package can also be used for comparison of methods and of original versus perturbed data not only by measuring information loss but also by various comparison plots.

1 Motivation

Nowadays there are various methods for dealing with confidential microdata developed to make such data accessible to researchers and users. Roughly speaking, these methods can be clustered into 3 categories:

(i) Remote Access based on the idea of hidden data and checked outputs.

(ii) Remote Access based on the idea of a free view of the data without the possibility to download the data.

(iii) perturbed microdata publicity as *public* or *scientific use files* or different variants between public and scientific use files.

In the following there is a short motivation of point (iii) since many researchers do not see a future in this point (iii) and therefore do not see the necessarity for a (new) software on microdata protection.

Regarding (i) researchers can apply models on data which can not be seen or where data and particularly outliers are perturbed [12, see e.g. in] but they can apply models. This approach is often called *confidential preserving model servers*. But it is the results of a models which are the objects of interest not the underlying data [22]. Considering a *Remote Execution* framework these results are checked by the census staff and, for example, in case of regression analysis synthetic residuals are often provided. But then, you see synthetic residuals based on residuals from really worst estimates because of using least squares regression on data which, naturally, includes outliers. Furthermore, you would never have the ability to find a good model without applying the whole range of diagnostic tools on robust estimates. This is in contradiction to [11] and others (e.g. [22]) because for such robust estimates it is improbably to provide synthetic residuals (because you always see large or heavy perturbed residuals/outliers), for example, and there is no method available for the detection of leverage points. It is also well known in robust statistics literature that you need robust methods for the detection of outliers which are essential and which cannot be provided. We can conclude that model servers are not compatible with a modern statistical world unless the underlying data is of a multivariate normal distribution or can be transformed into one which turns out to be unrealistic for real complex data.

In the second approach (ii) researchers may look at the data and can choose a suitable method for analysing the data, but it is not possible to download the data in any way (for applications in (ii) see e.g. [14] or [2]). Remote access can only partially be applied in some countries depending on the discipline from which the data originates. This is due to different legacy for data coming from different disciplines. In Austria, for example, only microdata protection is reasonable for official statistics because of the legal situation which prohibits a view on "original" data.

In (iii) users and researchers have access to perturbed data. In this paper we will show that such a perturbation can be easily performed with newly developed and extremely flexible software package by minimizing information loss and re-identification.

In addition to that, we will also concentrate on the design of this new package for microdata protection and illustrate the open source philosophy of this project.

2 Using R for SDC

R [20] is an open source high-level statistical computing environment subjected to the General Public License and therefore freely available and expendable. Furthermore, **R** has become the standard statistical software and thousands of people are involved in the development of **R** both at universities and companies. More than 1000 add-on packages have been built in the last years.

In the following the usefulness of **R** related to SDC is presented. The most important and most essential feature of a software for SDC is the reproducibility of results. The most effective way of anonymising microdata is by doing the anonymisation steps in an explorative and in some sense iterative way, since we can apply various methods on various variables with different parameters producing different effects on the data and while looking for sufficient anonymisation of the data with respect to low information loss. Therefore, we must have the ability to reproduce any step of the anonymisation process easily. This is fulfilled by running a script with all the commands included. It is then easy to change and adapt this script and get all the new results "in real time". Of course, this is fulfilled by many software tools, but the real advantage of using **R** is that we can interactively "play" with the data, i.e. we have access to all the objects in the workspace of **R** any time and can change, display or apply operations on it on the fly. This is very useful during the anonymisation of data and a quite different concept than to write a "batch file" which can then be executed.

In addition to that, we recommend the use of a powerful easy-to-grasp visualization tools to see the effect of the anonymisation on the data instead of computing various measures of information loss (see e.g. in [23]).

But dynamical reports can also be used for documenting the anonymisation when using **R** in combination with `Sweave` ([17]). Reports can then be generated very effectively and quickly for particular steps in the production process.

Since many different possible data formats are used by users it is very comfortable to have the opportunity to import and export data formats from data base software and statistical software like `SPSS`, `SAS`, `Stata`, `DBF`, `Excel`, `ASCII`, and many more. Sometimes it is also useful to run the anonymisation tools via batch mode, which can be easily derived with package `sdcMicro` [24].

It is also very important to provide a software for SDC which is platform independent and works on all common operating systems.

All these things are supported when using **R**.

3 Design Goals of Package `sdcMicro`

The advantages of an object-oriented programming language become apparent in the `scdMicro` package. However, the methods and class approach from **R** is not a class-oriented programming language (like `C` or `Java`) but a richer one as it is a function- and class-oriented programming language.

In **R** everything is an object and every object is related to a specific class. The class of an object determines how it will be treated and generic functions perform either a task or an action on its arguments specific to the class of the argument itself. The class mechanism offers the programmer the facility of designing and writing generic functions for special purposes which is extensively used in *sdcMicro*. Nearly

all functions, e.g. the one for the individual risk methodology or the frequency calculation, produce objects from a certain class. Different *print*, *summary* and *plot* methods are provided for each of these objects depending on their class. `plot(ir1)` produces a completely different result than `plot(fc1)` assuming that the objects `ir1` and `fc1` are objects from different classes, i.e. resulting from different functions in package *sdcMicro*.
This object-oriented approach allows a simple usage of the package for any user, independently of the proficiency in **R**. Furthermore, users can try out different methods with different parameters and they can easily compare the methods with the implemented summary and plot methods.

Note that no metadata management needs to be done by the user, even not after importing the data into **R**. You can apply the methods directly on your data sets or on objects for certain classes. At first you must only determine which of the variables are the key variables, the weight vector and the numerical ones.
An online documentation is included in the package containing all explanations on all input and output parameters. Furthermore, various examples are included for each of the functions. These examples can be easily executed by the users.

To be able to deal with large data sets extensive computational calculation steps are implemented in `C++` and included in **R** via the `R/C++` interface. The calculation of the frequency counts is one of the most critical calculation steps regarding to the computation time. Figure (1) shows the calculation time carried out by an one year old personal computer[1] with Windows XP operating system. The computation time can not be shown in this figure because it turns out to be too large when using loops in **R** for the calculation of the frequency counts because it is too large. Only when using functions in **R**, which are said to be implemented in `C` or `Fortran`, it is possible to calculate the frequency counts with up to 6 key variables by using this relatively small data set with only 4000 observations. This is not possible for the μ-Argus system ([15], version 4.1.0) which runs out of memory with 5 or more key variables. The developed `R/C++` can also handle a very large number of key variables[2] in reasonable time even for much larger data sets than the μ-Argus test data set. In Figure (1) you can easily see that the computation time is always less than 1 second for this small data set and it is also low for larger data sets.

4 Implemented Methods and Data

Methods like global recoding, local suppression and the individual risk methodology (see e.g. in [10]), top- and bottom-coding, PRAM ([16]), more than 10 microaggregation methods (mdav, pca-methods, robustified pca-methods with fast algorithms

[1] Intel x86 based system with 3Ghz and 1 GB memory
[2] to choose such a large number of key variables is, of course, not always meaningful.

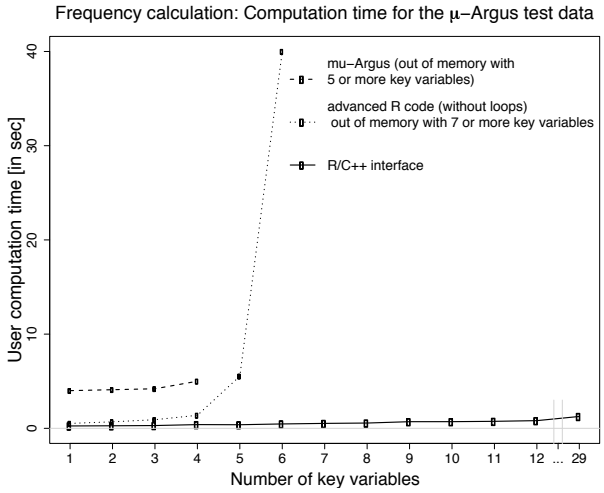


Figure 1: Frequency count computation time for the μ-Argus test data set with 4000 observations.

and clustering, individual ranking, methods based on robust Mahalanobis distances, ...) (more information on these methods and on microaggregation can be found in [23], [1], [9], [7], [6], [8]), rank swapping [5], 5 well known adding noise methods (see e.g. in [3]) including ROMM [25] and various other methods are implemented. There is also a method for fast generation of synthetic data included [18] with which multivariate normal distributed data can be generated regarding to the covariance of the original data, but, of course, this does not reflect the distribution of real complex data. Further methods will be implemented in near future (see below in section *Future Developments*).

Several test data sets are implemented in the package. There are some very small test data sets which used by other authors for demonstration in the past. The test data set from μ-ARGUS [15] can also be used as well as the test data sets from the CASC project [13].

4.1 New Methods

Some new algorithms for microaggregation are proposed and implemented in package `sdcMicro`. A simple approach is to cluster the data at first and then sort the data in each group by the most influential variables in the groups or by sorting the observations in each group by the first robust principal component using e.g. the MCD-estimator [21].

We proposed a new algorithm for microaggregation called RMDM (**R**obust **M**ahalanobis **D**istance based **M**icroaggregation where MDAV [8] is adapted in the following way:

1. Compute the robust center of the data. This can be the L_1-median or the coordinate-wise median.

2. Consider the most distant observation x_r to the robust center using robust Mahalanobis distances. The MCD-Estimator can be used to calculate the robust covariance matrix that is needed for the calculation of the robust Mahalanobis distances.

3. Find the most distant observation x_s by calculating robust Mahalanobis distances with the center in x_r.

4. Choose $k-1$-nearest neighbors from x_r and also from x_s using robust Mahalanobis distances computed with center x_r and center x_s. Aggregate x_r and its $k-1$ nearest neighbors with an average as well as x_s with their $k-1$ nearest neighbors. The average can be the arithmetic mean but also robust measures of location.

5. Take the previous dataset minus the aggregated data from the last step as a new dataset and continue with (1.) until all observations are microaggregated.

Note, that there are special rules at the end of the algorithm which are described in [8]. This proposed algorithm is more natural than the original MDAV algorithm since we deal with multivariate data taking the covariance structure of the data into account. For larger data sets usual distances can be chosen to find the nearest neighbors in item (4.) of the algorithm (we name this adaption of the algorithm RMDM2). These distances must be calculated only once at the beginning of the algorithm.

Robust versions of principal component methods for microaggregation are implemented, too.

The *clustpppca* and the *RMDM2* algorithms for microaggregation worked best for most of the data sets (like the Tarragona data set from the CASC project (see section 5)). The *clustppca* algorithm is described in [23].

Top and bottom coding can be replaced by (multivariate) outlier detection via robust statistics. Only those observations are recoded/perturbed which are outliers and therefore of high re-identification risk. Such a method is included as an adding noise procedure.

5 A Small Tour in sdcMicro

Within the limitation of pages only a small tour in *sdcMicro* can be done excluding most of the graphical results and some steps of recoding variables. Comments are marked with #, the output from **R** with **R**. For further details, please have a look at the examples and documentation which are included in package *sdcMicro*.

Supposing you have already downloaded and installed **R** and have also installed package `sdcMicro`, which can be installed directly from **R** or downloaded directly from *http://cran.r-project.org/src/contrib/Descriptions/sdcMicro.html*, you can load both the package and the data and print[3] the first seven columns and the first four rows of the data with the following command in **R**:

```
library(sdcMicro); data(free1); xtable(free1[1:4, 1:8])
```

[3] `xtable()` produces a LaTeX-styled print output.

	REGION	SEX	AGE	MARSTAT	KINDPERS	NUMYOUNG	NUMOLD
1	36.00	1.00	43.00	4.00	3.00	0.00	0.00
2	36.00	1.00	27.00	4.00	3.00	0.00	0.00
3	36.00	1.00	46.00	4.00	1.00	0.00	0.00
4	36.00	1.00	27.00	4.00	1.00	0.00	0.00

For this demonstration the μ-Argus test data was chosen but you can simply use another data set from the package or your own data.

In the following, the frequency counts are calculated as described in [4] and allocated to object `fr1` which is now automatically of class `freqCalc`. Several methods are available for this class. Object `fr1` can then be used as an input for the individual risk computation. A new object `ir1` will be produced which is of class `indivRisk`. Several methods are again available for this class and, for example, function `plot` will automatically know which plot method must be used. Figure (2) is generated by this plot method. The implementation of this plot method for individual risk methods is quite similar as in μ-Argus.

```
fr1 <- freqCalc(free1, keyVars=1:3, w=30)
rk1 <- indivRisk(fr1)
class(rk1)
R> [1] "indivRisk"
methods(class = indivRisk)
R> [1] plot.indivRisk  print.indivRisk
plot(rk1)
```

The script shown above can easily be adapted (exclude the **R** results), e.g. by adding the function `globalRecode()`. `globalRecode()` recodes several categories of a variable into less categories or discretize a numerical variable. So, this function checks the class of the variable and does the recoding based on the class of the variable. But you still have the ability to manipulate your data in a explorative way. You might try out recoding a variable to observe the influence of your recoding on the frequency count calculation and the individual risk computation. Note that you can easily reproduce all your steps easily either by running a part of your script or the whole script.

After minimising the re-identification risk you can apply local suppression (function `localSupp()`) on object `indivf` and `fr1` to delete the last unsureness about risky observations and make then use of the implemented print and summary methods. But, also global recoding and local suppression can also be used in an alternated manner, of course.

In addition to that, you can simply microaggregate numerical variables with more than 10 different methods. Furthermore, you can also use rank swapping and adding noise methods. There are a lot of comparison plot methods available to compare the perturbed data and the original data. You can also easily compare the different methods themselves. We will show this on another data set, the *Tarragona* data

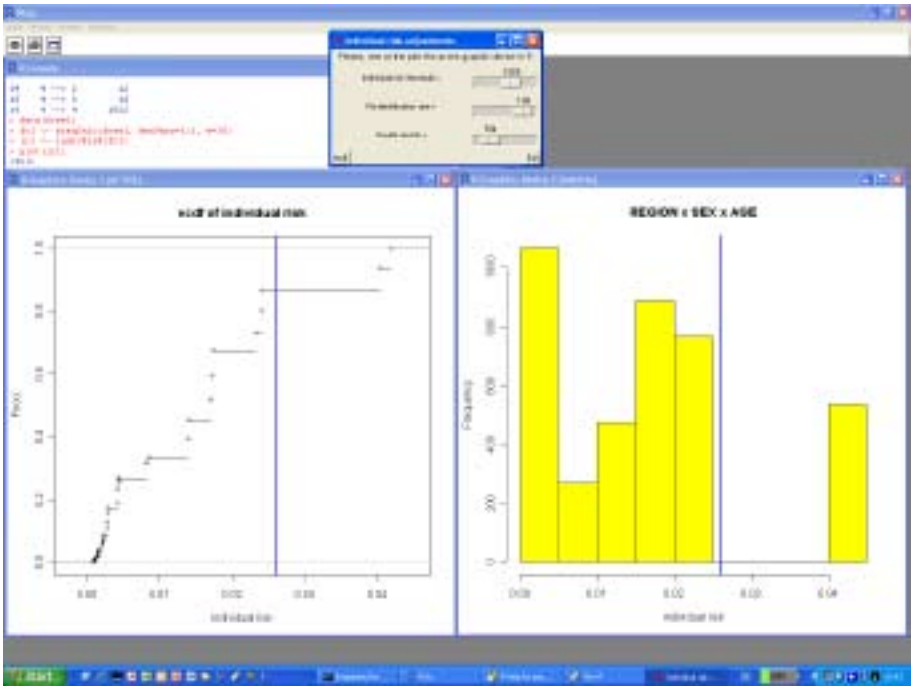


Figure 2: Individual risk for the first 3 variables of the μ-Argus test data set. In the upper region of the figure you will see a helpful slider which is directly linked with the graphics.

set[4], because the μ-Argus test data does include faked numerical variables which follow a multivariate uniform structure without correlations between the variables.

Microaggregation and a comparison of different methods with various measures of information loss can be done in the following way:

```
m1 <- microaggregation(Tarragona, method="onedims", aggr=3) # individual ranking method
# now you can use some comparison plots... (see examples of package sdcMicro)
summary(m1) # or the implemented print and summary methods (suppressed)
xtable(valTable(Tarragona, method=c("addNoise: correlated2", "swappNum", "simple",
"onedims", "pca", "mdav", "clustpppca"))[,c(1,2,3,5,6,7,12)]) # measures of inf. loss
```

	method	amean	amedian	devvar	amad	acov	apcaload
1	addNoise: correlated2	0.09	3.20	0.18	5.89	0.09	8.97
2	swappNum	0.20	0.13	9.10	0.23	4.55	28.75
3	simple	0.00	3.50	3.62	1.11	1.81	20.69
4	individual ranking	0.00	0.02	13.71	0.03	6.85	19.45
5	pca	0.00	2.62	2.77	1.10	1.39	12.99
6	clustpca	0.00	2.34	2.69	1.11	1.35	11.72
7	mdav	0.00	4.18	3.44	1.84	1.72	8.76
8	rmdm2	0.00	1.51	1.76	0.58	0.88	12.00
9	clustpppca	0.00	3.64	2.79	1.55	1.40	10.69

The generated table from function `valTable()` shows that the proposed method RMDM2 (which is explained in section 4.1) performs best (see a few description on

[4]http://neon.vb.cbs.nl/casc/testsets.html

the information loss measures used in [23]) on this data set which is not surprisingly because the multivariate structure of the data is taken into account when using Mahalanobis distances.

In addition to that, various measures of risk and data utility can be applied on these results, e.g. with function `dRisk()` and `dUtility()`.

Another method for categorical variables is PRAM which can be easily applied with function `pram()`. In the following, variable `MARSTAT` form the μ-Argus test data set will be perturbed with the invariant PRAM methodology. A lot of information is stored in object `MARSTAT`, e.g. the invariant transition matrix. Summary and print methods provided as well.

```
MARSTAT <- pram(free1[,"MARSTAT"], p=0.8, alpha=0.5)
summary(MARSTAT)
 ----------------------          ----------------------       ----------------------
 original frequencies:         transitions:                 invariant transition matrix
                                   transition Frequency             [,1]    [,2]    [,3]   [,4]
   1    2    3    4             1     1 --> 1      2448      [1,] 0.9584 0.00903 0.00951 0.0230
2547  162  171 1120             2     1 --> 2        27      [2,] 0.1421 0.73943 0.02836 0.0900
                                 3     1 --> 3        28      [3,] 0.1416 0.02687 0.74079 0.0906
 ----------------------          4     1 --> 4        44      [4,] 0.0523 0.01302 0.01383 0.9207
 frequencies after perturb.:     5     2 --> 1        33
                                 6     2 --> 2       118
   1    2    3    4             7     2 --> 3         4
2571  160  178 1091             8     2 --> 4         7
                                 9     3 --> 1        20
                                 10    3 --> 2         3
                                 11    3 --> 3       130
                                 ...    ...          ...
```

6 Open Source Initiative

As mentioned above, the whole code is free and can be downloaded on *http://cran.r-project.org*. So you can learn from this code, can change code for yourself or develop it further. Instead of keeping the developed code to yourself you are invited to contribute to this package. Every response and bug reports will be helpful in achieving a higher quality for the package. Note that the quality of the package was highly improved by comments and bug reports from many users from statistical offices and companies up to now.

Every function has its own author the copyright of the function is help by the author. This copyright means that nobody can use your function for a commercial software product and in the other hand the intellectual property is also ensured. But, at the same time all the functions are open-source and everybody can use it.

7 Conclusion

This package allows a flexible and explorative use of the most well known methods plus of various new methods. It allows the use of various comparison plots which are more informative than usual measures of information loss. New functionality like additional methods for synthetic data generation, blanking and imputation, etc. will be implemented soon. The potential capacity of this package can be very high, the package has a realistic chance to become the most important implementation for SDC in microdata protection. The respond of users from all over the world is very positive and the package is already used in the production process (see also [19]). The users are still satisfied with the command line interface and so an implementation of a graphical user interface has not been made yet. In addition to this package, the entire power of **R** can be used to boost the results in any way. Everybody is invited to contribute to this package, especially in funded future research projects.

References

[1] N. Anwar. Micro-aggregation - the small aggregates method. In *Internal report*. Luxembourg: Eurostat, 1993.

[2] L. Borchsenius. New developements in the danish system for access to micro data. In *Monographs of official statistics, Work session on statistical data confidentiality*. Eurostat, Luxembourg, 2005.

[3] R. Brand. Microdata protection through noise addition. In *Privacy in Statistical Databases. Lecture Notes in Computer Science. Springer*, pages 347–359, 2004.

[4] A. Capobianchi, S. Polettini, and M. Lucarelli. Strategy for the implementation of individual risk methodology into μ-ARGUS. Technical report, Report for the CASC project. No: 1.2-D1, 2001.

[5] T. Dalenius and S. Reiss. Data-swapping: A technique for disclosure control. In *Proceedings of the Section on Survey Research Methods*, volume 6, pages 73–85, 1982.

[6] D. Defays and Anwar M.N. Masking microdata using micro-aggregation. *Journal of Official Statistics*, 14(4):449–461, 1998.

[7] D. Defays and P. Nanopoulos. Panels of enterprises and confidentiality: the small aggregates method. In *Proceedings of the 1992 Symposium on Design and Analysis of Longitudinal Surveys*, pages 195–204. Statistics Canada, Ottawa, 1993.

[8] J. Domingo-Ferrer and J.M. Mateo-Sanz. Practical data-oriented microaggregation for statistical disclosure control. *IEEE Trans. on Knowledge and Data Engineering*, 14(1):189–201, 2002.

[9] M. Elliot, A. Hundepool, E.S. Nordholt, J-L. Tambay, and T. Wende. Glossary on statistical disclosure control, 2005.

[10] L. Franconi and S. Polettini. Individual risk estimation in μ-ARGUS: a review. In *Privacy in Statistical Databases. Lecture Notes in Computer Science. Springer*, pages 262–272, 2004.

[11] J. Heitzig. The 'jackkife' method: confidentiality protection for complex statistical analyses. In *Joint UNECE/Eurostat work session on statistical data confidentiality, Geneva, Switzerland*, 2005.

[12] J. Heitzig. Using the jackknife method to produce safe plots of microdata. In *Privacy in Statistical Databases. Lecture Notes in Computer Science. Springer*, pages 139–151, 2006.

[13] A. Hundepool. The casc project. In *Privacy in Statistical Databases. Lecture Notes in Computer Science. Springer*, pages 199–212, 2004.

[14] A. Hundepool and P. De Wolf. Onsite@home: Remote access at statistics netherlands. In *Joint UNECE/Eurostat work session on statistical data confidentiality, Geneva, Switzerland*, 2005.

[15] A. Hundepool, A. Van deWetering, Ramaswamy R., L. Franconi, A. Capobianchi, P-P. DeWolf, J. Domingo-Ferrer, V. Torra, R. Brand, and S. Giessing. μ-argus version 4.1 software and users manual, 2006.

[16] P. Kooiman, L. Willenbourg, and J. Gouweleeuw. A method for disclosure limitation of microdata. Technical report, Research paper 9705, Statistics Netherlands, Voorburg, 2002.

[17] Friedrich Leisch. Sweave, part I: Mixing R and LaTeX. *R News*, 2(3):28–31, December 2002.

[18] J.M. Mateo-Sanz, A. Martínez-Ballesté, and J. Domingo-Ferrer. Fast generation of accurate synthetic microdata. In J. In: Domingo-Ferrer, editor, *Privacy in Statistical Databases, Lecture Notes in Computer Science*, pages 298–306. Springer, 2004.

[19] B. Meindl and M. Templ. The anonymisation of the CVTS2 and income tax dataset. an approach using R-package sdcmicro. In *to appear in: Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality. Monographs of Official Statistics.*, 2007.

[20] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2007. ISBN 3-900051-07-0.

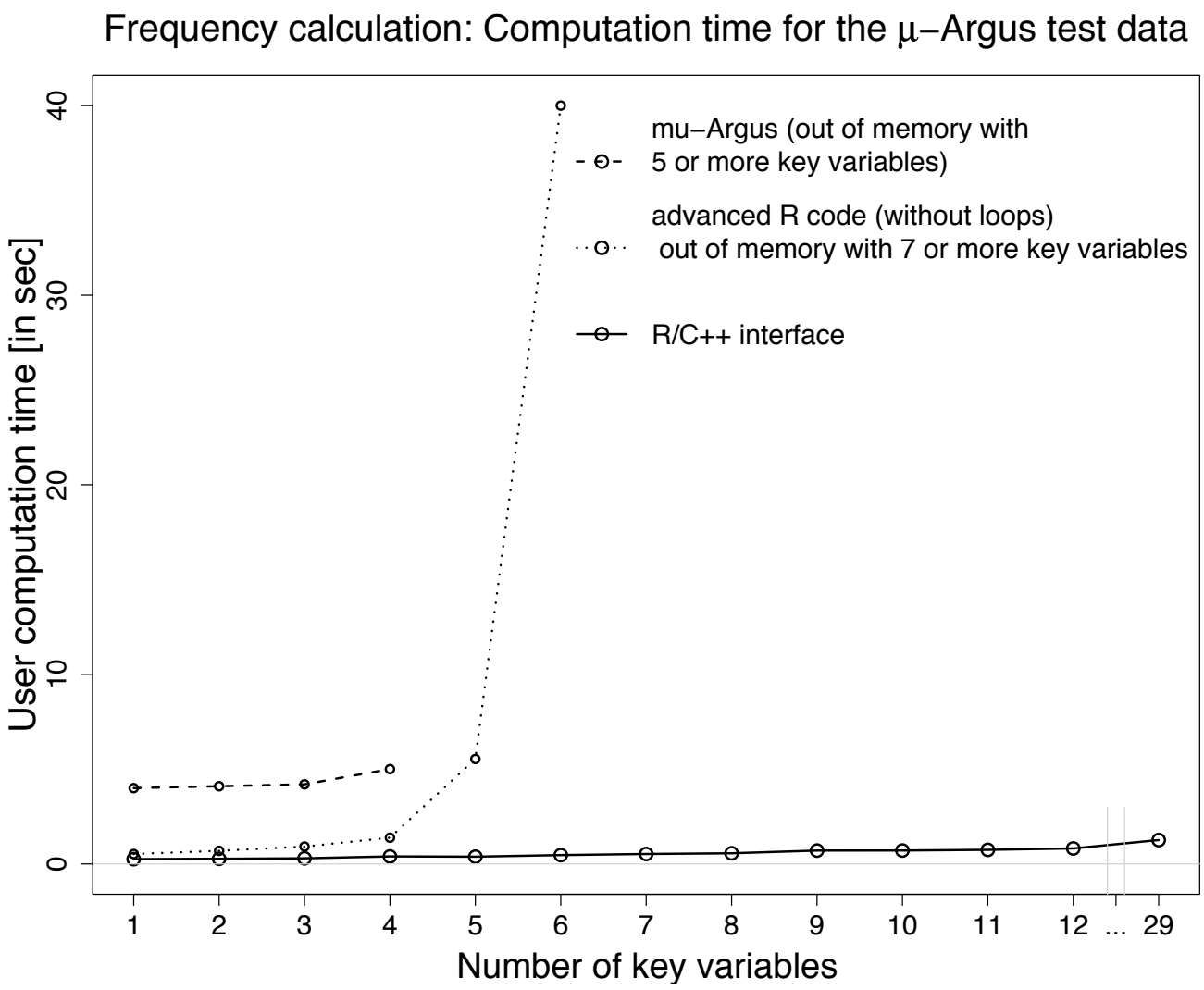
[21] P.J. Rousseeuw. Least median of squares regression. *Journal of the American Statistical Association*, 79:871–880, 1984.

[22] P. Steel and A. Reznek. Issues in designing a confidential preserving model server. In *Monographs of official statistics, Work session on statistical data confidentiality*. Eurostat, Luxembourg, 2005.

[23] M. Templ. Software development for SDC in R. In *Privacy in Statistical Databases. Lecture Notes in Computer Science. Springer*, pages 347–359, 2006.

[24] M. Templ. *sdcMicro. Manual and Package.* Statistics Austria and Vienna University of Technology, Vienna, Austria, 2007. *http://cran.r-project.org/src/contrib/Descriptions/sdcMicro.html.*

[25] D. Ting, S. Fienberg, and M. Trottini. Romm methodology for microdata release. In *Monographs of official statistics, Work session on statistical data confidentiality*. Eurostat, Luxembourg, 2005.


Frequency calculation: Computation time for the µ−Argus test data
User computation time [in sec]
0
10
20
30
40
1
2
3
4
5
6
7
8
9
10
11
12
... 29
Number of key variables
mu−Argus (out of memory with
5 or more key variables)
advanced R code (without loops)
out of memory with 7 or more key variables
R/C++ interface

Applying the EC Regulations about the Dissemination of Unidentified Individual Data for Scientific Purposes in the practice of NSI

During the last few years there is a growing interest by the research community on providing access to unidentified individual data. EC regulations and the Law on Statistics of the Republic of Bulgaria allow confidential statistical data dissemination for scientific purposes at definite conditions.

1. Legislation and legal basis

The Law on Statistics guarantees statistical confidentiality. It contains a special Chapter 6, entitled "Protection of Secrecy", in which data protection principles are laid down. According the amendments in 2005, confidential data could be provided for the purposes of the scientific work to the higher schools or to legal persons, the main activity of which are the scientific studies. There are some obligations to be carried out by the users in order to obtain micro data:

- To create the conditions, as stipulated by normative provisions for data protection;
- All persons, who will be acquainted with these data, sign a sworn declaration for protection of statistical secrecy;
- The data should be transmitted in a form, which does not allow direct or indirect identification of the person they refer to.

In NSI there are Rules on Dissemination of unidentified individual data for scientific purposes. The Rules are prepared in accordance with the Law on Statistics as well as the EC regulations relevant to statistical confidentiality (incl. Commission Regulation № 1000/2007). According the Rules, access to non-identified individual data could be provided for scientific and research purposes to users with a proven name in the field of science and scientific studies. These users should fulfill conditions defined by NSI, in order to receive an access to data. Data

is transmitted after data acceptance–transmission protocols and sworn declarations for protection of the principles of statistical secrecy are signed.

2. Institutional organization

A Council for Data Protection was established as a body with specified functions and tasks on data protection. The NSI's President chairs the Council, while its members are the directors and leading experts in the field of statistical data protection.

The NSI's personnel, when being nominated to a given work position signs a sworn declaration on confidentiality. The confidential data access is limited only to the personnel that work directly with the data in carrying out its everyday official engagements.

Punishments relevant to the offences of statistical confidentiality are legally envisaged.

3. Rules and procedures

In the Rules on Dissemination of unidentified individual data for scientific purposes are described:

- Information about the agencies, organizations and institutions, who my apply for access to confidential data for scientific purposes;
- Statistical surveys, the result of which may be provided for the purposes of the scientific work;
- Procedures for admissibility request;
- Criteria and principles concerning the request for admissibility decision:
  - the primary purpose of the organization,
  - the organisational arrangements for research in the organization,
  - the safeguards in place in the organization,
  - the arrangements for dissemination of the results of research.

A questionnaire for admissibility request is included in the Rules as an annex. The applicant has to fill in the questionnaire with information concerning:

- Identification, legal status and main purpose of the organization;

- Organizational and financial arrangements for research within the organization;
- Safeguards in place – obligations assumed by the applicant concerning data protection by the personnel, who will deal with confidential data;
- Policy on dissemination of the research results of the organization.

In addition, the following documents are required: legal act creating the organization, list of the Board of Directors, organization chart, the staff members who are responsible for the research department and other additional information if it is needed.

The members of Council for Data Protection discuss the filled in questionnaire and enclosed documents and concede or refuse an access to confidential data.

4. Restriction on unidentified individual data providing

The restrictions on individual data providing are described in the Rules on dissemination of unidentified individual data for scientific purposes:

- Unidentified individual data, which summarize the information on less than three persons (units) or in which the relative share of a person (unit) is over 85% of the total volume will not be provided;
- Unidentified individual data, which can be combined in a way leading to the identification of a given person (unit) will not be provided.

The main principles, laid down in the legal documents are in compliance with the EC legislation. The concepts of statistical confidentiality, prohibition to disseminate individual and personal data are defined as well as aggregated data, which refer to groupings of less than 3 units or when one of the grouped units has more than 85% relative share in the group. The Rules on work and dissemination of confidential data are amended and improved according to the last amendments to the Law on Statistics.

Work session on statistical data confidentiality
Manchester 17-19 December 2007

Analytical validity and confidentiality protection of anonymised longitudinal enterprise microdata – Survey of a German Project

Maurice Brandt[1], Rainer Lenz[2] and Martin Rosemann[3]

Research Data Centres of the Federal Statistical Office[1],
University of Applied Sciences Mainz[2]
Institute for Applied Economic Research[3]

Folie 1

RESEARCH DATA CENTRE

DESTATIS wissen.nutzen.

Overview

1. Introduction
2. The data sets of the project
3. Anonymisation methods and analytical validity
4. Approaches to assessing anonymity
5. Conclusions

RESEARCH DATA CENTRE

DESTATIS wissen.nutzen.

1. Introduction

"Business Panel data and de facto anonymisation"
new project since the beginning of 2006

- improve the data infrastructure in Germany regarding longitudinal data on local units and enterprises

- guarantee the access of the scientific community to the panel data of economic statistics

- the formerly project "De facto anonymisation of business microdata" has shown that de facto anonymisation can be achieved on a cross-section basis

Folie 3

RESEARCH DATA CENTRE

DESTATIS wissen. nutzen.

1. Introduction

- In this project different business statistics are linked to longitudinal datasets
- it is planned to complement the data with information from the official business register
- the data sets can already be used for scientific work
- the final aim is to produce a scientific use file

Folie 4

RESEARCH DATA CENTRE

DESTATIS wissen. nutzen.

2.1 The data sets of the project

Units of analysis are the local units in manufacturing and mining
Complete enumeration of local units with 20 or more employees

Monthly reports
- years from 1995 to 2005
- Information about employees, wages, salaries, turnover

Survey of investments
- years from 1995 to 2005
- Information on highly different types of investments

Survey of small units
- years from 1995 to 2002
- Local units with 19 or fewer employees

Folie 5

RESEARCH DATA CENTRE

DESTATIS wissen.nutzen.

2.2 The data sets of the project

Cost Structure Survey

Stratified sample of enterprises with 20 or more employees in the manufacturing and mining sector

- years from 1995 to 2005
- all together over 43.000 enterprises
- Information on output, production factors, employees
- from 1999 to 2002 13.300 enterprises available in the whole period
- studies regarding investments in research and development are possible

Folie 6

RESEARCH DATA CENTRE

DESTATIS wissen.nutzen.

2.3 The data sets of the project

Turnover Tax Statistics

- Very large data set of a total of 4.3 million enterprises years from 2000 to 2004 (1.8 million for the whole period)

- Information on all taxable turnovers, turnover tax, prior tax and of tax liability

IAB Panel of local units

- Information on employment trend, staff structure, hours worked, turnover, export share, investments and innovation

- Since year 1993 various waves on about 4.300 to a max. of 16.000 local units

Folie 7

RESEARCH DATA CENTRE

DESTATIS wissen.nutzen.

3. Anonymisation methods and analytical validity

Anonymisation methods

- methods reducing the information (suppression of variables or presenting key variables in broader categories)
- methods modifying the values of numerical data (data perturbating methods)

Data perturbating methods for panel data

- *Micro aggregation:* (a) separately for all variables and all periods (Individual Ranking), (b) separately for all variables but jointly for all periods, (c) separately for all periods but jointly for all variables and (d) jointly for all periods and all variable
- *Multiplicative stochastic noise: mixture distribution* (approach of Höhne)
- *Multiple Imputation*

Folie 8

RESEARCH DATA CENTRE

DESTATIS wissen.nutzen.

3. Anonymisation methods and analytical validity

In Focus

Impacts of data perturbating methods on

- descriptive distribution measures
- the estimation of econometric panel models, particularly on the within-estimator to control for individual unobservable heterogeneity

First Results

- the within estimator is consistent in the case of anonymisation by individual ranking
- Project team derived consistent within-estimators in the case of anonymisation by multiplicative stochastic noise (including the method of Höhne) and no autocorrelation
- Case of autocorrelation: work in progress
- Multiple Imputation: separate speech on this conference

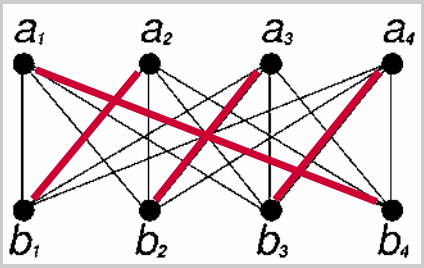
Folie 9

RESEARCH DATA CENTRE

DESTATIS wissen.nutzen.

4. Approaches to assessing anonymity

Linear Assignment Problem (AP)



We calculate distances $d(a_i, b_j)$

and obtain:

(AP) Minimize $\sum_{i=1}^{n} \sum_{j=1}^{n} d(a_i, b_j) x_{ij},$

s.t. $x_{ij} \in \{0,1\}$ for $i, j = 1,...,n,$

$\{a_1,...,a_n\}$ Records of external data

$\{b_1,...,b_n\}$ Records of target data

$$\sum_{j=1}^{n} x_{ij} = 1 \quad \text{for} \quad i = 1,...,n \quad \text{and}$$

$$\sum_{i=1}^{n} x_{ij} = 1 \quad \text{for} \quad j = 1,...,n.$$

RESEARCH DATA CENTRE

DESTATIS wissen.nutzen.

4. Approaches to assessing anonymity

to evaluate the degree of anonymity of anonymised micro data a technique for simulating data intrusion scenarios was necessary

- **Conventional distance based approach**
- **Correlation based approach**
- **Distribution based approach**
- **Collinearity based approach**

RESEARCH DATA CENTRE

DESTATIS wissen.nutzen.

5. Conclusions

Within the scope of the project the panel data sets can be used by

- remote data processing
- safe scientific work stations in the office

They are already used in some research projects

First scientific use files for data use on one's own workstation are probably available at the beginning of 2009

Folie 12

RESEARCH DATA CENTRE

DESTATIS wissen.nutzen.

Thank you for your attention

Folie 13

Analytical validity and confidentiality protection of anonymised longitudinal enterprise microdata – Survey of a German Project

Maurice Brandt*, Rainer Lenz** and Martin Rosemann***

* Federal Statistical Office Germany, Gustav-Stresemann-Ring 11, 65189 Wiesbaden, maurice.brandt@destatis.de
** University of Applied Science, Holzstrasse 36, 55116 Mainz, rainer.lenz@fh-mainz.de
*** Institute for Applied Economic Research, Ob dem Himmelreich 1, 72074 Tübingen, martin.rosemann@iaw.edu

Abstract: The access of the scientific community to cross-section data in the field of business statistics in Germany has considerably improved over the last few years. The purpose of the new project on "Business Panel data and de facto anonymisation" is to extend the data infrastructure in Germany for longitudinal data on local units and on enterprises, so that business statistical data can be made available to empirical researchers for use on their own workstations. This paper gives an overview of the project, describes the data sets and outlines the work done so far to assess the analysis potential and de facto anonymity of the data, after that first results of the project are presented.

1. Introduction

The bases for anonymising enterprise microdata were developed in the project on "De facto anonymisation of business microdata" (Lenz et al. 2006, Statistisches Bundesamt 2005a). A major result of the project was that de facto anonymisation of business statistical data can be achieved on a cross-section basis. De facto anonymisation means, that the cost of trying to reidentify records in the dataset must be higher than the benefit of the disclosed information. In this case a rational data intruder would not even try to deanonymise the dataset, because he would have to put an enormous amount on work, time, manpower and specialized knowledge in the data attack. A new project on "Business Panel data and de facto anonymisation"[1] started at the beginning of 2006 and is intended to clearly improve both the data infrastructure in Germany regarding longitudinal data on local units and enterprises and the access of the scientific community to the panel data of business statistics by the research data centres (Zühlke et al. 2004). The project deals with an improvement of the data supply by longitudinal linkage of statistics which so far have been used mainly on a cross-section basis. The focus is on the cost structure survey in

[1] The project is carried out jointly by the Institute for Employment Research (IAB), the Institute for Applied Economic Research (IAW), the Research Data Centre (FDZ) of the statistical offices of the Länder and the Research Data Centre of the Federal Statistical Office.

manufacturing, the monthly reports in manufacturing, the survey of investments, the industrial small units survey and the turnover tax statistics, which are processed as longitudinal data sets as part of the project. The local units panel of the Institute for Employment Research was selected for anonymisation by means of multiple imputation.

As another important element of the project, it is planned to complement the data linked longitudinally by information from the official business register. The main purpose of that work is to identify by means of the business register (cf. Sturm 2006) reasons for missing data, specially demographic information about enterprises, in longitudinal terms and thus to increase the analysis potential of the data. This will permit, for example, to find out on the basis of the business register whether a reporting unit has no longer been included in the survey because it changed to another economic branch or because its number of employees decreased under the cut-off limit. In these cases its ruled out that the enterprise has been shut down or has merged with another enterprise. As regards turnover tax statistics, the turnover data have already been complemented – on the basis of the business register – by employees data for the years 2000 to 2004.

Longitudinal and panel data are demanded more and more often by scientific users because only with such data is it possible to show the dynamics, changes and processes over time. Another advantage of panel data is that unobservable heterogeneity can be considered. However, the positive aspects provided by longitudinal data for research evaluations might also prove to be an additional challenge to anonymisation. This is because, across several waves, a structure in the data can be detected which gives additional knowledge to a potential data intruder that is helpful in reidentification attempts (Lenz 2008).

With a view to maintaining the analysis potential of panel data it must be ensured that developments over time can adequately be analysed also by means of anonymised data and that panel-econometric methods continue to produce consistent estimates (Biewen/Ronning/Rosemann 2007).

One of the questions to be answered by the project is the extent to which the anonymisation methods developed for cross-section data must be further developed for the anonymisation of panel data and what impact such methods have on data protection and on the analysis potential of the panel data of business statistics.

The outline of the paper is as follows. Chapter 2 contains a description of the datasets and the editing of the data in this project. Chapter 3 goes into the anonymisation methods of panel data and the analytical validity of the anonymised panel data. The Chapter 4 gives an overview about the possibilities to measure the disclosure risk to achieve de facto anonymity of the panel data. The paper ends with a summary and outlook on further projects.

2. The data sets of the project

For the longitudinal linkage and the subsequent anonymisation, business data were selected for which some experience is available regarding cross-section anonymisation and which are demanded most often by researchers.[2]

2.1 Monthly reports, survey of investments and survey of small units

Based on the local units as a unit of analysis, the monthly reports in manufacturing, mining and quarrying are a longitudinal linkage of the years from 1995 to 2005. They contain information on employees, wages and salaries, and turnover (Statistisches Bundesamt 2007a). The survey of investments, however, provides information on highly different types of investments (Statistisches Bundesamt 2007c) and basically contains the same local units as the monthly reports. The monthly reports represent a complete enumeration of the local units with 20 or more employees.[3] The range of data is complemented by the survey of small units of the years 1995 to 2002, which supplies information from local units with 19 or fewer employees.

For the panel data set, the individual data supplies have been aggregated to form an annual data supply. The data contain information on employees, turnover (domestic and foreign turnover), hours worked, wages and salaries, and investments (Konold 2007). Wagner (2007) contains some examples of comments on the research potential of the monthly reports.

2.2 Cost structure survey

The cost structure survey is a stratified sample with almost 18,000 enterprises each year. The data of the cost structure survey in manufacturing, mining and quarrying are designed as a longitudinal data set for the years from 1995 to 2005. The cost structure survey is suited for manifold structural analyses (Fritsch et al. 2004) and provides comprehensive information on output, the production factors used, and on the value added of enterprises with at least 20 employees (cf. Statistisches Bundesamt 2007d). The longitudinal data set contains a good 43,000 cases for the years 1995 to 2005. The way of processing allows to perform analyses both on a cross-section basis for the reference year and on a longitudinal basis. For the period from 1995 to 2005, there are a good 2,000 enterprises which were questioned every year. A large part of those enterprises come from areas fully covered (branches with few cases, large enterprises). For the years 1999 to 2002, there are still just under

[2] In consequence of the project on "Anonymisation of business microdata", further enterprise statistics such as the structure of earnings survey, could be anonymised (cf. Hafner and Lenz 2007).
[3] An exception is 14 economic branches with 10 or more employees (cf. Statistisches Bundesamt 2007b/c).

13,300 enterprises which were questioned every year, thus providing sufficient potential for scientific analyses and shall cover the period for the scientific use file.

2.3 Turnover tax statistics

The longitudinal linkage of turnover tax statistics comprises a data set of a total of some 4.3 million enterprises, about 1.8 million of which can be linked for the period from 2000 to 2004 to form a real panel data set. In a first step, the panel data set for 2000-2004 was established for special analyses at the Federal Statistical Office and for remote data purposes. For every case, the file contains a data set with a total of 156 variables for 5 reference years, with differing numbers of variables actually occurring, depending on the existence of the enterprise in the relevant year. Turnover tax statistics contains information on all taxable turnovers, turnover tax, prior tax, and duration of tax liability (Statistisches Bundesamt 2005b).

2.4 IAB panel of local units

The IAB panel of local units is a representative survey among employers on local unit items influencing employment and covers a stratified sample of all local units with at least one employee subject to social insurance contributions in Germany. The panel contains information allowing to perform analyses of the development of labour demand on the labour market in Germany. Items covered include information on the employment trend, weekly hours worked, turnover, and export share, investments and innovation in the local unit, public subsidies, staff structure, vocational training and apprenticeship positions, staff recruited and staff leaving, search for new staff, wages and salaries, hours worked in the local unit, advanced training and continuing education. The local units panel has been produced every year since 1993 in western Germany and since 1996 in eastern Germany by the IAB research unit "Local units and employment". The local units panel contains information of the various waves on about 4,300 to a maximum of some 16,000 local units (Bellmann 2002).

3. Anonymisation Methods for Panel Data and the Analytical Validity of anonymised Panel Data

In the last decade a broad variety of anonymisation methods is described in literature (see for example Brand (2000), Höhne (2003), Statistisches Bundesamt (2005a) and Rosemann (2006)). Anonymisation methods may be subdivided into two groups: methods reducing the information, and more recent methods modifying the values of numerical data (data perturbating methods). When an anonymisation concept for business micro data is developed a mix of these two approaches often seems to be the best solution. Information reducing methods such as the suppression of variables or

presenting key variables in broader categories should be preferred, provided that the analyses of interest to the users can still be made. However, if it seems inevitable to additionally apply anonymisation measures which modify the data, a method has to be agreed upon and the parameters of that method need to be balanced appropriately (Lenz et al. 2006).

In Statistisches Bundesamt (2005a) most known anonymisation procedures have been rated both with regard to data protection and to informational content left after perturbation. In particular micro aggregation or stochastic noise has been found convenient for continuous variables whereas "Post Randomization" (PRAM) can be recommended with some reservations for discrete variables. Additionally, most recently multiple imputation has been suggested by Rubin (1993) for data protection.

The basic idea of (deterministic) micro aggregation is to form groups of similar objects and to substitute the original values by the arithmetic mean of this group (Mateo-Sanz and Domingo-Ferrer 1998).[4] The variants of deterministic micro aggregation principally differ with regard to the question whether the micro aggregation is performed jointly for all numerical variables or separately for each variable.[5] In the first case therefore the same groups are formed for different variables when determining the averages. In the second case (individual ranking) the groups are formed for the several variables separately.

In the case of panel data we have r variables, T periods and N observations. So we can perform the micro aggregation (a) separately for all variables and all periods (Individual Ranking), (b) separately for all variables but jointly for all periods, (c) separately for all periods but jointly for all variables and (d) jointly for all periods and all variables.

Micro aggregation preserves the expected values original but leads to a decreasing variance. Therefore Höhne (2004a) develops a variant of individual ranking that preserves the variances too. He builds up groups of size four. Then for two of these observations in group i anonymised values are given by $x^a_{i,1/2} = \bar{x}_{i.} - sd(x_i)$ whereas for the two other anonymised values $x^a_{i,3/4} = \bar{x}_{i.} + sd(x_i)$ is used where $\bar{x}_{i.}$ is the average of the variable x in group i and sd(xi) is the standard deviation of x in this group.

The alternative approach of addition or multiplication of stochastic noise is one of the most important data perturbating methods. In the additive case the noise variable usually is assumed to be normally distributed with expectation zero. To increase the data security one can use a mixture distribution of normal distributions where the expectations of the underlying component distributions are unequal to zero. In the

[4] For stochastic micro aggregation see Rosemann (2006).
[5] Also used are variants where the set of numerical variables is subdivided into groups first and where the variables of a group are then micro aggregated jointly (Statistisches Bundesamt 2005a).

case of anonymisation we can restrict ourselves to a mixture distribution of two normal distributed components with expectations −μ and μ (Roque 2000, Yancey et al. 2002, Höhne 2004b and Statistisches Bundesamt 2005a).

We achieve better protection for larger firms if we use multiplicative noise (Statistisches Bundesamt 2005a). In this case the expectation of the noise variable should be one and the values of the noise variable should be limited to the positive band. Several distributions can be used, e.g. lognormal or uniform distribution. As an alternative, also in the multiplicative case a mixture distribution of two normal distributions is used, where the expectations are 1−f and 1+f. The parameter f as well as the standard deviations of the two components (which equal each other) are chosen in such a manner that the values of the noise variable remain positive.

A special variant of a mixture distribution was proposed by Höhne (2004b). The main idea of this approach is that for one observed unit all values are scaled down or scaled up. In other words, for every unit the probability to draw from a normal distribution with expectation 1−f is 0.5 and corresponds to the probability to draw from a normal distribution with expectation 1+f. If we adopt this anonymisation method on the case of panel data we can distinguish several variants for the multiplicative noise variable w_{ijt} of observation i, variable j and period t.

$$w_{ijt} = 1 + d_i f + \varepsilon_{ijt} \qquad (3\text{-}1)$$

$$w_{ijt} = 1 + d_{ij} f + \varepsilon_{ijt} \qquad (3\text{-}2)$$

$$w_{ijt} = 1 + d_{it} f + \varepsilon_{ijt} \qquad (3\text{-}3)$$

$$w_{ijt} = 1 + d_{ijt} f + \varepsilon_{ijt} \qquad (3\text{-}4)$$

In all cases we assume $\varepsilon_{ijt} \sim N(0, \sigma_\varepsilon^2)$ and the variable d takes on +1 and −1 with probability 0.5.

Another auspicious method to anonymise panel data is multiple imputation (Rubin 1993, Raghunathan et al. 2003). In 1993 Rubin suggested to generate fully synthetic data sets to guarantee confidentiality. His idea was to treat all the observations from the sampling frame that are not part of the sample as missing data and to impute them according to the multiple imputation framework. Afterwards, several simple random samples from these fully imputed datasets are released to the public.

However, the quality of this method strongly depends on the accuracy of the model used to impute the "missing" values. If the model doesn't include all the relationships between the variables that are of interest to the analyst or if the joint

distribution of the variables is mis-specified, results from the synthetic data set can be biased. Furthermore, specifying a model that considers all the skip pattern and constraints between the variables can be cumbersome if not impossible

To overcome these problems, a related approach suggested by Little (1993) replaces observed values with imputed values only for variables that are publicly available in other databases or for variables that contain especially sensitive information leaving most of the data unchanged. This approach has been adopted for some data sets in the US. In our project both approaches are tested in time with data of the IAB establishment panel (first results can be found in Drechsler et al. (2007) and Reiter and Drechsler (2007)).

The methods described above should ensure confidentiality of panel data at the same time the usefulness of data should be gained. The analytic potential is limited on the one hand by the fact that certain analyses are excluded from the start by the anonymisation procedures it selves because either the issue in question cannot be analysed anymore or the method to be used and equivalent methods cannot be applied anymore. This could be the main problem in the case of using methods reducing the information. On the other, such limits result form anonymised data producing results which differ from those based on the original data. When anonymisation procedures are assessed which modify the data, the focus is on the second aspect.

When we use data perturbating methods we have to ensure that distributional properties of the data do not change too much. However, in the project "Business Panel data and de facto anonymisation" the impacts of data perturbating methods on analysis using special qualities of panel data are in focus. On the one hand the project analyses the impacts of the described data perturbating methods on descriptive distribution measures where cross-sectional measures are supplemented by special aspects of panel data, for instance measures relating to the rates of change. On the other hand we focus on the effects of these methods on the estimation of econometric panel models, particularly if we use the within-estimator to control for individual unobservable heterogeneity. These analyses include theoretical derivations as well as simulation experiments and examples with data of official statistics. First results of this work are available.

Biewen et al. (2007) show that the within estimator is consistent in the case of anonymisation by individual ranking. These results correspond to the results of Schmid (2006) for the OLS-estimator. Biewen (2007) derives a consistent within-estimator in the case of anonymisation by multiplicative stochastic noise. The paper focuses on the case of no autocorrelation as yet. Ronning (2007) deals with the effects of stochastic noise using a mixture distribution, for instance the method proposed by Höhne (2004b). In the case of panel data he focuses on the variant described in formula (3-1). However such a distribution will imply correlation of measurement errors. This is of special concern if linear (or nonlinear) models are

estimated from data anonymised in such a way. This case so far had not received much attention since usually measurement errors are assumed to be independent across variables. It can be shown that the measurement error of the dependent variable in this case no longer can be considered as harmless to estimation. A consistent fixed effects estimator using the method of Höhne can be found in Ronning (2007) as well as in Biewen (2007).

4. Approaches to assessing de facto anonymity

In order to evaluate the degree of anonymity of previously anonymised micro data, it was necessary to develop a technique for simulating data intrusion scenarios a potentially attacking data intruder might apply. One important constellation is the so-called database cross match scenario. In a database cross match scenario, an attacking data intruder tries to assign as many external database units as possible (additional knowledge) uniquely to units of an anonymised target database in order to extend the external database by target database information.

In a first phase, the database cross match scenario was mathematically modelled as a multicriteria assignment problem, which was then converted, by way of suitable parameterisation, into an assignment problem with one target function to be minimised. Then, the main concern was to choose the best-fitting coefficients of this target function. Whereas in the past a distance measure, generated across all matching variables of the two data sources (key variables and overlaps), proved to be well suited for the examination of cross-sectional data (Lenz 2006), it turned out that the examination of panel data requires the use of additional, more elaborated measures. As the information on variables, which in the case of panel data is available to a potential data intruder, has been collected in several waves, it seems obvious that this more complex structure should be reflected in the coefficients of the linear program as well. With that goal in mind we have implemented and tested several promising approaches. A more detailed description of these approaches can be found in Lenz (2008).

4.1 Conventional distance based approach

For every numerical key variable vi and every pair of records *(a,b)* in the two data sources, the standardised square deviation is calculated. Afterwards, these component deviations are summed up. It may be advisable in some cases to assign additional weights to the various deviations on variable level. However, a weakness of that measure becomes apparent in cases where the definition of some key variable slightly differs between the two data sources, for example, if a variable such as “number of employees" relates to the number of all employees in absolute terms in one data set, whereas that number is converted into full-time workers in the other data set.

4.2 Correlation-based approach

Let $v^e_1,\dots, v^e_k$ and $v^t_1,\dots, v^t_k$ be ordinal key variables of the external and target data, respectively. We define $\boldsymbol{v}^e$ and $\boldsymbol{v}^t$ as variables from which k realisations have been drawn and calculate the empirical correlation $corr(\boldsymbol{v}^e; \boldsymbol{v}^t)$ using Spearman's coefficient. The less this coefficient deviates from 1 the more likely the record pair (a,b) belongs to the same enterprise. Note that this coefficient can be applied either in case of numerical (and hence also ordinal) variables or in case of categorical variables, whose range forms a well-ordered set.

4.3 Distribution based approach

In a panel data situation we can take it for granted that an attacking data intruder will have information over several years for every key variable, for example, total turnover of an enterprise from 1999 to 2002. In general, we can assume the existence of a bias between the two sources of data in these variables. In order to counteract this problem, we consider the annual changes of a key variable and treat them like a frequency distribution of a discrete variable. Hence, we can apply statistical methods in order to measure the "similarity" of the frequency distributions on either side, external and target data.

4.4 Collinearity approach

A data intruder might have information on two key variables over a period of n years in both sources of data, e.g., "total turnover" $(u_1, \dots, u_n)$ and "number of employees" $(b_1, \dots, b_n)$ of an enterprise. If we interpret the pairs of values (u_i,b_i) as realisations of two random variables, those units that belong together in the different data sources can be expected to reveal empirical correlation cofficients that are 'similar'. It should, however, be considered that what is measured by correlation is just the linear interrelation of two variables. In special cases the two estimated correlation coefficients can diverge from each other very clearly, even if the variables are linked by a direct functional relationship.

4.5 Combination of approaches

Because of the mentioned weaknesses of the various measures described above they are combined in a suitable way. Here we distinguish between two types of combination, *hybrid* and *composite* matching, see Lenz (2008). Once the coeffcients $d(a_i,b_j)$ are calculated, one can solve the linear assignment problem using classical established methods such as the simplex method. For larger data blocks (typically generated when dealing with tax statistics) it is recommendable for reasons of efficiency that approximation heuristics should be used. Fortunately, the usage of

appropriate heuristics yields results near the optimum solution of the assignment problem, see Lenz (2003).

5. Outlook

Already within the scope of the project on "Business Panel data and de facto anonymisation", some panel data sets were supplied. They are already used in some research projects. The cost structure survey for the years 1995 to 2005, the monthly reports from 1995 to 2005, the survey of investments from 1995 to 2005 and the survey of small units for the years 1995 to 2002 in manufacturing as well as the data of the turnover tax statistics for 2000 to 2004 are available through remote data access and by using safe scientific workstations at the statistical offices.

First Scientific Use Files for data utilisation on one's own workstation will presumably be made available at the beginning of 2009. The project should permit to automate the processing and anonymisation of other business statistics over time. Also, the experience thus acquired will be used for further projects such as the integration of business data from various surveys and years.

References

Bellmann, L. (2002). *Das IAB-Betriebspanel. Konzeption und Anwendungsbereiche. In Allgemeines Statistisches Archiv, Bd. 86, H. 2, 177-188.*

Biewen, E. (2007). *Within-Schätzung bei anonymisierten Paneldaten, IAW-Diskussionspapier Nr. 34.*

Biewen, E., Ronning, G. and Rosemann, M. (2007). *Estimation of Linear Panel Models with Anonymised Business Data. In: IAW-Report 1/2007, 87-114.*

Brand, R. (2000). *Anonymität von Betriebsdaten – Verfahren zur Erfassung und Maßnahmen zur Verringerung des Reidentifikationsrisikos, Beiträge zur Arbeitsmarkt- und Berufsforschung, 237.*

Drechsler, J., Dundler, A., Bender, S., Rässler, S., and Zwick, T. (2007). *A new approach for disclosure control in the IAB establishment panel. Multiple imputation for a better data access.Tech. rep., IAB Discussion Paper, No.11/2007.*

Fritsch, M., Görzig, B., Hennchen, O. and Stephan, A. (2004). *Cost Structure Surveys in Germany, Schmollers Jahrbuch / Journal of Applied Social Science Studies 124, 557-566.*

Hafner, H.-P. and Lenz, R. (2007). *Die Gehalts- und Lohnstrukturerhebung: Methodik, Datenzugang und Forschungspotential, FDZ-Arbeitspapier Nr. 18.*

Höhne, J. (2003). *Methoden zur Anonymisierung wirtschaftsstatistischer Einzeldaten, in:* Gnoss, R./Ronning, G. (Eds.): *Anonymisierung wirtschaftsstatistischer Einzeldaten,* Wiesbaden, *pp. 69-94.*

Höhne, J. (2004a). *Weiterentwicklung von Mikroaggregationsverfahren, Arbeitspapier des Projekts `Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten´.*

Höhne, J. (2004b). *Varianten von Zufallsüberlagerungen, Arbeitspapier des Projekts `Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten´.*

Konold, M. (2007). *New possibilities for economic research through integration of establishment-level panel data of German official statistics, Schmollers Jahrbuch / Journal of Applied Social Science Studies 127.*

Lenz, R. (2003). *Disclosure of confidential information by means of multi-objective optimisa- tion. Proceedings of the Comparative Analysis of Enterprise Data Conference (CAED),* London. *(CD-ROM publication, see http://www.statistics.gov.uk/events/caed/abstracts/lenz.asp)*

Lenz, R. (2006). *Measuring the disclosure protection of micro aggregated business microdata - An analysis taking the example of German Structure of Costs Survey. Journal of Official Statistics 22 (4),* Sweden, *681-710.*

Lenz, R., Rosemann, M., Vorgrimler, D. und Sturm, R. (2006). *Anonymising business micro data - results of a German project. Journal of Applied Social Science Studies (Schmollers Jahrbuch) 126 (4), 635-651.*

Lenz, R. (2008). *Risk Assessment Methodology for Longitudinal Business Micro Data. AStA- Wirtschafts- und Sozialstatistisches Archiv, to appear.*

Little, R. (1993). *Statistical Analysis of Masked Data, in: Journal of Official Statistic, Vol. 9; pp. 407-426.*

Mateo-Sanz, J., Domingo-Ferrer, J. (1998). *A Method for Data-Oriented Multivariate Microaggregation, in: Statistical Data Protection, Proceedings of the conference Eurostat 1999.*

Raghunathan, T., Reiter, J., Rubin, D. (2003). *Multiple Imputation für Statistical Disclosure Limitation, Journal of Official Statistics, 19, pp. 1.16.*

Reiter, J and Drechsler, J. (2007). *Releasing Multiply-Imputed Synthetic Data Generated in Two Stages to Protect Confidentiality. IAB Discussion Paper, No.20/2007.*

Ronning, G. (2007). *Stochastische Überlagerung mit Hilfe der Mischungsverteilung, IAW-Diskussionspapier Nr. 30*

Roque, G. (2000). *Masking Microdata Files with Mixtures of Multivariate Normal Distributions, Ph.D. thesis, University of California, Riverside.*

Rosemann, M. (2006). *Auswirkungen datenverändernder Anonymisierungsverfahren auf die Analyse von Mikrodaten. IAW-Forschungsbericht, Nr. 66,* Tübingen.

Rubin, D. (1993). *Discussion: Statistical Disclosure Limitation, in: Journal of Official Statistics, 9(2), pp. 461-468.*

Schmid, M. (2006). *Estimation of a linear model under microaggregation by individual ranking, Allgemeines Statistisches Archiv, Vol. 90 Nr. 3.*

Statistisches Bundesamt (2005a). *Statistik und Wissenschaft, Handbuch zur Anonymisierung wirtschaftsstatistischer Mikrodaten, Band 4.*

Statistisches Bundesamt (2005b). *Umsatzsteuerstatistik, Qualitätsbericht.*

Statistisches Bundesamt (2007a). *Monatsbericht für Betriebe des Verarbeitenden Gewerbes sowie des Bergbaus und der Gewinnung von Steinen und Erden, Qualitätsbericht.*

Statistisches Bundesamt (2007b). *Produktionserhebungen, Qualitätsbericht.*

Statistisches Bundesamt (2007c). *Investitionserhebung bei Unternehmen und Betrieben des Verarbeitenden Gewerbes sowie des Bergbaus und der Gewinnung von Steinen und Erden, Qualitätsbericht.*

Statistisches Bundesamt (2007d). *Kostenstrukturerhebung im Verarbeitenden Gewerbe, im Bergbau sowie in der Gewinnung von Steinen und Erden, Qualitätsbericht.*

Sturm, R. and Tümmler, T. (2006). *Das statistische Unternehmensregister - Entwicklungsstand und Perspektiven. In: Wirtschaft und Statistik 10/2006, 1021-1036.*

Wagner, Joachim and Ulrich Kaiser (2007). *Neue Möglichkeiten zur Nutzung vertraulicher amtlicher Personen- und Firmendaten. FDZ-Arbeitspapier Nr. 20.*

Yancey, W., Winkler, W. and Creezy, R. (2002). *Disclosure Risk Assessment in Perturbative Micro Data Protection, in*: Domingo-Ferrer, J. (Ed.): *Inference Control in Statistical Databases – From Theory to Practice,* Berlin, *pp. 135-152.*

Zühlke S., Zwick, M., Scharnhorst S., Wende, T. (2004). *The research data centres of the Federal Statistics Office and the statistical offices of the Länder, Schmollers Jahrbuch 124, 567-578.*

Dealing with Confidentiality in Dissemination: The experience of the Basque Statistics Office

Marta Mas[1] and Cristina Prado[2]

[1] Technical Assistance, Vitoria-Gasteiz, Basque Country (SPAIN)
[2] Basque Statistics Office (EUSTAT), Vitoria-Gasteiz, Basque Country (SPAIN)

Abstract. One of the main goals of a statistical agency is to maintain and provide statistical confidentiality for its respondents. Confidentiality should be preserved in all the stages of statistical production and especially in the dissemination phase. If we consider the wide range of formats in which statistical information is available (tables, microdata, metadata, etc.) and the detailed classifications and fine-scaled geographical levels released, the problem of data protection has become a far from trivial issue lately. This paper describes not only the experience of the Basque Statistics Office (EUSTAT) in providing protection for its published products, but also the development of a comprehensive policy that includes the establishment of standard protection criteria, the constitution of an expert group and a commitment to future tasks.

Keywords. Confidentiality, Identification, Statistical disclosure control, Microdata protection, Tabular data protection, On-site access

1 Legal framework and preliminary issues

One of the main goals of a statistical agency is to maintain and provide statistical confidentiality for its respondents. The sole use of information for statistical purposes should be also guaranteed. Privacy rights are preserved by our Constitution and considered in statistical laws. Specifically, Chapter IV of the Basque Statistics Law (23rd April, Law 4/1986) concerns the duty to keep statistical secret and the type of data protected:

> *"[...]the duty to keep statistical secret protects any identifiable data as belonging to an specific person [...]"*

In addition, at national level, the Organic Law of Personal Data Protection (13th December, Law 15/1999) guarantees the protection of personal data, defining this concept as:

> *"[...] any information related to an identified or identifiable person"*

But, what does *identifiable* mean?. This question was partially answered and defined by European Directive 95/46/EC which considers identification by *direct* or *indirect* means. This is to say that not only direct identifiers (ID numbers, names, surnames, addresses, telephone numbers, etc.) must be protected against disclosure but also indirect identifiers (sex, age, marital status, relation to activity, etc) or combinations of them should be considered to avoid identification.

Since it was founded in 1986, EUSTAT has implemented physical and technological measures in order to protect published products against direct and indirect identification. As a result of the internal statistical project *"Research and Development in Statistical Data Protection Techniques"*, several actions have been taken during the last ten years:

Period	Action	Output
1988-1999	Research fellowship on data protection techniques and statistical confidentiality	Technical notebook on "*Statistical Data Protection Techniques"* edited by EUSTAT.
April 2000	International Seminar on "Confidentiality and statistical data protection techniques" organized by EUSTAT. Lecturer: L.H. Cox	Publication: *"Confidentiality and statistical data protection techniques"* L.H. Cox edited by EUSTAT.
September 2000	Security Analysis of Census Tables	Internal report about sensitive crosses and dissemination proposal
October 2000	Participation in OFISTAT (Official Statistics List of distribution) Seminar on Statistical Confidentiality	Discussion about the proposed document: *"Statistical Secret protection: basic elements of a data protection system"* by A.Garín, J. Urrutia
2001	Participation in The Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality (Skopje, Macedonia, 14-16 March)	Article: *"A comparative test for several threshold values in frequency tables: A Tau-Argus performance example."*

2002	Tabular Data protection of preliminary results of the Census 2001, using Tau-Argus (optimal method).	Publication of suppression patterns for frequency tables with fine geographical levels.
2003-2004	CASC project pursuit.	Testing of Argus software.
June 2004	Attendance of PSD (Privacy in Statistical Databases) Conference. (Barcelona, Spain, 6-9 June)	
2005	Staff training on disclosure control and protection software.	Internal Workshop on SDC techniques and ARGUS.
2006	Work on standard safety criteria	Internal report about analysis of sources and internal situation.
December 2006	Attendance of PSD Conference. (Rome, December)	Feedback and contacts.

Table 1.1 Summary of actions in *"R&D in Statistical Data Protection Techniques"*

A lot of work has been done but data protection practices are still applied by statisticians as a part of a "non-written" code of practice, based more on "know-how" rather than on stated rules. Therefore, standards should be discussed and fixed. In fact, one of the recommendations on Statistical Confidentiality included in Principle 5 of the European Statistics Code of Practice[1], refers to this point:

> *"[…] Instructions and guidelines are provided on the protection of statistical confidentiality in the production and dissemination processes. These guidelines are spelled out in writing and made known to the public […]"*

Since 2006, important efforts have been devoted to fulfilling this principle. Finally, in April 2007, a first draft on standard safety criteria was agreed after months of discussion. This whole process, the decisions adopted and the actions carried out are described in the following sections.

[1] Adopted by Statistical Programme Committee on 24 February 2005

2 Current situation

2.1 Confidentiality Board.

One of the main points of this process is the need for an expert group to make decisions about data protection and to give advice on confidentiality matters. Such a group should integrate members from all areas of EUSTAT in order to cover both technical and legal aspects.

The Confidentiality Board has been constituted this year at EUSTAT with the highest representative of each area and the General Direction. These are some of its main duties:

- To establish rules and criteria about confidentiality issues.
- To establish and make decisions concerning sensitive topics and sensitive variables.
- To discuss and approve public-use microdata structure.
- To decide about on-site access conditions.
- To solve specific queries (research-use microdata, etc.)
- To advise other statistical agents from the Basque statistics system on confidentiality matters and data protection procedures.
- To keep a coherent and updated system.

2.2 Establishing confidentiality criteria in dissemination.

2.2.1 Research of sources and other experiences.

A small group of experts (mainly from the methodological area) was constituted in order to make a preliminary analysis of the situation. In the first stage, this analysis consisted of an external search of sources and experiences from other statistical offices, regarding their policies on reporting and implementing confidentiality. The results of this phase were as varied as the sources consulted, but general conclusions could be drawn from the study:

- Legal framework is available to all sources consulted and it is considered essential as a starting point. In addition, guidelines about confidentiality treatment were found in all of them.
- It is less common to find information about the sensitivity rules applied and the values for the parameters of such rules, which are, in most cases, confidential.
- Disclosure control methods are applied to tables and microdata in most cases.
- Geographical thresholds are applied in many cases with diverse values and mainly in microdata releases.
- Almost all the sources provide microdata products (for research use and/or public use)

In the second stage, a summary of the most common data protection practices used by EUSTAT in dissemination was included. On the one hand, only the economic statistics area applies a standard procedure to protect the released data. In spite of the fact that any counting (frequency table) of establishments and companies is allowed by the Basque Statistics Law, no economical magnitudes are published if a cell frequency is less than three (only two or less establishments or enterprises contribute to one cell). Recoding of categories and manual suppressions are applied in order to avoid disclosure. On the other hand, the socio-demographic statistics area often applies 'ad-hoc' protection for each particular case, if a problem of a breach of confidentiality arises.

From the general EUROSTAT guidelines, the experience of other statistical offices and our own practices, an initial proposal on confidentiality criteria for standard dissemination has been developed at EUSTAT.

2.2.2 Microdata protection rules

Although there is no standard release of microdata at this moment in EUSTAT, some rules have been developed to be taken into account for specific demands of information and future public-use files:

- Microdata files released should not include, in any case, either direct identifiers or personal data.
- In general, microdata files will not include geographical indicators referring to areas under a fixed threshold (10,000 inhabitants).
- Aggregation level for other variables included in the file will depend on geographical level released and sensitivity of the variable itself. Therefore, the more geographical detail, the less conceptual level and the greater the sensitivity of the variable the greater the aggregation of categories.
- As an additional protection, disclosure control techniques (perturbation methods, record swapping, noise addition, etc.) could be applied to microdata, always preserving the statistical properties of data.

2.2.3 Tabular data protection rules

Some new rules have been added to the current uses of table protection at EUSTAT. The existing ones have been specified in more detail or modified in some way:

- Low values should be avoided in frequency tables with multiple crossings, where at least one of the variables is sensitive and the geographical indicator refers to an area under a fixed threshold (10,000 inhabitants).
- Dominant contributions should be avoided in magnitude tables in order to prevent accurate estimation of sensitive data belonging to a contributor in a cell. Sensitivity rules (minimum frequency rule or concentration rules (n, K – rule, pq-rule, etc.)) will be applied to detect sensitive cells.

- Appropriate protection techniques for tabular data (recoding of variables, primary and secondary cell suppression, etc.) will be applied in order to protect sensitive cells from disclosure.
- As a general rule, specific demands for information should respect the same protection criteria as standard dissemination. However, certain cases could be studied and discussed by the Confidentiality Board and specific measures might be taken.

2.3 Checking confidentiality criteria.

Having made a proposal on safety criteria, the next step consists of checking these rules against the data published at this moment by EUSTAT. Nowadays, the main results of statistics and data products are released through our website (www.eustat.es) by means of statistical tables and the data bank. Both sources will be revised.

Throughout this process, we shall focus on two main aspects: the geographical scope and the sensitivity of the variables used in each revised table. According to the confidentiality rules recently approved, low frequencies should not be published if the geographical detail refers to areas under a fixed threshold and at least one sensitive variable is involved. Therefore, each table considered should fulfil both conditions. In addition, dominant contributions in magnitude tables (mainly in economical surveys) will be checked.

At the moment of writing this contribution, we are in full checking process of our published products. The results of the checking will be shown to the members of the Confidentiality Board, who will have to discuss and decide about the problems or weaknesses found.

3 Future tasks

3.1 Towards a safe-standard microdata structure

A step forward in EUSTAT dissemination policy is the general release of microdata as a statistical product. Apart from the Economical Activity Directory, which is public by law[2], only two anonymised microfiles containing social survey data have previously been ceded in response to specific demands for information. However, the objective is to develop a standard product which will be available for the general public but with all the guarantees of confidentiality protection.

[2] The releases of Directory files containing identifying information about economical establishments, enterprises and other entities are not covered by the duty to keep statistical secret. (Art.20.3 Law 4/1986 of 23 April - Basque Statistics Law)

This is not a merely trivial issue. In fact, the establishment of a "safe" microdata structure requires the consideration of multiple "intruder" scenarios and many other aspects enumerated below:

- Type of statistics (census or sampling survey)
- Hierarchical structure of the data (i.e.: families and individuals)
- Geographical indicators included
- Identifying variables and possible combinations (identifying keys)
- Level of detail (number of categories) of the variables included
- Sensitive variables included (if any)
- Risk indicator
- Disclosure control methods to be applied
- Information loss measure (Utility measure)

Nevertheless, EUSTAT is in a good position to face this challenge and it has already developed some confidentiality rules for future microdata releases which will be the starting point of this complex task.

3.2 On-site access facility

An alternative to microdata accessibility consists of providing users (mainly researchers) with an 'in-situ' workstation where microdata could be accessed under specific conditions. Recently, EUSTAT has been asked about the possibility of accessing health data in order to perform multivariate analysis and develop a mathematical model to prevent child leukaemia. It is being considered as a pilot experience for a future on-site facility.

4 Conclusions

EUSTAT has been working for a long time on the implementation of a complete data protection system which considers all the phases of statistical production. In fact, a more general report has also been produced this year which considers the treatment of confidentiality in the whole production process, from data collection to physical protection measures and computer security.

However, in this paper we have focused on the dissemination phase and the development of standard criteria to protect statistical products: microdata and tabular data. Reaching an agreement about these confidentiality rules has been a hard process, and the discussion about what should be considered as identifiable or sensitive is still ongoing. However, these criteria should be in continuously updated, reflecting changes in the legal framework, in the technological environment and in social reality.

References

Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on *The Protection of Individuals with regard to the Processing of Personal Data and on the Free Movement of such data.*

Basque Statistics Office - EUSTAT (1999) *Statistical Data Protection Techniques*. Technical notebook.

Basque Statistics Office - EUSTAT (2007) *Treatment of Confidentiality in EUSTAT statistical operations*. Confidentiality protocol.

Garín, A., Urrutia, J., (2000). *Statistical Secret protection: basic elements of a data protection system*. OFISTAT Seminar.

National Institute of Statistics - INE (1994) . *Population and Households Census 1991: Methodology*. ISBN: 84-260-2889-6. Madrid.

Law 4/1986 of 23 April - *Basque Statistical Law.*

Law 15/1999 of 13 December - *Organic Law on Personal Data Protection.*

Statistical Programme Committee (2005) *European Statistics Code of Practice and Commission Recommendations.* Brussels.

Improving our knowledge of metaheuristic approaches for cell suppression problem

Andrea Toniolo Staggemeier[1], Alistair R. Clark[2], James Smith[3], and Jonathan Thompson[4]

[1] Information Management (Strategies), Office for National Statistics, Newport, United Kingdom, andrea.staggemeier@ons.gsi.gov.uk
[2] Principal Lecture in Operational Research at Computing, Engineering and Mathematical Sciences School, University of the West of England – Bristol, United Kingdom
[3] Reader at Computing, Engineering and Mathematical Sciences School, University of the West of England – Bristol, United Kingdom
[4] Principal Lecture in Operational Research at Maths Institute, University of Cardiff – Wales, United Kingdom

Abstract. This work will discuss further tests carried out using a pool of operational research and artificial intelligence techniques to solve the cell suppression problem. Existing solutions to the problem available through Tau-Argus software are mathematically demanding and only enable a solution to the problem for small table sizes. The approaches investigated here are pseudo-optimum in quality but enable handling of large size tables with complex structure. The test bed for this work used artificially created data which represent real-world scenarios found at ONS, and a sample of real data created from IDBR[1] sources. These data type are summarised as magnitude data, in both hierarchical and non-hierarchical formats, including 2 sets of sensitivity (2% and 10% sensitivity) and 2 sets of sparsity measures (5% and 25% of the table's cells contain zero values).
Among the approaches discussed in this paper are: a hybrid Evolutionary Algorithm,;Ant Colony Optimization; and Greedy Randomising Adaptive Search Procedure (GRASP). The table safety criterion was met using the Attacker Model by Salazar and Fischetti (2001). A relaxed feasibility criterion was also used on the Ant Colony and GRASP approaches in order to try to accelerate the evaluation process. Initial results showed that all approaches are able to handle larger data sets than existing mathematical programming routines. However a trade-off analysis between time taken to solve and data size indicated that we still have to improve the total time, perhaps not by using a single cell pass for table safety evaluation but multiple cells at a time.

Keywords: Tau-Argus, Statistical Disclosure Control, Cell Suppression, Mathematical Programming, Evolutionary Algorithms, Ant Colony Optimization, Greedy Randomising Adaptive Search Procedure (GRASP)

[1] Inter-Departmental Business Register

1 Introduction

Cell Suppression is just one of many Statistical Disclosure Control methodologies for protecting individual respondents when publishing tabular statistical outputs. This particular methodology lends itself better to magnitude data, but because of the perception some statisticians have when their outputs are modified in any form, e.g. health data experts, there was a desire to apply the methodology to frequency data also . The problem is to find an ideally optimal pattern of suppressed cells for which key sensitive cells remain non-disclosive. This is done such that the objective for good solutions is measured in terms of the total information loss caused by the required additional suppressed cells.

This paper will discuss some of the problems of the existing suppression implementations available in Tau-Argus, and describe some initial results found when using alternative solutions using traditional mathematical programming techniques.

1.1 Business Problem

There are currently four suppression methodologies available within Tau-Argus: Network flow; Hypercube; Modular; and Optimal. For further details on all the relevant methodologies available in Tau-Argus, refer to the manual[2]. The ONS has a variety of data that needs to be confidentialised, and since the existing methodologies were not suitable to all magnitude and frequency data types, the Neighbourhood Statistics programme, in collaboration with Methodology Directory and Information Management Group, started an investigation to understand the potential of alternative approaches from different areas of Operational Research (OR) and Artificial Intelligence (AI).
This paper will present a short description of each alternative solution implemented, and highlight the findings of initial investigations. It will also discuss some of the issues that we consider outstanding from a modelling and algorithmic perspective, and a discussion on how some of these problems might be addressed.

1.2 Comparison of Existing Methods in Tau-Argus

A direct comparison between the existing Tau-Argus cell suppression approaches is not possible, due the inconsistent way each of the solutions incorporates protection levels and objective function definitions. This is certainly an excellent point of improvement for Tau-Argus contributor's community. However, this would only be of interest if the aim of the tool was to provide a framework for researching different approaches for the same problem in an easy way. At present, the plug-and-play architecture of Tau-Argus is reflected in the way each of the Tau-Argus contributors have their own preference on programming language and style, mathematical

[2] http://neon.vb.cbs.nl/casc/Software/TauManualV3.2.pdf

programming technology, and levels of documentation available. This in turn results in difficulties determining when is best to use one approach over another.
Other work from ONS, proposing quality measures for confidentialised tables, also did not take into account the fact that different solutions have slight different way in approaching the definition of cell safety, and the quality of the suppression pattern used. For example, some methodologies report on the number of suppressed cells, and others on the total of weighted sum for a suppression pattern. The ONS work , however, assesses the risk of disclosiveness of a table and trade-off against statistical quality of the outputs.
In other words, although the end result is the same, they are achieved by different formulations of the problem, and unless it is clearly defined how one protection interval is compared to another, we are not able to compare the quality of the outputs produced.
During the evaluation of alternative solutions, researchers from the University of the West of England and Cardiff University were given the task to follow the same rules of protection as per the Optimal Cell suppression routine (Salazar, 2001), and develop alternative methods that were capable of dealing with larger instances of the data, and also subject to the protection levels as defined in the Optimal model.
A first phase of this work concluded that there were some issues in the way the current model was expressed which meant not all the experiments were completed in terms of understanding the capability of the proposed methods. For highlights on the issues found please refer to the Issues section of this paper.
A second phase was then commissioned to work within the restrictions of the model found in phase one, but focussing on the capability of the approaches rather than definitions of protection. The remit of each investigation was to compare results when using different parameters, different operators, the table safety check, i.e. reliability of the approaches, and the implications for each approach when running on hierarchical data sets. This paper focuses on the results obtained from phase two.

1.3 Proposed Solutions

Three concepts of the optimization algorithms in use need to be considered before the brief explanation of the approaches we developed. they are the notion of:
An optimal solution is described when the upper and lower bounds for a problem are meeting the same results, i.e. can't be further improved and very often referred to as a global optimum. Traditionally this technique uses mathematical modelling implementations such as Linear Programming and Mixed Integer Programming, which are time consuming in computational terms as the problem sizes increase. Optimal Suppression is an example of this type of solution approach.
A Heuristic solution, on the other hand, is a technique which does not involve mathematical proof for finding optimal results. It can be of two types; Constructive or Improvement; and works by searching through the solution space for a representation of the problem which satisfies its constraints. Heuristic techniques are

mostly described as quick search algorithms. An example of this type of approach is a greedy[3] algorithm to select candidate secondary suppressed cells.

A Metaheuristic technique, however, tries to overcome a potential local optimum solution from a starting solutions by adding learning inputs to the process. They also don't have a mathematical proof for optimality but often can deal with large datasets in reasonable computational times. Examples of metaheuristics applications in SDC are Tabu Search, and Simulating Annealing for the Controlled Rounding problem, by James P. Kelly.

Using Salazar definitions, the cell suppression problem involves ensuring that a table of cells is protected, so that certain cells, denoted primary cells, cannot have their values deduced from the published values in the table. Each cell has an associated **weight** and the objective is to find the set of cells to suppress to ensure that confidentiality is maintained, but minimising the sum of the **weights** of the suppressed cells. This is different to the **values** of the cells.

The algorithms that were evaluated in the first phase of investigation were:

(a) four variations of Greedy heuristic;
(b) local search algorithm (Descent method);
(c) Greedy Randomised Adaptive Search Procedure (GRASP);
(d) Ant Colony Optimization (direct and indirect models); and
(e) Evolutionary Algorithm.

Due to some of the imposed requirements for a solution to be available in a "reasonable" amount of time, an alternative feasibility check was created to speed-up some of the computational problems we encountered during phase one of the project in 2006. This is referred to as **relaxed feasibility check** as opposed to **strict feasibility check,** when following Salazar's incremental attacker model (2001), and were applied to options developed by ONS and Cardiff University.

The aims of phase two of the project were to provide more insights on how the approaches behaved when exposed to other settings of the algorithm, and to try to tune them for best results. For that a subset of the problems investigated in phase one were used. Table 1 describes the factors we were interested in analysing when changing the parameters, operators, verifying the approach for safety and for the hierarchical variable cases. In phase two only magnitude data was used because in phase one we revealed the methodological problem of the definition of protection when using frequency data (see section 4 in this paper for a summary).

Label	Number	Rows	Columns	Sensitive	% zero cells	Av. No Primary
A	5	200	5	10%	25%	60
B	5	200	5	2%	5%	60
C	5	200	50	10%	25%	553

[3] Greedy algorithm works by choosing the cheapest cell in a row/column which minimise the information loss whilst still guaranteeing protection of the primary cells.

D	5	200	50	2%	5%	526
E	5	4000	10	10%	25%	2387
F	5	4000	10	25%	5%	2201
G	2	654	14	19%	16%	1913
H1	1	14	1433	16%	14%	3680
H2	1	14	1433	16%	14%	3641
H3	1	712	10	6%	49%	407
H4	1	712	10	8%	49%	495
H5	1	712	19	11%	35%	1432
H6	1	712	19	13%	35%	1616

Figure 1: Table of artificial and a sample of real data created for this work when variables were non-hierarchical (A-G) and hierarchical (H1-H6).

2 Experiments Design

2.1 Parameter Optimisation

The methods used in phase one of the project depended on a number of parameters; for example, GRASP requires a candidate list size and number of cycles, and Ant Colony Optimisation requires an evaporation value, weights on the visibility, trails, and possibly a candidate size. The first set of experiments will focus on taking the best performing heuristics from the previous research, seeking to produce a more thorough parameter optimisation. In this way, we can ensure that the proposed methods are as efficient as possible.

From an Evolutionary Algorithm (EA) perspective, the first set of experiments was designed to determine whether there was any benefit to the use of a population-based approach as opposed to a simple local search method. A second goal was to determine the effect of changing the way in which solutions are perturbed by mutation (in the EA) or in the Local Search (LS) routine.

2.2 Operator Approach

For the GRASP approach, the algorithm was extended to consider the protection levels of the primary cells, meaning that cells in rows and columns that have been chosen to become secondary suppressed are no longer the ones that possessed the lowest cost, but also have to assure the protection limits are ensured.

For the EA approach, three different neighbourhood generation operators were used for the Local Search/Mutation steps, namely:

- Insertion: pick two random values in the permutation, and move the second to just behind the first, moving the intermediate elements along to accommodate the change;
- Swap: pick two random elements in the permutation and swap their positions; and

- Inversion: pick two random elements and invert the entire sub-permutation between them.

2.3 Table Safety

There were many difficulties in phase one of the project, particularly ensuring that a table was completely protected. Eventually, a working incremental attacker heuristic model implemented in a mathematical solver did enforce feasibility. This work will look again at the difference between a safe solution to the relaxed cell suppression problem, and a safe solution to the tight cell suppression problem. The intention would be to attempt to identify where the solutions to the relaxed problem are not feasible, and to see if the definition of the relaxed variant can be improved so that solutions to the relaxed problem are more likely to be truly protected. This required analysis of datasets, looking at the differences between the relaxed solution and tight solution, and working out means of reducing the gap between the two feasibility definitions.
One aspect that could have a dramatic effect on this work would be a simpler feasibility check.

2.4 Hierarchical Tables

The methods (a) to (e) considered in phase one of this project were not designed to work for hierarchical tables. For phase two, the current methods were adjusted to ensure they were sufficiently robust to deal with hierarchical and non-hierarchical datasets. An EA approach was implemented in a way that is transparent to the algorithm whether or not the data was hierarchical. However, better understanding of how the approach works under these circumstances will be the focus of attention.

3 Issues

Many issues with the Cell Suppression model have arisen from phase one, and others were further identified during phase two. This section highlights some of the findings and suggests points of further research we intend to pursue.

3.1 External bounds in attacker model and tight intervals set in Tau-Argus

Note that an attacker is assumed to know the values lb_i and ub_i of the lower and upper "external bounds". This may not be a realistic assumption. The values of lb_i and ub_i supplied in the Tau-Argus JJ-format file are currently specified as 0.5 and 1.5 times a cell's nominal value respectively. This is an issue that should be further considered.

3.2 Upper, Lower, and Sliding Protection levels set in Tau-Argus

Protection levels are only a feature of primary cells. However, if the levels are allowed to be defined as in the Fischetti & Salazar (2001) model, it may be possible

to identify a contributor to a primary cell due to the secondary chosen not pursuing insufficient boundary gap.
In other words, the "less/more than or equal to" inequalities in expression (3) from Fischetti and Salazar (2001) paper need to be replaced by "strictly less/more than" inequalities as in:

$$f^k_{i_k} < a_{i_k} - LPL_k \quad and \quad g^k_{i_k} > a_{i_k} + UPL_k \quad and \quad g^k_{i_k} - f^k_{i_k} > SPL_k \qquad (3^*)$$

thus consistent with Tau-Argus' Optimal Suppression protection limits. This is not a trivial distinction given that many table data and protection limit values tend to be small integers. The result is usually a distinctly larger set of secondarily suppressed cells when the table has many integer values, i.e., frequency tables and certain magnitude tables. The discovery of these problems in Fischetti & Salazar (2001) obliged us to modify our method accordingly and rerun experimental tests.
Knowing the external bounds lb_i and ub_i for all cells $i = 1,\ldots, n$ and which cells have been suppressed in the published table, an attacker will try to discover the minimum and maximum possible values, $f^k_{i_k}$ and $g^k_{i_k}$, of each sensitive cell i_k. The attacker can do this "*by solving a linear program in which the values* y_{ij} *for ... [specific] missing cells (i, j) are treated as unknowns*" (Fischetti & Salazar, 2001, page 1009).
For a given sensitive cell i_k, the minimum possible value $f^k_{i_k}$ can be found by solving the following linear programme (LP):

$$\begin{aligned} &\text{minimise} && y_{i_k} && (4)\\ &\text{such that} && \boldsymbol{M}\boldsymbol{y} = \boldsymbol{b} \\ & && lb_i \le y_i \le ub_i && \text{for all } i \in SUP \\ & && y_i = a_i && \text{for all } i \notin SUP \end{aligned}$$

Similarly, for a given sensitive cell i_k, the maximum possible value $g^k_{i_k}$ can be found by solving the same LP, but maximising y_{i_k}, *i.e.,* replacing objective function (4) by:

$$\text{maximise} \quad y_{i_k} \qquad (5)$$

Fischetti & Salazar (2001) state that the sensitive cell i_k is sufficiently protected if the solutions to (4) and (5) satisfy:

$$\min(y_{i_k}) \le LPL_k \qquad \text{and} \qquad UPL_k \le \max(y_{i_k}) \qquad (6)$$

However, to conform to the TauArgus Optimal Supression protection definition, the solutions to (4) and (5) should be strictly outside the interval [LPL_k, UPL_k]. In other words, rather than (6), we should require

$$\min(y_{i_k}) < LPL_k \qquad \text{and} \qquad UPL_k < \max(y_{i_k}) \qquad (6^*)$$

This is not a trivial distinction, given that table data values and protection limit values tend to be integers and often small.

Fischetti & Salazar (2001) state that if this condition is satisfied for all sensitive cells i_k then the whole table is feasible, i.e., sufficiently protected. However, given that the attacker will not know which of the suppressed cells are the sensitive ones, this condition should really be satisfied not just for each sensitive cell i_k, but also for each secondarily suppressed cell within the set SUP. If not, then the values of certain secondarily suppressed cells might be guessed, subverting the protection of the sensitive cell. This issue merits further investigation and research than was possible within the resources and time frame of the current project.
The Sliding Protection Level SPL_k was zero for all cells in all the JJ-format files supplied for testing purposes.

3.3 The Incremental Attacker Heuristic

Fischetti & Salazar (2001) state that their branch-and-cut (BC) approach finds an optimal set of secondarily suppressed cells that guarantees protection for all sensitive cells in a table. The approach is sophisticated, time-consuming, and identifies optimal solutions only for moderately sized tables. However, the authors do make use of a fast heuristic to find incumbent solutions at each node of the BC tree, based on a heuristic procedure from Kelly et al. (1992) and Robertson (1995). The heuristic starts by taking as input:
a given sequence of all the sensitive cells $\{i_1, \ldots, i_p\}$ to be protected; this sequence is heuristically determined according to decreasing weight in Fischetti & Salazar (2001), but in our method it is the key decision, as it defines the solution space in our Evolutionary Algorithm. A set SUP of suppressed cells that is initially equal to the set sensitive cells $\{i_1, \ldots, i_p\}$
The set SUP of suppressed cells is then augmented by solving a series of Linear Programmes (LPs), two per sensitive cell i_k in the order of the given sequence. The LPs use the cell weights, consistency equations, upper & lower bounds, and upper & lower protection limits provided by the JJ files output by Tau-Argus. Note that this does not necessarily minimise the number of secondarily suppressed cells in SUP, but rather their total weight.
The first LP, known as the UPL *incremental attacker problem*, identifies which cells need to be added to the set SUP in order to guarantee that the sensitive cell i_k is protected with respect to its upper protection limit UPL_k. For a given sensitive cell i_k, the LP is:

$$\text{minimise} \quad \sum_{t=1}^{n} c_t (y_t^+ + y_t^-) \tag{7}$$

$$\text{such that} \quad M(y^+ - y^-) = b \tag{8}$$

$$0 \le y_t^+ \le UB_t \qquad \text{for all } t = 1,\ldots,n \tag{9}$$

$$0 \le y_t^- \le LB_t \qquad \text{for all } t = 1,\ldots,n \tag{10}$$

$$y_{i_k}^- = 0 \quad \text{and} \quad y_{i_k}^+ = UPL_k \tag{11}$$

where $y_i = a_i + y_i^+ - y_i^-$ is the attacker's estimate of the value of sensitive cell $i \in \{1,\ldots, n\}$ so that the non-negative decision variables y_i^+ and y_i^- are respectively the deviations above and below of y_i from the cell value a_i. $UB_i = ub_i - a_i \ge 0$ is the relative external upper bound on y_i^+. $LB_i = a_i - lb_i \ge 0$ is the relative external lower bound on y_i^-. The objective function coefficient $c_i = 0$ for all $i \in$ SUP and $c_i =$ cell weight w_i for all $i \notin$ SUP.

After solving LP (7)-(11), the set SUP is augmented with all cells $i \notin$ SUP for which $y_i^+ + y_i^- > 0$ in the optimal solution.

Setting $c_i = 0$ for the set SUP's newly added cells i resulting from the solution of (7)-(11), the second LP similarly identifies which cells need to be added to SUP so that sensitive cell i_k is protected with respect to its lower protection limit LPL_k. This LP constitutes expressions (7)-(10), but with (11) replaced by:

$$y_{i_k}^+ = 0 \quad \text{and} \quad y_{i_k}^- = LPL_k \tag{12}$$

Fischetti & Salazar (2001) state that: "this guarantees the fulfilment of the upper/lower protection level requirement for i_k with respect to the new set SUP of suppressions." However, our experimental tests found exceptions to this statement. It was agreed that, from ONS's perspective, the upper and lower protection limits, UPL_k and LPL_k, have to be strictly obeyed, i.e., < and > rather than <= and >=, contrary to Fischetti and Salazar (2001) and as discussed in section 3.2.1 above. This meant that expressions (11) and (12) were respectively replaced by

$$y_{i_k}^- = 0 \quad \text{and} \quad y_{i_k}^+ = UPL_k + 1 \tag{13}$$

$$y_{i_k}^+ = 0 \quad \text{and} \quad y_{i_k}^- = LPL_k + 1 \tag{14}$$

For tables with integer cell values this generally resulted in:

1. a substantial increase in the number of secondarily suppressed cells.
2. sufficient protection for the primarily sensitive cells, as defined by expression (6*)
3. occasionally insufficient protection for the secondarily suppressed cells, as defined by expression (6*)

With respect to this last observation (#3), it is possible that some or all of the insufficiently protected secondarily suppressed cells are redundant, (i.e., not needed for primary protection). This merits further investigation beyond the scope of the current project, and is an issue on which we would like to continue to collaborate.

Arising from phase one issues, a major concern towards the end of phase two was the behaviour of the system in that it was not clear whether tables were being adequately protected. Typically it was noted that for the best solutions found, the min and/or max attacker problems would be reported as "infeasible" for several of the primary cells while the suppression set was being incrementally built up. Also, when the problems were re-solved to check for protection using the complete suppression set, most, or indeed all, of these problems would have disappeared.

During detailed discussions it became apparent that exactly the same behaviour was being observed with the Dash Xpress-MP version used by Cardiff University, which was implemented completely independently by ONS and Cardiff University. This may be to do with the way in which the models were specified in the original Fischetti and Salazar paper, although the reasons would seem to be rather subtle.

Given the success of the joint ONS-UWE bid for an EPSRC three year CASE studentship to study this issue, it was decided that it would be more valuable to spend the remaining allocated time considering further improvements to the way in which the each of the (a) to (e) approaches worked.

At the end of the first project it was suggested that it might be worth amending the constructive heuristic, so that instead of incrementally the primary cells one-by-one and generating new suppression sets, it might be possible to treat the primary cells in groupings of some form.

It was also noted during the analysis of the initial results that on these problems the EA evaluated both the row-order heuristic, and the weight-ordered heuristic, and that the latter never gave the best results found. This suggests that there may be some merit to treating together groups of cells belonging to a common marginal total.

This idea has been discussed in some detail and it was pointed out that although this might potentially greatly reduce the number of max/min attacker Linear Programming problems to be solved, the complexity of each one would increase which might make the overall run-time little different. It was agreed that UWE would investigate the feasibility of this approach, but that it would be considered supplementary to the original specification since it might involve considerable modifications to the way that the problems were specified.

4 Conclusion

This paper stresses the importance of keeping a close link between ISIs, NSIs and universities so that creative thinking is applied to the challenges of large tables with multiple hierarchies and varying densities of zeros and sensitive cells. All this work is being developed in close partnership with two UK universities, namely the University of the West of England (UWE, Bristol-UK) and Cardiff University. Dr. Alistair R. Clark and Dr. James Smith (both from UWE-Bristol) lead the work on

Evolutionary Algorithms (Clark and Smith, 2006 and 2007) and Dr. Jonathan Thompson (Cardiff) leads the work on Ant Colony Optimization and GRASP algorithms (Thompson 2006 and 2007).

The experiments have shown that GRASP is the preferred solution method, as Ant Colony Optimisation requires too much time for learning to take place. Even GRASP had to be simplified for run times to be reasonable for the larger datasets.

Various parameters have been considered and it was shown that performing additional cycles was unlikely to significantly improve solution quality. The results from the GRASP method were then assessed by the Incremental Attacker model (Salazar 2001) and here, there is little that can be done to improve the run times. On the smaller datasets, the run times are well within the desired times, and indeed, it has been shown that several solutions can be assessed by the incremental attacker model in a relatively short time. However the larger datasets are different, as just assessing one solution required up to 9 hours of run time. It is difficult to assess solution quality without knowing the optimal results but they appear to be encouraging.

This work has also produced a solution method for hierarchical datasets, similar to the GRASP method for non-hierarchical datasets. This again worked well but required considerable run times.

There are considerable gaps in some instances between the solutions to the relaxed problems and to the real problem, however in many cases the solutions generated by GRASP were already feasible and the incremental attacker model did not add any further suppressions.

On many tables, the EAs find solutions with between 75-90% of the cost of the heuristic solutions. In some cases the cost is only 28% of the heuristic cost. On most types of tables one of the EA-based approaches gives the lowest mean and minimum cost. On most types of tables the Inversion mutation operator gives the lowest mean and minimum results.

Analysis shows that on tables such as 14x1433 and 712x12 where Local Search algorithms are more effective, the EAs are stopping because of the inbuilt convergence threshold. Given that the LS algorithms often find better solutions after a large number of unsuccessful attempts, this suggests that this parameter has been set to terminate the EA too quickly.

Statistical analysis by ANOVA suggests that the Local Search is marginally preferable to a population size of ten, and that the Swap operator is best. However the overwhelming factor is the observed difference in results comes from the choice of "seed" used to create the tables. Thus, for example, in some cases (200x5, 200x50, 4000x10) there is considerable difference between the minimum costs tables for different instances (seeds), and the number of runs for each method may be different, so comparisons based on variance and absolute costs must be treated with a certain amount of caution.

For the 4000 x 10 tables with more sensitive cells, the algorithms do not have time to evaluate sufficient solutions to find major improvements, but even so cost reductions of between 8% and 24% are observed with Local Search.

This project was intended to assess the viability of the approaches to cell suppression developed in the first phase of the project. To that end we have conducted an extensive and highly computationally intensive set of experiments, the results of which have been described above.

In terms of the quality of the results obtained we have demonstrated that both the Local Search and Evolutionary Algorithm approaches are able to systematically improve on the quality of the solutions provided by the initial heuristics used. In some cases the improvements the improvement is dramatic – for example cost savings of up to 72% have been reported.

We have further analysed the differences between the two approaches tested, and reported on the combination of settings that gives the best results across the fairly broad set of problems used, namely a "steady state" Genetic Algorithm with population size 10, inversion mutation, and termination of runs if the population has remained converged for 5000 iterations.

However, these results have clearly demonstrated that for the larger tables with many sensitive cells, using a constructive heuristic to build a suppression set by treating each primary cell in turn is not a time-effective approach. While improvements of up to 24% were still obtained with the Local Search, each solution took on average 20 minutes to evaluate, which is not promising as a scalable approach unless significant computational resources are available.

While not in the original scope of this project, we have developed an alternative "grouping" approach which considers a whole row or column at once. This has been implemented and initial results show that the scalability issue seems to be largely solved. The benefit of this approach is that it requires absolutely no modification to the way that the Evolutionary Algorithm functions, and should not affect the validity of the findings contained in this report concerning parameter settings.

There remains one aspect that it was originally intended to consider, and which was not possible. This was an analysis of the degree of protection afforded by the evolved solutions. As discussed in Section 4, results obtained by both UWE and Cardiff University showed that there appear to be further problems with the formulation of the min/max attacker problems within the constructive heuristic. Consideration of these issues would have taken considerably more time than was budgeted for, with no guarantee of successful resolution. Therefore in conjunction with ONS it was decided to leave this issue for further work. We are of course pleased to report that we have obtained funding from the Engineering and Physical Sciences Research Council for a three year project to focus specifically on this issue.

5 Acknowledgements

The authors would like to thank the Statistical Disclosure Control Branch in Methodology Directorate and Neighbourhood Statistics Programme at the ONS, as

well as the Scottish Executive for their collaboration and input throughout the project, specially to Neighbourhood Statistics and the Scottish Executive for funding phase one of the project and the Information Management Group at ONS for funding phase two of this project.

6 References

Clark, A. R and Smith, J. EA and the Cell Suppression Problem – ONS Internal report – phase 1, (2006)
Clark, A. R and Smith, J. Further Experiments to investigate the Cell Suppression Problem – ONS Internal report – phase 2, (2007)
Fischetti, M. and Salazar, J.J. The cell suppression problem on tabular data with linear constraints, Management Science 47, 7, (2001).
Shlomo, N. and Young, C. Quality Measures for Statistical Disclosure Controlled Data, Proceedings of Q2006 - European Conference on Quality in Survey Statistics, 2006. See http://www.statistics.gov.uk/events/q2006/downloads/W21_Shlomo.doc
Thompson, J, Metaheuristics for Cell Suppression Problem – ONS Internal report – phase 1, (2006)
Thompson, J, Further experiments to investigate the Cell Suppression Problem – ONS Internal report – phase 2, (2007)

The Review of the Dissemination of Health Statistics in England

Jane Longhurst, Carole Abrahams, Ann Blake, Nirupa Dattani (ONS)*
Mary Grinsted (Department of Health)**
Gwyneth Thomas (Welsh Assembly Government)***

* Office for National Statistics, Segensworth Road, Fareham, UK, Jane.Longhurst@ons.gov.uk
** Department of Health, Skipton House, 80 London Road, London, Mary.Grinsted@dh.gsi.gov.uk
*** Welsh Assembly Government, Cathays Park, Cardiff, Gwyneth.Thomas@wales.gsi.gov.uk

1. Introduction

Health statistics support a wide range of work to improve and protect our health, they inform patients and the public. Many of these areas of work require detailed figures, which may raise issues about data confidentiality. Producers of health statistics must ensure that their statistics meet the needs of users while at the same time protecting confidentiality. The Review of the Dissemination of Health Statistics in England was initiated in 2005 to address disclosure issues around health statistics. The aim of the review was to produce guidance for handling health statistics in a way that ensures the public interest in the figures is met while managing data confidentiality risks.

The review was led by the Office for National Statistics (ONS) and involved representatives from the Health Departments in England, Public Health Observatories and the devolved administrations (Wales, Scotland and Northern Ireland). The approach adopted for the review was a two-stage process. The first part of the review focused on developing guidance for published tables of abortion statistics. Specific guidance for these outputs was released in July 2005, ONS (2005) and has been subsequently implemented by the Department of Health. The scope was then extended to provide more general guidance on disclosure issues for all published tables of health statistics. Throughout the development of the guidance key stakeholders were consulted via a series of workshops and in addition the guidance was released for public consultation. Following quality assurance and approval from the National Statistician and Health Minister the final guidance from the review was published in October 2006 on the National Statistics website, ONS (2006).

This paper provides an overview of the final guidance in Section 2. Section 3 describes work that is being undertaken to support the implementation of the guidance across the health domain. Sections 4, 5 and 6 detail three specific examples of practical implementation of the guidance from the Department of Health, the Welsh Assembly Government and ONS.

2. The Guidance

2.1. Scope

The principles and approach outlined in the guidance apply to all health statistics. However, the review is focused on tables derived from registration processes, administrative sources and statistical returns. It does not deal with confidentiality issues concerned with record-level information. The guidelines replace previous practices that had been adopted within the health field, such as the rule of thumb to suppress all values in tables less than 5.

The review was established specifically for published health statistics, where following release there is no control over their further use. The guidelines should be used to protect statistics released as part of a production process, however, ad-hoc releases and in particular Freedom of Information (FoI) requests are also within scope. Throughout the review it was therefore necessary to consider the implications of FoI and the guidelines were developed taking into consideration what it is or is not appropriate to withhold under the Act. Since the original publication of the Guidance, the Statistics and Registration Services Act 2007 (SRSA) has been passed by Parliament. This comes into force on 1 April 2008 and introduces new legal considerations

2.2. Format of the Guidance

The final guidance has been published on the National Statistics website as seven pdf documents; a main document, five working papers and a summary of the responses to the public consultation. The main document describes an approach that data providers should follow based on a general framework for addressing the question of confidentiality protection (see section 2.3). No single solution or rule is recommended instead guidance is provided, based on the steps in the framework, on how to develop solutions for different datasets. Examples are used throughout the document and more technical advice is provided in the five working papers:

1. Legal and Policy Considerations.
2. Risk Assessment
3. Risk management
4. Glossary
5. References and other Guidance

2.3. Framework for Confidentiality Protection

The guidance advises producers of statistics of six main steps to be taken in considering disclosure control in relation to tables of health data. The guidance works through each step, giving details, examples and useful references. The six steps are:

1. Determine users' requirements for the published statistics
2. Understand the key characteristics of the data
3. Are there circumstances where disclosure is likely to occur?
4. If so, would disclosure represent a breach of public trust, the law, or National Statistics policy?
5. If required, select appropriate disclosure control methods to manage this risk
6. Implement and disseminate.

2.3.1 Determine users' requirements for the published statistics

The first priority for producers of statistics should be that their publications meet the needs of users. It is therefore vital to identify the main users and understand why they need the figures and how they will use them. The disclosure protection used needs to have the least possible adverse impact on the usefulness of the statistics.

2.3.2 Understand the key characteristics of the data

It is important to have a good understanding of the data that may require protection to assess any risk of disclosure. Issues to consider include:

- the source of the data underlying the statistics
- sensitive variables

- the age of the data; older data may carry less risk of disclosure
- quality of the underlying data
- small groups of statistical units
- whether the data is event-based or residence-based

It is also important to consider the characteristics of the tables. Where tables are very simple and presented at a high level of aggregation (including geography), disclosure issues are unlikely to arise. When tables become more detailed, and the counts in individual cells are small, the risk of identification may increase and protection may be needed. If the spread of values is skewed across a table, the risk in particular cells may increase above an acceptable level. In addition, issues may arise with linked tables where risk of disclosure can increase by differencing or through combining with other data.

2.3.3 Are there circumstances where disclosure is likely to occur?

The answer to this question is the risk assessment. Risk is a function of likelihood (related to the design of the table), and impact of disclosure (related to the nature of the underlying data). Decisions on likelihood and impact should be made by those who have a detailed understanding of the statistics and experience of the interest in the figures. In order to be explicit about the disclosure risks to be managed one should consider a range of potentially disclosive situations and take action to prevent them. The situations should be used to identify those parts of the table that could lead to disclosure. Appropriate confidentiality rules should be applied to these cells. The guidance provides examples of disclosive situations but notes that it is not possible to protect against all risks and that this is a risk management rather than a risk elimination exercise.

In practice it is likely that producers of statistics will find that outputs can be placed into one of three broad risk categories and recommendations are made on the level of protection required for these three risk categories.

- Low Risk: For some health statistics the likelihood of an attempt at identification may be considered to be low if tables are disseminated at a high level of aggregation and only limited tables are produced from the one database, i.e. no risks from linking between current and future releases. A high level of aggregation reflects a reduction in disclosure risk as the size of the population of the statistic increases. Health statistics in this category will not usually require any protection beyond good table design. However, in order to prevent attribute disclosure care should be taken where rows or columns are dominated by zeros and in particular where a marginal total is a 1 or 2.
- Medium Risk: In order to ensure protection from disclosive situations for the majority of health statistics it will be sufficient to consider all cells of size 1 or 2 unsafe. Care should also be taken where a row or column is dominated by zeros.
- High Risk: For some health statistics the likelihood of an identification attempt will be higher, and the impact of any successful identification would be great, e.g. statistics on abortions, AIDS/HIV, STDs. In order to ensure protection all cells of size 1 to 4 are considered unsafe and care should be taken where a row or column is dominated by zeros. Higher levels of protection may be required for small geographical levels or for particular variables with an extremely high level of interest and impact.

The guidance outlines situations where these recommended levels of protection may need to be increased.

2.3.4 If so, would disclosure represent a breach of public trust, of the law, or of National Statistics policy?
When establishing whether confidentiality protection is required for a particular health statistic, it is necessary to consider public trust and cooperation, and legal rights and obligations, as well as national and international standards for statistics. More legal and policy considerations are provided in the relevant working paper.

2.3.5 If required, select appropriate disclosure control methods to manage this risk
An introduction is given to five main statistical disclosure control methods. Table redesign (grouping/collapsing categories, aggregating across geographies or time) is recommended as a simple method that will minimise the number of unsafe cells. If unsafe cells remain in the table, further protection methods such as rounding, cell suppression or cell perturbation (e.g. barnardisation) should be considered. If a data provider has access to the individual record level data then disclosure control methods can be implemented that adjust the data before tables are designed, e.g. record swapping. The different methods are compared and contrasted to assist the selection of the appropriate disclosure control tool. The guidance also recommends alternative methods for presenting the data, e.g. graphs, commentaries or analytical output, as an approach for providing users access to information without disclosing the underlying data.

2.3.6 Implement and disseminate.
The guidance will allow data providers to set disclosure control rules and select appropriate disclosure control methods to protect different types of health statistics that are to be published. The most important consideration is maintaining confidentiality but these decisions will also accommodate the need for clear, consistent and practical solutions that can be implemented within a reasonable time and using available resources. The methods used will balance the loss of information against the likelihood of individuals' information being disclosed.

The guidance has been developed taking into account the implications of the Freedom of Information (FoI) Act and therefore confidentiality policy developed using this guidance can be used to help decide which exemptions in the Act are relevant, and which should be cited when withholding confidential statistical information. However, FoI requests should always be considered on a case by case basis and there may be cases when decisions about a case are different to the general policy for the publication of statistics. This does not mean that the policy is wrong, since it has been developed for use in a production process. Whilst confidentiality must always be maintained, a decision made under FoI to provide information in a form different to the published outputs is compatible with this guidance.

Guidance is provided more generally on implementation, in particular the information that should be provided to users concerning the confidentiality rules and disclosure control methods.

3. Implementation

The guidance produced from The Review of the Dissemination of Health Statistics as outlined above is intended for anyone in the health community involved in the publication of health statistics. The Office for National Statistics (ONS) have given assurances that the

proposals in the guidance will be implemented for their outputs. In addition the ONS will work collaboratively with the Information Centre for Health and Social Care and the Department of Health to support the implementation of the recommendations more widely.

A project within ONS has been established to coordinate this implementation, the aim being to ensure that health statistics releases throughout England are consistent with the new confidentiality guidance on disclosure and to encourage the adoption of the guidance for the other UK countries. The project board includes representatives from ONS, the Health Departments, Public Health Observatories and the devolved administrations.

Each key producer of health statistics represented on the project board is drawing up and executing plans to implement the guidance in relation to their outputs from April 2007. It is not expected that the guidance will be applied retrospectively unless there is an exceptional reason to do so. Templates (see below, 3.1) have been developed to record basic information on outputs and timetables for release and more detailed information relating to the steps within the framework for confidentiality protection and proposals for disclosure control. Where statistics are produced by organisations outside the representation of the project board, these organisations will be encouraged to follow the guidance.

The project board is responsible for coordinating these high level implementation plans and timetables. In addition the board will provide expert advice and support for implementation and facilitate consistency and sharing of best practice.

3.1 Risk assessment templates

The Welsh Assembly representative on the project board provided the risk assessment template which had been developed and was being used in Wales (see section 5). This was slightly modified by the project board members representing ONS health statistics to produce two templates, a high-level one and a low-level one.

The high-level template is an Excel spreadsheet on which the data providers can enter details of each output which it is planned to publish during the year 2007/08. Work on populating these has begun. The information entered onto this template includes, for each output:

- Source of the data, e.g. “cancer registrations”; “calculated using life tables & census data”; “smoking cessation questionnaire”.
- Output type, e.g. HSQ (Health Statistics Quarterly) article; web release; quarterly statistics.
- Name of the output
- Area coverage
- Date of publication
- Key characteristics, e.g. “contains potentially sensitive information”; “may contain low numbers”; “low sensitivity”.
- Disclosure issues and risks – High, Medium or Low
- Impact of disclosure e.g. whether it would be a breach of public trust
- Statistical disclosure controls applied – Yes or No
- Link to low level report - this is an electronic link to the corresponding low-level template

The low-level template is a Word document, and there will be one for each planned publication. This template is based around the framework described in section 2.3. The details to be entered are:
Background to the data source

- Legal issues (collection and dissemination) - Reference to any relevant statutory arrangements relating to the data, e.g. population statistics act
- Key characteristics of the output - List of sensitive variables, age and quality of data, coverage, population base, possibility of linking to other published tables
- Evidence of risk in the output - A summary of the risk assessment including:
  - a note of disclosure scenarios considered
  - potentially unsafe cells
  - arrangements with data provider (e.g. if data are provided by another government department). Colleagues' views should also be recorded here.
- Proposals for mitigating risk in publishing output, including list of options
- Conclusions and details of disclosure control methods to be used: Which option was chosen and why, and description of methods to be used.
- Review process
-

The next three sections of the paper provide practical examples of how the guidance is being implemented for different outputs.

4. Abortion Statistics

The Department of Health (DH) undertakes the statistical processing and analyses of notifications of abortions for England and Wales. This includes the release and publication of statistics derived from the information contained within the notification. Abortion data is very sensitive and therefore the impact of any identification or disclosure from these statistics is considered to be high. Abortion data attract a lot of attention from the media, MPs and Peers, the public and lobby groups and is likely to be scrutinised closely, particularly at its margins.

DH receive a lot of requests for abortion data some of which are potentially disclosive, e.g. numbers to girls aged 11, 12 & 13 years old, medical conditions of late abortions. From experience of high profile cases in 2001/02 it was known that statistics like these, possibly used with other information, could be used to identify and target individuals. As a result 2002 data due to be published in 2003 was held back and only a skeleton publication was released. This limited publication and refusals to release requested information, resulted in DH being seen as overly cautious and accused of hiding information. Clearly guidance was needed in interpreting the National Statistics Code of Practice in order to balance the data confidentiality risks with the public interest. The first part of the Review of the Dissemination of Health Statistics focused on abortion statistics and attempted to address these concerns.

The guidance provided details on how to identify cells within tabulated statistics where the risks of a breach of confidentiality were unacceptable ("unsafe cells"). The risks within the publication were identified and were reduced largely by redesigning tables. However, where table redesign proved to be impossible then suppression was applied to cells with fewer than 5 cases at National level or fewer than 10 cases at sub-national level and to

highly sensitive variables such as gestation weeks in tables of terminations by medical grounds. The same principles were also applied to tables showing rates and percentages and to ad hoc requests for data.

The guidance is followed for all requests and for the vast majority it is unquestionably useful. It is especially useful to be able to point questioning customers towards the protocol on the internet and for them to know it is a health statistics wide issue rather than a data provider's decision. However, in a very few instances the information suppressed does seem overly cautious, more so to the customer than to the data provider, who understands that the rules work for the majority of cases.

4.1. Example - Abortions performed by the British Pregnancy Advisory Service

In order to illustrate the disclosure issues related to abortion statistics the following example of a Parliamentary Question is described. Through contractual arrangements with Primary Care Organisations, some approved independent sector places of termination perform NHS-funded abortions. The query related to the number of early medical abortions performed by the British Pregnancy Advisory Service (BPAS) at the request of the National Health Service (NHS) in each of the last five years, broken down by (a) age of the woman, (b) gestation of the pregnancy and (c) region.

There are three key items relating to an individual abortion which need to be protected. These are: the details of the woman whose pregnancy was terminated, the identity of the practitioner who carried out the termination, and the identity of the hospital or clinic where the abortion took place.

Therefore in answering the query the following were considered:

- Confidentiality of the patient - control for small numbers, e.g. counts less than 10.
- Confidentiality of the doctor - check original documents for doctors' names to make sure there were at least three doctors performing terminations at any one clinic.
- Confidentiality of clinic - check that there was more than one BPAS clinic per region.
- Confidentiality of the data - make sure no similar data had been previously published that could be used deliberately or inadvertently to disclose small numbers.

An extract from one example table from the release is shown below, suppressed cells are indicated by '..':

Table of counts of abortions performed under 8 weeks at BPAS clinics

Region	2002	2003	2004
East	17	43	..
East Midlands	..	..	78
London	249	362	582
North West	52	74	..
North East	..	..	97
South East	409	507	913
South West	44	85	192
West Midlands	376	348	597
Yorkshire and Humber	83	173	171
Total	**1252**	**1636**	**2683**

The data was grouped within the tables in various ways in order to maximise the amount of information provided and minimise the number of cells that had to be suppressed. The data was compared with similar data extracts for clinics run by other agencies in case an equivalent request was made for these other agencies. Data also had to be checked for disclosive cells in groups that could be derived from other published data, e.g. privately funded abortions, NHS hospital terminations within the same regions and more than 9 weeks' gestation. In some cases small counts were suppressed and in order to protect these counts further, secondary suppressions of counts greater than 10 may have also been suppressed. Care will need to be taken with future releases to ensure that these counts and other totals are not revealed, otherwise the protection for this release could be compromised.

5. Implementation of the Guidance by the Welsh Assembly Government

During the time when the Health Statistics confidentiality guidelines were being prepared, the Statistical Directorate of the Welsh Assembly Government was also considering statistical disclosure issues in health statistics and across the range of subject areas covered in the Directorate. Internal guidance has been written and is being used, based largely on the Health Statistics guidance but widened in scope to cover all statistical areas dealt with by the Directorate.

A key part of the guidance is the process of assessing and documenting risk and, as an aid to this process, a risk assessment template has been devised in consultation with staff. It is felt to be important for two main reasons. Firstly so that consistent decisions about disclosure control can be taken for each dataset in advance of the receipt of ad-hoc requests; and secondly, so that there is documentary evidence that disclosure risk has been considered. The risk assessment details evidence of risk in the dataset based on the kind of issues described in the Health Statistics guidance, such as, sensitivity, geography and denominators, low cell counts, zero value cells, age and quality of the data and so on. Users of the template are encouraged to seek the views of colleagues working in a practical way in the subject area, look for examples of disclosure control in similar or related datasets and to think about scenarios where disclosure might occur. Options for mitigating risk are considered and conclusions drawn about future practice and methodology for disclosure control if neces-

sary. Because they may detail intended methods for disclosure control, risk assessments are intended to be internal documents only. As a public acknowledgement of the issue the Directorate is planning to add a standard phrase to publications stating that statistical disclosure risk has been considered, risk is felt to be low/medium/high and as a result data has been modified/not modified.

The template is being gradually introduced for all datasets in the Statistical Directorate. It has been employed so far in such diverse areas as health statistics and economic statistics. The format has worked best where the responsible statistician is also the data collector, where it provides a useful aide memoir to the thought process of considering the risk of disclosure. It is less useful where data dissemination is already governed by the rules of a third party data provider. The guidance used in the Directorate together with the risk assessment template will be subject to a review during the next few months. For health statistics it is working well and following a workshop on disclosure control organised by the Welsh National Public Health Service and the Welsh Health Analysts' Network, partner organisations in Wales have expressed an interest in utilising it as a basis for their own statistical disclosure control decisions.

5.1 Example - Community Contraception Statistics

As an example of how the risk assessment template has been used, the publication of community contraception statistics is considered. Data is collected annually from NHS Trusts in an electronic form. The dataset is aggregate data and relates to the numbers of first face-to-face contacts of patients with the service in the financial year by age (individual year of age for women aged less than 20), reasons for attending the clinic and method of contraception chosen.

The dataset contains items which may be sensitive in some cases, for example, the contraception methods chosen by young girls. Low cells and real zeros may be disclosive where many dimensions are tabulated but disclosure is unlikely otherwise. It is possible that potential intruders have access to local information which would help them find, say, individuals who have chosen unusual contraception methods given their age but it is not clear what that information might be since clinics would not break their patients' confidentiality.

A degree of uncertainty is present in this dataset since the coverage is only a proportion of all contraception consultations; it only relates to community clinics and excludes consultations with GPs and private clinics. Also, the data relates only to first contacts in the financial year which again means that not all visits are included introducing further uncertainty. The lowest level of disaggregation is NHS Trusts and there is no possibility of differencing.

After consideration and discussion with NHS professionals, it was felt that where two dimensions are involved e.g. method and age or NHS Trust and age, risk is fairly low but where all three are involved the risk increases with a determined intruder. Thus it was concluded that in routine and ad-hoc tabulations only 2 of the 3 possible dimensions would be used. Trusted users might be allowed access to more detailed data but only having signed a confidentiality agreement.

6. Disclosure control for health data with rare events: Development of a method for presenting conceptions to girls aged under 18 by small area

As outlined in Section 3 the ONS will be reviewing all outputs of health data as part of the implementation of this guidance. One area where work has already been carried out is for conception statistics. In order to meet a requirement for information to support action at local level towards achieving the Public Service Agreement target to reduce the under 18 conception rate by 50 per cent by 2010 a project was undertaken to produce an output presenting under 18 conceptions by ward.

There are around 40,000 conceptions to girls aged under 18 in England and Wales each year. But the number of cases at ward level is relatively low. Therefore a method of presenting these data has been developed which will provide useful information by small area whilst protecting confidentiality of individuals.

Due to the sensitive nature of the data, the small area of the geography, and the focused nature of the age group of interest the methodology combines three statistical disclosure control methods.

First the data for three years were combined. This served two purposes:

- It smoothed the data and thereby reduced the impact that natural variation in rare events can have on understanding trends over time
- It increased the ambiguity of small numbers so that it is not possible for a potential intruder to identify individuals

The data for England and Wales were then divided into five bands (quintiles). Wards with the lowest rates were allocated to quintile 1 and those with the highest rates to quintile 5. Data for wards with a population of fewer than 30 girls in the 15 to 17 age group were then suppressed. The upper and lower limits for each band were examined to assess whether publication of these, in combination with the availability of population estimates for the group of interest, may allow an intruder to unpick the data and calculate actual numbers of teenage conceptions in a particular ward. It was established that this may be a possibility, and therefore only the lower limit for the wards with the highest rates will be made available.

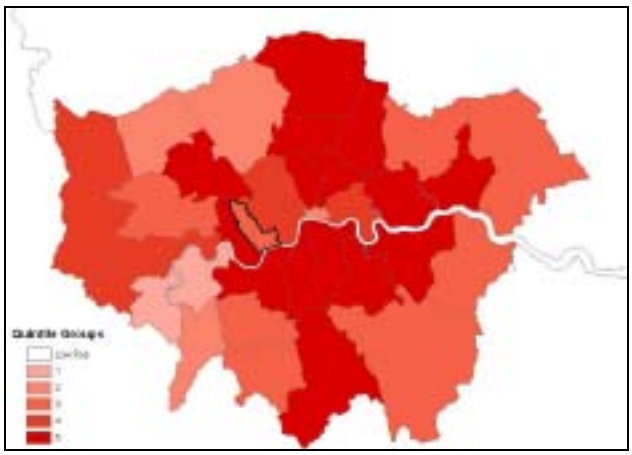
Users would be able to identify which quintile a particular ward falls into, and also make a comparison between wards in the country using the Neighbourhood Statistics mapping functionality on the ONS website. In this way the under 18 conceptions rates for individual wards are not published, but because they can see the geographic distribution of the data, users are still able to identify areas with high rates that they may wish to target. Further information on these data can be obtained from ONS (2007).

Using Kensington and Chelsea the following example shows how the ward and local authority (LA) level maps provide extra information about under 18 conception rates within a given area.

The first map provides a view of LAs within the London Government Office Region for 2001 - 2003. The map allows comparisons of rates between LAs and also shows LAs with

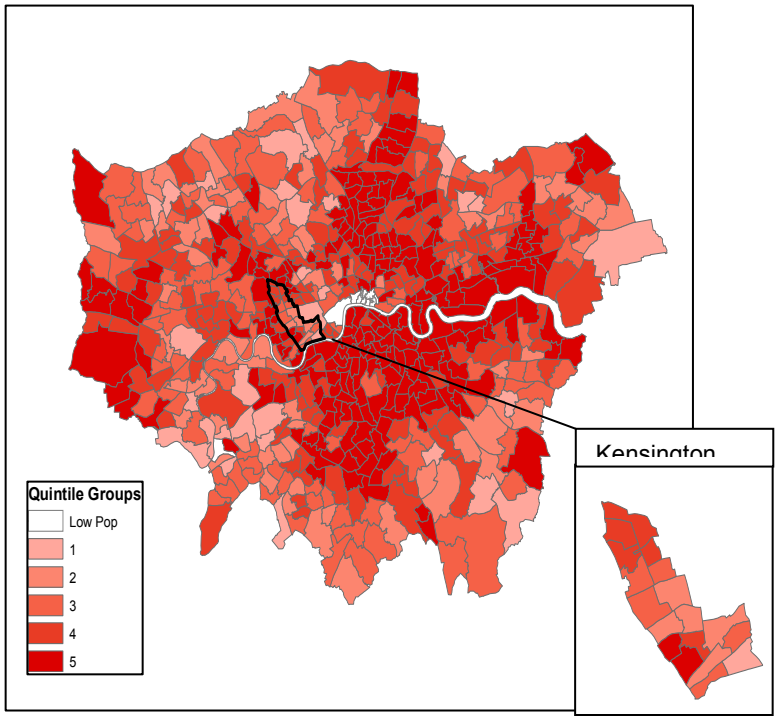
high rates of under 18 conceptions, i.e. those in quintile group 5. For Kensington and Chelsea (as highlighted) we can see that this LA is in quintile 3. Therefore the rate for this LA is in the middle quintile.

Under 18 conceptions data for Local Authorities in London, Jan 2001 – Dec 2003



The map below shows the information at ward level for the London region, for 2001-2003. This shows that there is significant variation in under 18 conception rates by ward within some local authorities. We can now see that there are also wards within Kensington and Chelsea with the highest rates of under 18 conceptions.

Under 18 conceptions data for wards in London, Jan 2001 – Dec 2003



7. Conclusion

In using and releasing health statistics there is a risk, generally with small numbers, of identifying individuals. To address this, the Department of Health in England asked the National Statistician to provide it with guidelines for disseminating health statistics, in a way that balances data confidentiality risks with the public interest in the use of the figures. Guidance has been developed based on a framework for addressing issues concerning confidentiality. No single solution or rule is recommended; instead data providers are encouraged to develop solutions for different sets of statistics based on the steps in the framework.

Work is being undertaken to implement the guidance. In particular the Welsh Assembly Government has developed templates to aid in the risk assessment process and to document the decisions made. The Department of Health are using specific guidance developed for abortion statistics for all annual and ad-hoc releases. The ONS has also developed new ways to release conception statistics without disclosing the underlying small counts. These are just three specific examples. Work on implementation is being coordinated by the Implementation Project Board to ensure consistency and encourage best practice.

References

ONS (2005) Disclosure Review for Health Statistics, 1st report, guidance for abortion statistics,
http://www.statistics.gov.uk/downloads/theme_health/abortion_stag_final.pdf)
ONS (2006) Review of the Dissemination of Health Statistics: Confidentiality Guidance, http://www.statistics.gov.uk/about/Consultations/disclosure.asp.
ONS (2007) Conceptions, Under 18's: Local Analysis
http://neighbourhood.statistics.gov.uk/dissemination/Info.do?page=news/newsitems/6-august-2007-conceptions---under-18s-local-analysis.htm
Statistics and Registration Services Act 2007,
http://www.opsi.gov.uk/acts/acts2007/ukpga_20070018_en_1

Disclosure detection in research environments in practice

Felix Ritchie*

* Office for National Statistics, Cardiff Road, Newport, South Wales NP10 8XG
Email: felix.ritchie@ons.gov.uk

Abstract: There is an increasing demand for access to raw confidential data, and NSIs have responded by setting up controlled research facilities. However, the most common approaches to statistical disclosure detection and control (SDDC) struggle to accommodate the infinite variety of outputs produced in research environments. The main problems are designing statistical disclosure control (SDC) rules for unknown transformations of the data, and in managing the potential volume of outputs needing review.

Research facilities need a different approach to SDDC. In the UK, ONS has developed an approach based around classes of output, where the time devoted to checking outputs can be concentrated on the more unsafe outputs.

Defining "safe" and "unsafe" outputs based on the functional form of the model improves the efficiency and security of confidentiality checking, but is not straightforward. This paper outlines the broad approach, and then takes specific examples to show how the UK rules on analytical outputs (and the conditions attached to them) have been developed.

1 Disclosure control in research environments

After falling out of popularity, in recent years there has been an increase in the provision of Research Data Centres (RDCs) and other research facilities by National Statistics Institutes (NSIs). These pose problems for disclosure control. RDCs are designed to be places where experts have access to very detailed data; they select, twist, transform and link it in interesting and different new ways; and they produce complex outputs which need to be assessed for disclosiveness.

A disclosure control system should

- be transparent
- be consistent
- guarantee a level of disclosure risk
- not unduly restrict research output

As noted in Ritchie (2007), automatic disclosure control and hard-and-fast rules do not really provide this for RDC outputs. Hence all NSIs operate manual disclosure checking for their RDCs, and have guidelines for NSI staff and researchers (see, for example, Enright et al (2006), or the NORC/NIST website at dataenclave.norc.org). However, the potentially infinite range of outputs is a problem: how can any set of

guidelines cover all types of outputs in enough detail to be secure, with enough flexibility to be useful, and with enough consistency to be fair?

Ritchie (2007) proposed that grouping outputs into certain classes would go a long way towards making a feasible RDC checking system. He noted that the Office for National Statistics (ONS) in the UK was already developing such a classification for outputs from its Virtual Microdata Laboratory.

The aim of this paper is to shed light on how some of the concepts raised in the earlier paper can be used in practice. In particular, it shows

- how definitions of safe and unsafe outputs can be turned into rules
- how those definitions need to be based upon functional form and not data
- some of the steps to define an effective classification for an output

It also shows how the most popular SDC guidelines fit into this model.

2 Classifying the research zoo

As noted in Ritchie (2007), designing a disclosure control mechanism for a research environment is like designing a zoo. There may be uncertainty about the specific animals, but it should be possible to classify the various animals into groups: those that swim, those that fly, those that need water, those that eat unwary keepers. A herpetarium may not be designed with any specific snake, or species of snake, in mind, but should be able to effectively contain and keep healthy most of the snakes the zoo intends to stock.

These types can be allocated to broad classifications of "safe" and "unsafe". The "safe" animals do not pose a significant danger to themselves or others. Hence, the zoo owner can then concentrate more time on the unsafe ones, in an efficient distribution of resources.

In terms of research outputs, "safe" and "unsafe" have clear interpretations for both researchers and NSI staff

- **Safe outputs**: these **will** be released **unless** the NSI staff can see some reason why they should be held back or adjusted.
- **Unsafe outputs**: these **will not** be released **unless** the researcher can demonstrate to NSI staff that the output meets the detailed criteria for this output

Note that the burden of proof shifts depending upon whether safe or unsafe outputs are being discussed.

For a safe output, the NSI team have decided that a certain class of output holds no disclosure risk in general. They may have concerns about a specific output which is an exception to the general rule. To enable the system to work well, these exceptions should be

- Small in number
- Well defined
- Comprehensible to and communicated to the researchers before research begins

The third bullet is essential. Developing an effective SDDC system for a research environment requires a positive relationship between researchers and NSI staff; there should be no surprises on either side. By clearly specifying exceptions, the researcher can be confident that the results produced will be acceptable for release.

For unsafe outputs, the NSI has decided that scope for disclosiveness in the output is such that it is, in general, unprepared to release the output. However, it leaves the door open for the researcher to argue a case as to why the decision should be changed.

Clearly, a researcher arguing that an unsafe output can be released needs to have a good awareness of the principles of disclosure control as well as the specific data and context. Hence, the researcher training necessary for effective SDDC should be

- focusing on and discouraging these unsafe outputs
- illustrating what can turn an unsafe output into a safe one
- informing researchers as to what needs to be demonstrated to make an unsafe outputs safe

3 Determining "safety"

For historical reasons, the SDC literature focuses on specific aspects of the data being released: dominance, outliers, use of public information etc. This is because the vast bulk of work on SDC has gone into making sure that either datasets have been anonymised effectively, or that aggregate tables are safe.

The use of these methods in research environments is inappropriate. The standard techniques are designed for a fixed input dataset and a finite set of outputs, against which a range of intruder scenarios can be tested. In research environments the input dataset and output data are not known when the SDC rules are being drawn up; and it is not practical to provide the same sort of detailed analysis of each output as is done for aggregate finite results.

The key to determining the "safety" of a dataset is to study the underlying functional form of an output. If there is no disclosure risk in an arbitrary dataset, then there cannot be any additional risk from having, for example, an "identifying" set of variables.

Note that this technique also helps to narrow down precisely where the risk arises. For example, in the linear regression model, the apparent risk arises from the coterminous publication of means and frequencies. Hence the linear regression model itself is safe, but supporting statistics may be problematic.

Each output type needs to be assessed for primary disclosure – that is, whether something can be inferred directly from the single output – as well as disclosure by differencing. Assessment should consider both cardinal and categorical variables.

It may not be easy to classify results. If something is fundamentally safe but has a large number of exceptions, it may be better to classify it as unsafe. For example, Table 1 shows examples of current classifications used in the ONS Virtual Microdata Laboratory (VML):

Safe	**Unsafe**	**Uncertain**
General linear regression[1]	Tables	General non-linear aggregations of data
Panel regression	Graphs	
Herfindahl indices[1]	Quantiles	Large high-frequency aggregate tables
Covariance matrices[1]	Cross-product matrices	

1 Restrictions apply; see below

Table 1 Examples of safe and unsafe outputs at ONS

Most of the "safe" outputs have further restrictions. These are where the exceptions come from, which the NSI uses to decide whether the output can be released. These are, as noted, limited in number and made known to the researcher. If those two conditions cannot be met, then the output would have been classified as "unsafe" or, at best, "uncertain".

The "uncertain" elements here arise from several factors. It may be that the model hasn't been studied yet; or that there is no simple statement of the exceptions for a safe output; or that there is no agreement yet on how to demonstrate safety in a way which is not labour intensive.

Before studying practical examples, two further considerations are needed. First, it is clear that, given a specific functional form, in theory a specific combination of data always exists that would allow a data point to be identified. A "safe" output is one where this theoretical possibility has no practical counterpart in analysis.

Following on, it needs to be assumed that the outputs are genuine statistical outputs. A malignant researcher could construct a statistic which appears a valid statistical result, but which has in fact been constructed simply to avoid detection. Dealing with deliberate cheating is outside the scope of this paper.

4 Examples

In this section we investigate some specific assessments of outputs . Only a selection of outputs is covered, to illustrate different aspects of the method. Further examples can be found or referenced in VML(2007).

4.1 Linear transformations of the data

For any linear combination of data,

$$\partial f(x)/\partial x = c$$
$$f(x) - f(y) = f(x - y)$$

where c is some constant. The first equation tells us that an individual data point can be assessed without reference to any other variable. Therefore all data points are a potential disclosure risk, and need to be assessed individually. The second equation notes that, if f(x) is a function which generates useful data when applied to a single observation, then there is a disclosure risk in the differencing of f(x).

All linear aggregates must therefore be classified as "unsafe": there is a high requirement on data checking, and a realistic risk of disclosure by differencing; and both of these are inherent in the mathematical form of linear combinations. This classification refers to all linear aggregates: tables, graphs, means, frequencies. It also covers quantiles, maxima and minima, which can be recast as tables.

This is why most SDDC literature in respect of the release of aggregate tables focuses on data problems, population uniques etc. The tables are linear combinations of data, and so cannot be made safe in their structure: safety must come through an appropriate choice of variables and sample. The alternative is to break the linear relationship between source data and output tables by, for example, recoding or rounding.

4.2 Linear regression coefficients

For a simple linear regression, consider the functional form of the estimated coefficients:

$$f(X, y) = \hat{\beta} = (X'X)^{-1} X'y$$

As Ritchie (2006) demonstrates, there is, in general, no danger from differencing; and the non-linear interactions mean that individual data points cannot be analysed. Hence this counts as a safe output.

This holds true for categorical variables as well as cardinal values. Although there appears to be a potential danger from differencing of models with categorical variables orthogonal to all others, the ability to identify observations relies on having the means available; and with the means available there are more direct ways to identify values.

There are some limited exceptions to be considered:

- If the explanatory variables are all categorical, then this is clearly a table and needs to be evaluated as such; or, if there are insufficient degrees of freedom for this to be a valid statistical model, exact values can be determined.
- If all the explanatory variables could be known to an intruder, then a value for an individual could be predicted; if the fit was particularly good, then potentially this could breach confidentiality restrictions by being close enough to a true value
- If the data comes from repeated observations on single unit, this could be informative, particularly in comparison with another unit

The first is simply a misclassification of a table as an analytical output. The second provides a theoretical problem, but in practice it seems that the fit needs to be infeasibly good (work by Statistics New Zealand suggests R^2 approaching 99%). Moreover, both a simple test for the accuracy of prediction and a counter-measure are easily available; see Ritchie (2006) for details.

The third exception is more interesting. While it is not clear what useful information could be derived, on a precautionary basis the VML currently bans regressions based on a single unit.

4.3 Cross-product and covariance matrices

Consider a cross-product matrix

$$M = X'X$$

This is an unsafe output. Frequencies and totals are identified by interactions with any constant or categorical variables. Hence this should be viewed as a linear aggregation.

Now consider the variance-covariance matrix generated by a simple regression

$$V = (X'X)^{-1}\hat{\sigma}^2$$

Should this be released? On the assumption that the estimated σ is available to the researcher, then it is a simple matter to turn V into a cross-product matrix, which is not safe. So this simple covariance matrix is unsafe.

However, this is not the case for the more general form

$$V = \left[(X'WX)(Z'Z)^{-1}(X'WX)\right]^{-1}\hat{\sigma}^2$$

Unless Z=X and W is the identity matrix, this cannot be unpicked. This holds even if W is known. This is a useful result because, for example, W will not be the identity matrix in any robust regression, let alone more complex models.

Can anything be inferred by combining V with the estimated coefficient vector? As:

$$V\hat{\beta}/\sigma^2 = (X'X)^{-1}(X'X)(X'y) = (X'y)$$

this is potentially a problem as a linear combination has been generated. Again, however, this is in general only true in the case of $V=(X'X)^{-1}\sigma^2$. For more complex forms of V, then the convolution of variables cannot be unpicked.

In summary then, variance-covariance matrices appear to be safe unless the model is simple unweighted OLS.

4.4 Herfindahl indices

The Herfindahl index reflects the dominance of one firm in an industry, as measured by turnover, employment etc:

$$H = \sum_i s_i^2 \qquad s_i = x_i / \sum_i x_i$$

On the face of it, this seems a safe output. As long as there are more than two firms in the market, individual values cannot be ascertained.

However, the use of the quadratic term causes a problem: unless the second largest firm is of a significant size, √H is a good approximate of the largest firm's share. The difficulty for SDC assessors is that the goodness of this approximation depends upon the relative sizes of the firms and the size of the tail. Table 2 illustrates this, with six sets of simulated values for the share of the two largest firms (S1 and S2) and the 'tail':

S1	***S2***	***S3-S50***	***H***	***S1-S2***	***Closeness of √H***
27%	1.5%	1.5%	.08	26%	8%
32%	20.0%	1.0%	.15	12%	20%
37%	15.0%	1.0%	.16	22%	10%
S1	***S2***	***S3-S10***	***H***	***S1-S2***	***Closeness of √H***
30%	10.0%	7.5%	.15	20%	27%
37%	15.0%	6.0%	.19	22%	17%
56%	40.0%	0.5%	.47	16%	23%

Table 2 H as an approximation to S1, the largest firm's share

Table 2 shows a range of values for the two largest firms and other firms in an industry, with 10 or 50 firms in the industry. There is not a simple relationship. Moreover, the last entry, which is safe in terms of the value of approximation of the largest value through √H, would usually fail a dominance test.

Therefore, although H is likely to be a safe statistic, it is difficult to state this categorically just on the value of H. The VML therefore allows Herfindahl indexes as long as the researcher demonstrates that

- there are more than two observations
- √H exceeds the largest value by a given percentage
- the dominance criterion is met

This is a pragmatic state of affairs. But it is not ideal: although these additional conditions do guarantee the safety of H, it requires three more pieces of information for researchers to provide and SDDC staff to check.

5 Conclusion

This paper has fleshed out some of the ideas in Ritchie(2007) about how to combine security, consistency and efficiency in a practical SDDC system. The examples here have demonstrated how a relative transparent assessment method can be applied to classes of output.

Many of the ideas here are already implicit in the SDDC manuals produced by NSIs; to some extent, the key purpose of this paper is to stimulate the development of a common framework for evaluating SDDC approaches. In the light of ongoing developments in creating RDC standards, it is intended that this approach be a step forward towards giving research outputs a 'risk rating', with the advantages that would give for establishing greater co-operation and transparency in RDC design.

Acknowledgements

I am grateful to Rhys Davies, Paul Allin and Philip Lowthian for comments.

References

Enright, J., McDonald, S., Corscadden, L., Jewell, E., O'Sullivan, J., Zeng, I., Nair, B., and Bentley, A. (2006) *Confidentiality Best Practices Manual*. Mimeo: Statistics New Zealand

ONS (2007) *Disclosure Control Standard for Business Surveys*. Mimeo: Office for National Statistics

Ritchie, F (2006) *Disclosure Control of Analytical Outputs*. Mimeo: Office for National Statistics

Ritchie, F (2007) *Statistical Disclosure Control in a Research Environment*. Mimeo: Office for National Statistics

VML (2007) *VML Default SDDC Methods*. Mimeo: Office for National Statistics

Integrated European Census Microdata (IECM) Samples: Enhancing the study of ageing with high precision over-samples of the oldest-old

Albert Esteve*, Joan Garcia*, Jeroen Spijker*, Robert McCaa**

* Centre d'Estudis Demogràfics, Campus Universitat Autònoma de Barcelona, Bellaterra, 08193, SPAIN aesteve@ced.uab.es, jgarcia@ced.uab.es, jspijker@ced.uab.es

** Minnesota Population Center, University of Minnesota

Abstract: A breakthrough in the tradeoff between privacy and data quality has been achieved for restricted access to anonymized population census microdata samples of Europe. As of September 2007, the IECM website, in partnership with the Minnesota Population Center, offers integrated microdata for 22 censuses, totaling more than 31,2 million person records, with 7 European countries represented. Over the next two years, the European collaboratory led by the Centre d'Estudis Demogràfics and the Minnesota Population Center, with major funding by the 6th Framework, United States National Science Foundation and the National Institutes of Health, will disseminate samples for more than 25 additional censuses. Thanks to high precision samples of one, five and even ten percent, much research on the ageing of populations can be accomplished. Nevertheless for studies of the oldest-old, over-samples of the elderly will often be required. The paper examines basic statistics on headship rates in the samples currently available and illustrates some of the limitations that can best overcome by oversamples for elderly populations.

1 The Integrated European Census Microdata (IECM) project

A vast quantity of census microdata covering Europe in the period since the 1960s survives in machine-readable form. Most of these data, however, remain inaccessible to researchers. The Centre d'Estudis Demogràfics located at the Universitat Autònoma de Barcelona along with other leading European research centers is involved in an international initiative, lead by the Minnesota Population Center (MPC), to create a harmonized and documented data series based on samples of over fifty Western and Eastern European censuses. The project is funded by the National Institutes of Health (NIH), a funding agency of the United States of America. In collaboration with the national statistical agencies of each country, we have negotiated redistribution agreements for the censuses of 17 European countries: Austria, Belarus, Bulgaria, Czech Republic, France, Germany, Greece, Hungary, Italy, the Netherlands, Portugal, Romania, Slovenia, Spain, Switzerland, Turkey and the United Kingdom (see Table 1). Combined, these countries account for over sixty percent of the population of Europe. The European census microdata series has an important chronological dimension. In all seventeen countries, the data span twenty years; for eight countries, thirty years; and for three countries, forty years. With over fifty million persons residing in over twenty million households, the integrated European dissemination system offer broader chronological scope and greater sample densities than any alternative data source. In most cases, the censuses are also the most representative source of information available about national populations.

Grants from the National Institutes of Health and the National Science Foundation of the United States cover the costs of finding and preserving microdata and documentation, negotiating dissemination agreements, developing data cleaning and sampling procedures, creating data conversion and dissemination software, and establishing design protocols for data and documentation. Additional funding from the Sixth Framework Programme cover the costs of

coordination, dissemination and harmonization tasks based in Europe. On March 2007, the first European samples were released. Data access is provided for Belarus (1999), France (1962, 1968, 1975, 1982, 1990), Greece (1971, 1981, 1991, 2001), Hungary (1970, 1980, 1990, 2001) Portugal (1981, 1991, 2001), Romania (1991, 2002) and Spain (1981, 1991, 2001). In 2008, samples for Austria, Netherlands and the UK will be launched.

Table 1. IECM-Europe: microdatasets entrusted by country, subsample precision and design For current data availability, see: http://www.iecm-project.org .

Datasets entrusted by subsample precision			Country	Sub sample design	2000s	1990s	1980s	1970s	1960s
10%	~5%	<=4%							
4			**Austria**	IPUMS	**2001**	**1991**	**1981**	**1971**	1961
1			**Belarus**	IPUMS		**1999**	1989	1979	1970
			Bulgaria (in process)		**2001**	**1992**	**1985**	1975	1965
	2		**Czech Republic**	IPUMS	**2001**	**1991**	**1980**	**1970**	1961
	5		**France** ('99 in process)	IPUMS	**1999**	**1990**	**1982**	**1975**	**1968, 62**
1			**Germany** (in process)	IPUMS	**2001m**	**1991m**	**1987, 81**	**1971, 70**	1961
4			**Greece**	IPUMS	**2001**	**1991**	**1981**	**1971**	1961
	4		**Hungary**	IPUMS	**2001**	**1990**	**1980**	**1970**	
			Italy (in process)		**2001**	**1991**	1981	1971	1961
		3	**Netherlands**		**2001m**			**1971**	**1960**
			Poland (negotiating)		**2001**		**1988**	**1978, 70**	1960
	3		**Portugal**	IPUMS	**2001**	**1991**	**1981**	1970	1960
2			**Romania** ('77 recovered)	IPUMS	**2001**	**1992**		**1977**	1965
			Slovenia (in process)		**2001**	1991	1981		
	3		**Spain**	IPUMS	**2001**	**1991**	**1981**	1970	1960
			Switzerland (in process)	IPUMS	**2000**	**1990**	**1980**	**1970**	1960
			Turkey (in process)		**2000**	**1990**	1985, 80	1975, 70	1965, 60
		2	**United Kingdom**		**2001**	**1991**	**1981**	**1971**	1961

Note: **bold country** = Agreement signed between University of Minnesota and National Statistical Authority
Year = census; **Bold year** = microdata survive; * = 100% microdata entrusted to IECM; m = microcensus
IECM systematic subsample design for private households: every n^{th} household stratified by enumeration district.

The database allows social scientists to make comparisions across European nations and opens the doors to European Statistical offices for transnational access to one of the most valuable assets for social science research. The database is an exceptionally valuable resource for researchers working on a broad range of topics. The large samples offered by European census microdata, are an invaluable resource compared to other sources such as demographic and labor forces surveys, which often offer greater subject coverage and detail, but less sample density, chronological depth, or geographic coverage. Thanks to high precision samples of one, five and even ten percent (see Table 1), much research on the ageing of populations can be accomplished. Nevertheless for studies of the oldest-old, over-samples of the elderly will often be required.

A comprehensive array of protections is in place to guarantee the privacy and statistical confidentiality of census microdata samples incorporated into the IECM database (McCaa, Esteve, 2006). These protections involve three elements:

1. legal: dissemination agreements between the University of Minnesota and each participating Official Statistical Institute.

2. administrative: licenses between the University of Minnesota and each user, specifying conditions and restrictions of use.
3. technical: perturbations of the data (swapping, recoding) to make exceedingly unlikely the identification of individuals, families or other entities in the data. Technical measures have the additional benefit that any assertion of absolute certainty in identifying anyone in the data is false.

The IECM project, by disseminating only integrated, anonymized microdata and restricting access to licensed academic users, shifts the risk-utility curve sharply upward. This approach provides access to microdata of high utility at the same time that confidentiality risks are minimized.

2 Enhancing the study of ageing with high precision over-samples of the oldest-old

2.1. Aging in Europe: the continuing need for better research and data

Europe is the continent with the highest proportion of elderly population, both in terms of the old (60-79) and oldest-old (80+). While the latter group only encompased 1,1% of the population in Europe in 1950, today they account for 3,5% and it is estimated that in 2020 5,1% will be 80 years or older and in 2050 9,6% (United Nations 2007). However, it is not only that a segment of the population is growing in both absolute and relative numbers because of falling birth rates, but it is also because they are doing better than ever in terms of lower levels of mortality (United Nations 2007) and morbidity (Murray and Lopez 1996). One consequence of the observed increase in people surviving to older ages is a much higher demands in health care (that are mainly for care, not cure), social welfare services, pension funds and housing specifically for elderly needs, etc (European Commission 2005), something that will only increase in future.

Besides the aforementioned growing welfare and housing needs, increased longevity has also had its impact on the oldest-old household composition and population structure by sex and marital status (more are still married, but there is also a growing group of divorced elderly). Changes in both household composition and marital status are, however, not only related to demographic factors, as changes in norms and values have made it both legally possible and socially acceptable for the widowed and divorced to remarry or to cohabit without formal marriage, whereby even among the oldest-old repartnering is no longer considered a taboo or something they themselves wouldn't consider given their age, at least for those who are in relatively good health (e.g. Lopata, 1996).

Finally, due to its cultural, historical and political diversity, there are still many differences between and even within the European countries as to both the household composition and welfare and household demands of the oldest-old population. One example of intra-European differences is with respect to the proportion of the oldest-old population that is institutionalised or that live in semi-autonomous elderly residential homes, which is more common in northern countries, while in the southern countries it is the family that is expected to take on the majority

of caring responsibilities of ailing or disabled elderly family members. In a similar manner, many rural areas, particularly those far away from mayor urban areas, lack the infrastructure for the supply of sufficient resources for the oldest-old, especially in the area of health care. Not surprisingly, population ageing and the continual increase in life expectancy, including of the oldest-old, has also excited a wide interest among researchers to study theories of ageing (e.g. Olshansky et al. 1990, 2005; Yashin 2001), to produce population projections of the oldest-old (e.g. Boleslawsky and Tabeau 2001; United Nations 1997, 2007), as well as to study specific implications of an ageing society in terms of changing household structure, changes in relationship forming, social policy, health care, disability and housing (see European Commission 2005).

However, a common problem is that there are few sources that permit a better and more accurate description of both current, past and changing household characteristics of the elderly, their living arrangements and living conditions. Moreover, the sample size for studies of the oldest-old is often small which undermines not only the research possibilities but also the reliability and validity of the obtained results. This is why we advocate greater use of census microdata as a means to study population characteristics of the elderly like their household structure, living arrangements and socioeconomic situation as it usually contains the entire population or a large sub-sample.

At the same time, we also advocate for future censuses where only low sample densities are made available for researchers that oversampling is applied to the oldest-old as this will provide the scientific community with an improved statistical basis for their research when the conventional sample size is not enough. Oversampling has been used in the past as a way to obtain a larger representation of small population groups that were of special interest for a particular study but where there were relatively few cases and therefore would cause larger statistical uncertainty in the results. For example, oversampling was recommended by Coleman et al. (2000) in order to better explore the mechanisms underlying some of the trends, patterns, and relations found in quantitative work on different types of relationship formations, that although still uncommon among the elderly today, is increasing in importance. This especially includes non-traditional living arrangements like Living Away Together or non-marital cohabitation after widowhood and divorce. In order to better understand the meanings of such experiences of people in the different cultural contexts (as we stress in this paper the importance of doing international comparisons) we could gain considerable insight from such qualitative approaches as in-depth interviewing when particular sub-sets of the population (in this case the elderly) is oversampled.

2.2. The IECM census samples and the oldest-old: Why are larger sample sizes needed?

To illustrate some of the limitations of the IECM census sample densities with regard to the oldest-old population, we examine basic statistics on headship rates in the samples currently available, and which can be best overcome by oversampling. Figure 1 shows the width of the 95% confidence intervals, that is, the range in percentage points between the lowest and the highest values. The confidence intervals have been calculated in the usual manner.

$$Z = 1{,}96 * \sqrt{\frac{N-n}{N} * \frac{p*(1-p)}{n}}$$

Where N equals to the total population; n to the sample size, and p to the observed rate.

We considered the most recent sample available of actual IECM countries that provided regional level data[1] (for this reason Hungary is excluded). For each sample we estimated sex specific headship rates by age, age and marital status, and age, marital status and region. With regard to the first case, we simply calculated confidence intervals by single age except for those aged above 95 years, which we grouped into an open ended age group. Secondly, we computed the width of the confidence intervals for each combination of age and marital status (single, married, divorced and widowed). The results that are given in the figure are weighted averages of the four marital statuses by age. A similar approach was applied to the age, marital status and regions combinations[2]. The regional dimension was included to test the possibilities of carrying out sub-national analyses.

As to be expected, confidence intervals increase with age and are wider for men than for women. Due to the specific old age population sizes, confidence intervals really begin to increase between ages 75 and 85. At the national level, with or without the inclusion of marital status, the range of the confidence intervals rarely exceeds 10 percentage points, except at very old ages. The lowest variations of headship rates are found among French and Spanish women, while Portuguese men generally show the highest range in age-specific confidence intervals. The impact of including a regional dimension on the width of the confidence intervals depends on the number of regions available in the sample for each country. By definition, the confidence interval is larger when region is included. In general, analysis by region increases the width of the confidence intervals to above 10 percentage points and among the oldest-old to more than 20 percentage points.

As noted above, one way to overcome the statistical limitations of small sample sizes that particularly affect old age groups is to over-sample this subset of the population. In our example, we show the effects of increasing sample sizes two-, three-, four- and five-fold, on the confidence interval widths for the sex specific headship rates by age, marital status and region (see Figure 2). As is clearly shown in the figures, increasing sample size dramatically shrinks the width of the confidence intervals, especially for those countries with either a small elderly population and/or small sample densities. However, due to the high variability in confidence intervals by age, sex and country it would be better to "customize" the specifications of high precision samples for the oldest-old. For example, according to the results for France and Portugal, oversampling would need to be performed across all ages, but in the case of Belarus and Rumania, perhaps only for ages 80 and over. One should bear in mind, however that this applies to age, sex and marital status headship rates that we used as an illustrative example. Other types of variables might yield smaller or even greater confidence intervals.

[1] Belarus 1999 (10%), France 1990 (4,2 %), Greece 2001 (10%), Portugal 2001 (5%), Romania 2002 (10%), and Spain 2001 (5%).

[2] Belarus 1999: 6 regions; France 1990: 22 regions; Greece 2001: 54 regions; Portugal 2001: 7 regions; Romania 2002: 8 regions; Spain 2001: 52 regions.

Figure 1. Width of the confidence interval (in % points) of headship rates by age, marital status and region

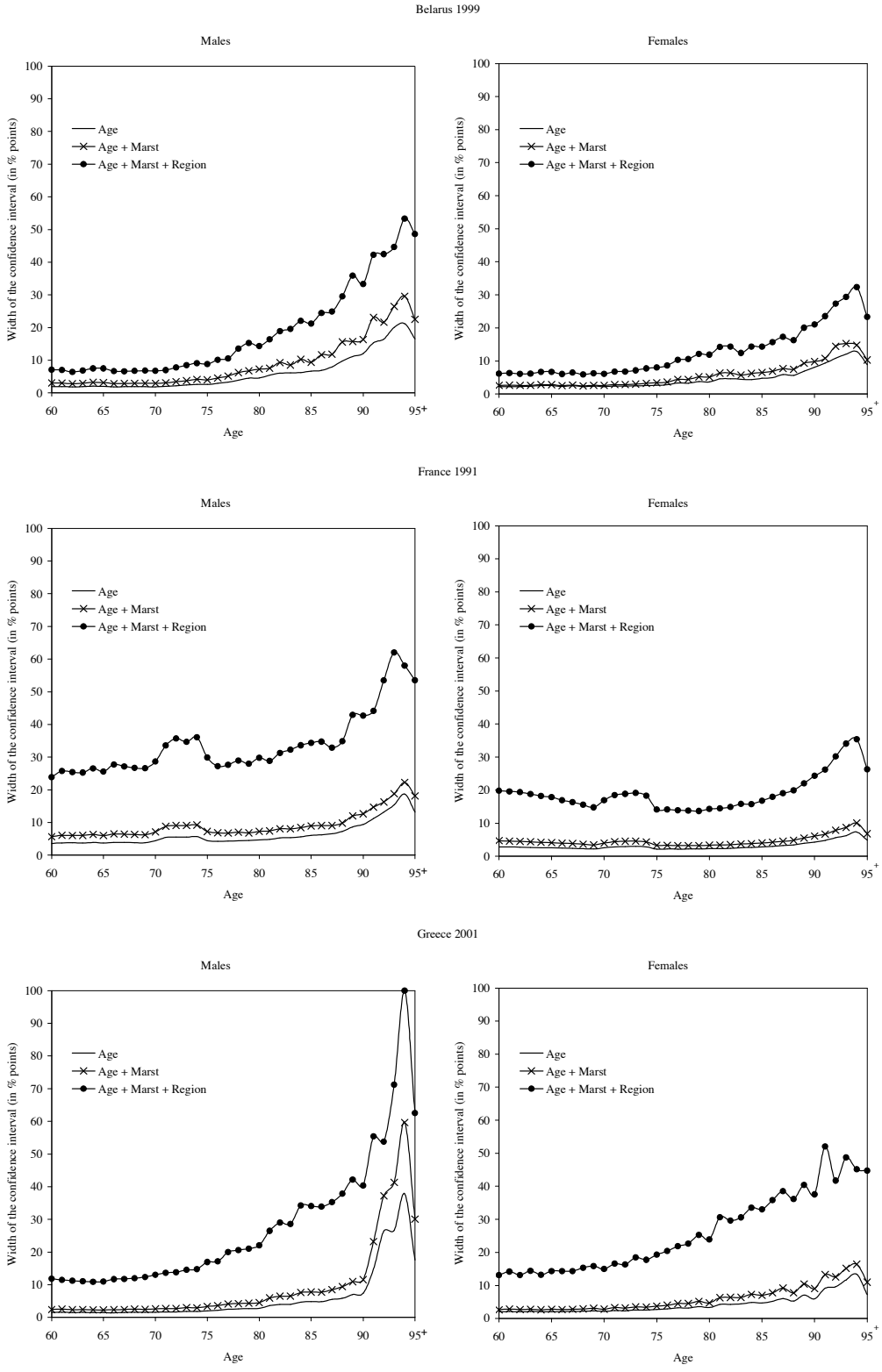


Figure 1 (continuation). Width of the confidence interval (in % points) of headship rates by age, marital status and region

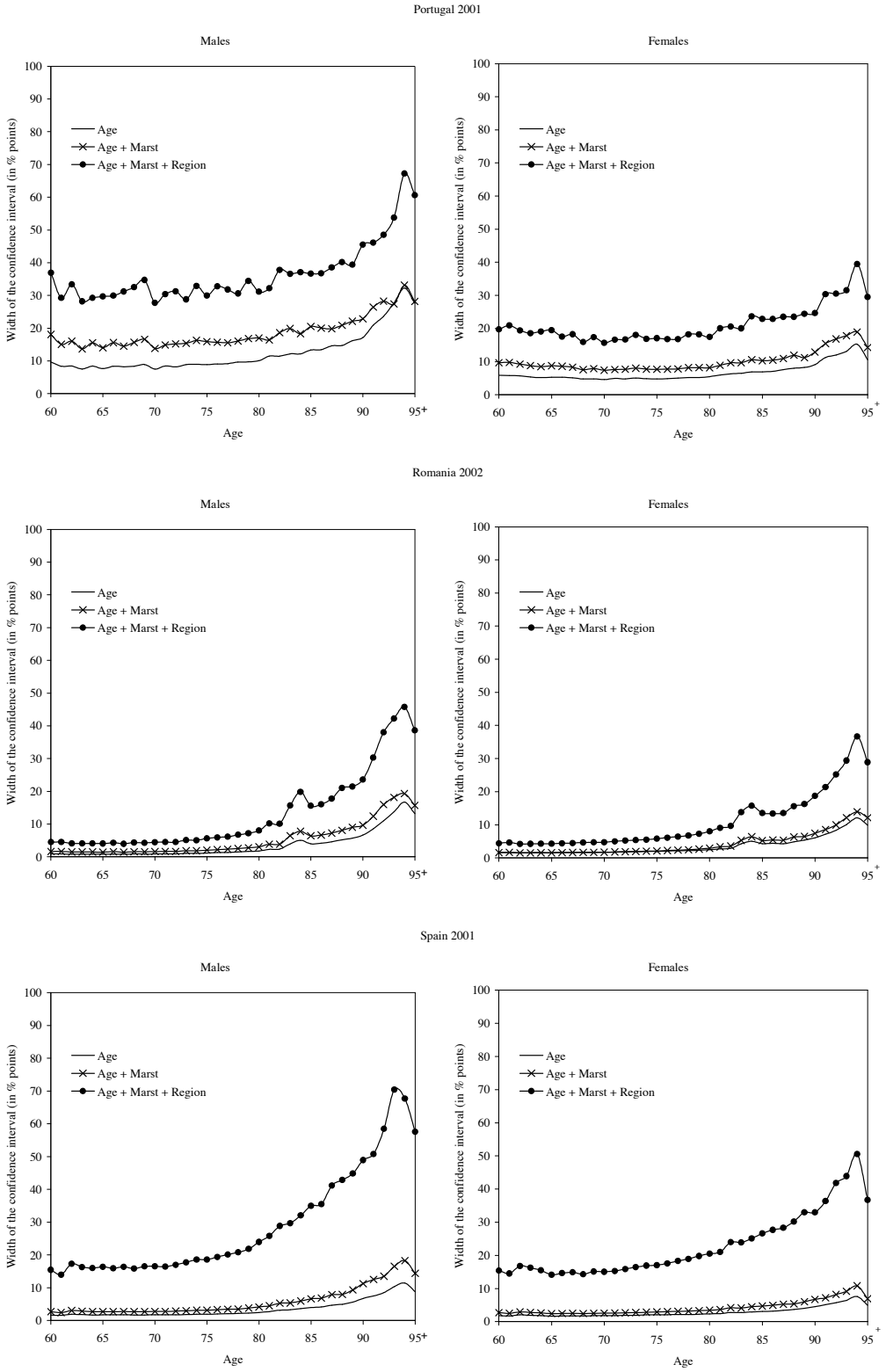


Figure 2. Width of the confidence interval (in % points) of headship rates by age, marital status and region according to hypothetical sample size increases

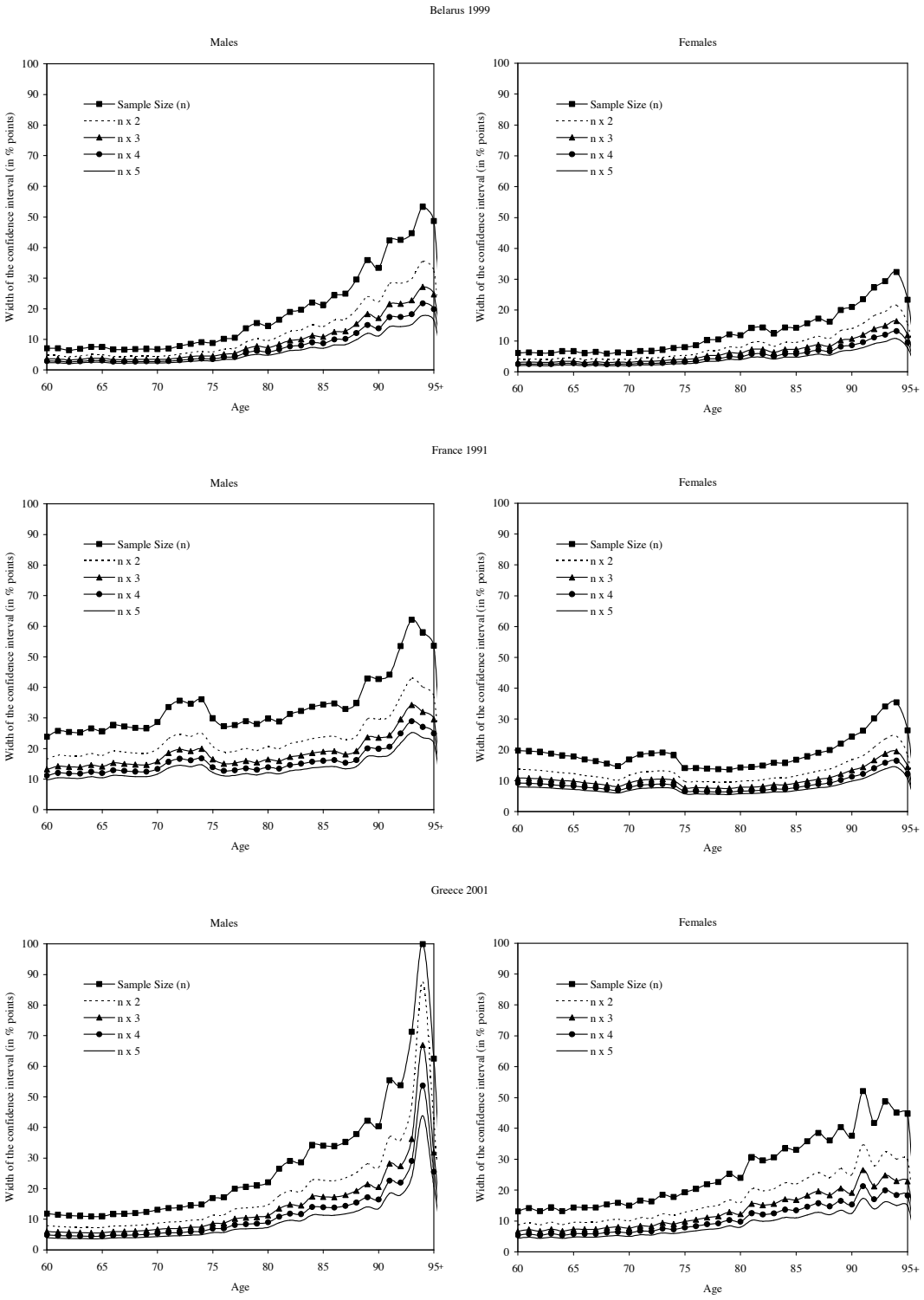
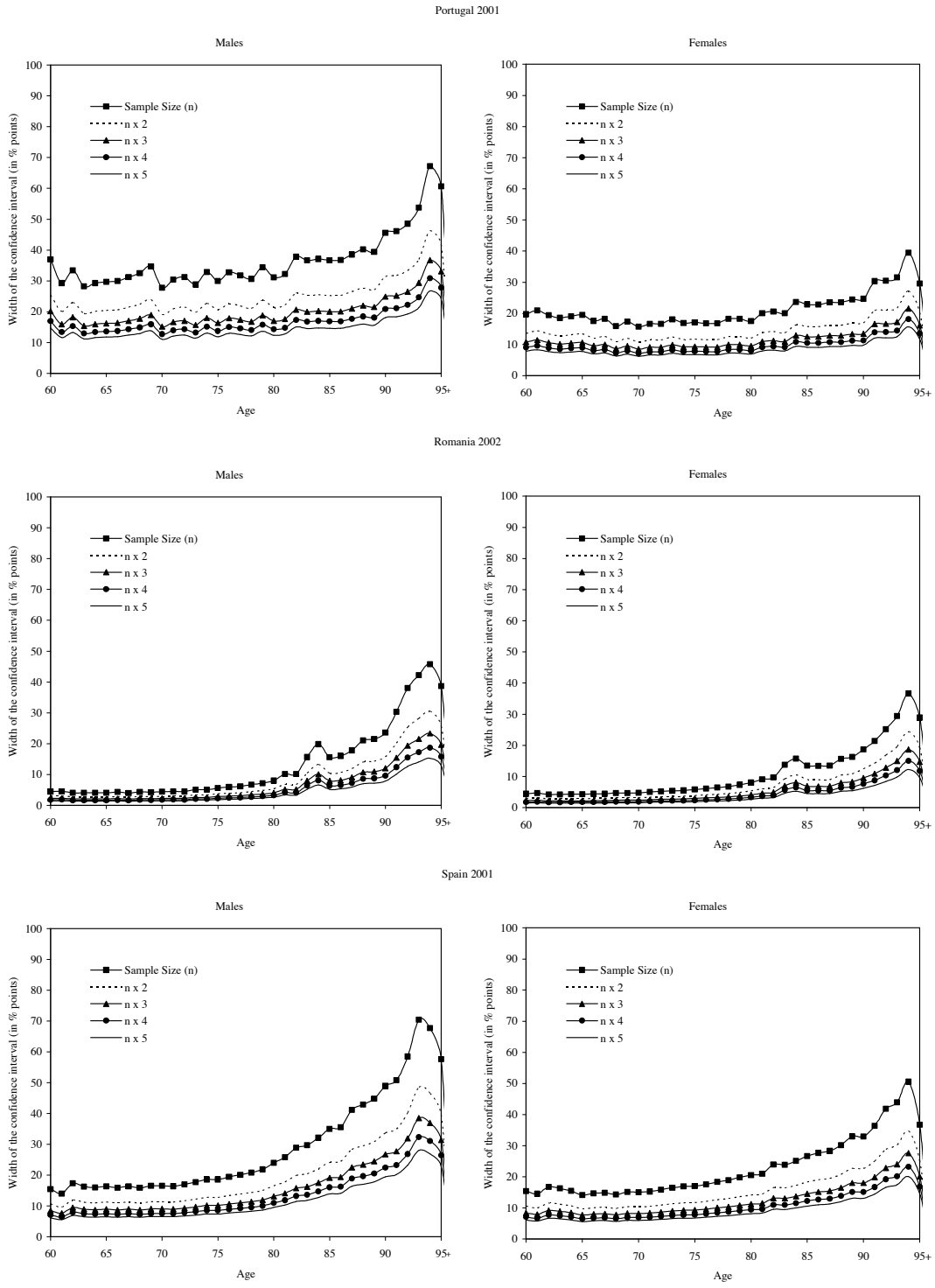


Figure 2 (Continuation). Width of the confidence interval (in % points) of headship rates by age, marital status and region according to hypothetical sample size increases



3 Conclusions

The IECM and IPUMS projects are fortunate to enjoy the support of many of the most respected National Statistical Offices in Europe. High precision household samples of 5-10% have been entrusted to the Barcelona-Minnesota team by the NSOs of eleven countries--Austria, Belarus, Czech Republic, France, Germany, Greece, Hungary, the Netherlands, Portugal, Romania, and Spain--and are expected soon from an additional four countries--Bulgaria, Italy, Switzerland, Turkey and the United Kingdom. While sample densities of 5-10% are adequate for most research purposes, this paper has shown that even higher densities—double, triple or even greater oversamples—are essential for the study of such important population sub-groups as the elderly. Otherwise sampling error alone will be so great as to deprive the results of significance. The Population Activities Unit (ECE), a precursor to the IECM project for the 1990 round of censuses, developed over-samples of up to 50% for Bulgaria, Czech Republic, Estonia, Finland, Hungary, Latvia, Lithuania, Romania, and Switzerland. In a second phase of the IECM initiative, we invite European NSOs to consider enriching the sample densities within a regimen of stringent controls for statistical confidentiality. By restricting access to a class of academic users, high-density microdata extracts can be provided to researchers at vanishingly low risk.

References

Boleslawsky L., E. Tabeau (2001), Comparing theoretical age patterns of mortality beyond the age of 80. In: E. Tabeau, A. Van den Berg Jeths, C. Heathcote (eds.), Forecasting mortality in developed countries: Insights from a statistical, demographic and epidemiological perspective, pp. 127-155. Kluwer Academic Publishers, Dordrecht.

Coleman, M., L. Ganong and M. Fine (2000). "Reinvestigating Remarriage: Another Decade of Progress." Journal of Marriage and the Family 62(4): 1288-1307.

European Commission (2005), Network for the Integrated European Population Studies, NIEPS project final report, Brussels, EUR n° 21529, ISBN 92-79-00426-3.

Lopata, H. Z. (1996). Current Widowhood: Myths & Realities. Thousand Oaks, CA, USA: Sage.

McCaa, R., Esteve,A. (2006) 'IPUMS-Europe: Confidentiality measures for licensing and disseminating restricted-access census microdata extracts to academic users' in *Monographs of official statistics*, Eurostat, pp. 37-47.

Murray, C.J.L., A.D. Lopez (eds.), Global burden of disease and injury series: volume I: The global burden of disease: a comprehensive assessment and disability from diseases, injuries, and risk factors in 1990 and projected to 2020.

Olshansky S.J., Carnes B.A. & Cassel C. (1990) "In search of Methuselah: estimating the upper limits to human longevity" Science 250:634-640.

Olshansky S.J., Grant M., Brody J. and Carnes B. (2005) "Biodemographic perspectives for epidemiologists" Emerging Themes in Epidemiology 2(10) (http://www.ete-online.com/content/2/1/10).

United Nations (2007), World Population Prospects: The 2006 Revision. [data downloaded from http://esa.un.org/unpp on November 05, 2007].

United Nations (1997), Projecting old-age mortality and its consequences. Report on the Working Group, New York, 3-5 December 1996.

Yashin A. (2001), Mortality models incorporating theoretical concepts of ageing. In: E. Tabeau, A. Van den Berg Jeths, C. Heathcote (eds.), Forecasting mortality in developed countries: Insights from a statistical, demographic and epidemiological perspective, pp. 261-280. Kluwer Academic Publishers, Dordrecht.

IV

Panel discussion on microdata protection versus remote access facilities

WP.39
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
(Manchester, United Kingdom, 17-19 December 2007)

Topic (iv): Panel discussion on microdata protection versus remote access facilities

PANEL DISCUSSION

MICRODATA PROTECTION VERSUS REMOTE ACCESS FACILITIES

Chair: Jane Longhurst, Office for National Statistics, United Kingdom

One of the main functions of National Statistical Institutes (NSIs) is to publish detailed information on a large spectrum of aspects of the society. For this they collect large amounts of data via questionnaires and by using other registers or administrative sources leading to very rich databases. These databases are the main source for compiling their statistical publications.

The classical outputs are sets of marginal tables, published in volumes accessible by all users. However, these large databases can serve a second very important need. The rich databases are ideal material for performing statistical research. The results of these research projects give new insights and contribute to the well-being of modern societies.

Therefore it is a natural task for the NSIs to facilitate this research in the most effective way by providing access to these microdata files. But on the other hand the NSIs have a strong commitment to safeguard the privacy of the individual respondents in their databases. This commitment is not only an obligation from a Statistical Act in many countries, but also part of the ethical code for statisticians. An important side effect is also to guarantee the cooperation of the respondents in the future. This is why NSIs take confidentiality issues very seriously.

Within this discussion session specific focus will be given to three different approaches to managing the disclosure risk of microdata releases; licensing, masking techniques and remote access. Each approach has pros and cons and impacts on the microdata detail and therefore subsequent analysis in different ways. Panelists will provide an overview of their experience and thoughts on microdata access and participate in discussion from the floor. The aim will be to determine future directions on this issue and establish where research should be concentrated.

In order to stimulate discussion the thoughts and position of each panel member are provided below.

Paul Jackson, Office for National Statistics, Segensworth Road, Fareham, UK,
paul.j.jackson@ons.gov.uk

- Decisions about modalities for research data access should **begin** with consideration of the purpose for providing access.

  UK experience (and we are not alone) is that the demand for a sophisticated quantitative evidence base for policy making is increasing dramatically. Pre-defined standard tabular outputs can not meet this demand.

  Interdisciplinary (cross-cutting) and longitudinal analysis is needed to describe complex phenomena such as 'social capital' or 'child development and well-being'. Micro-data is essential to this research.

  Regional policy requires a tailored response to local issues. A central NSI can not hope to meet the specific demands of local evidence-based policy making.

  Analysis for policymaking is not founded on one-off reports, but is continuous.

- Decisions about modalities should then **acknowledge** the increasing competence and capacity in data management and analysis outside the NSI.

  Typically, academia now rivals or exceeds the statistical analysis competence of NSIs.

  Concepts for social and economic analysis (such as social capital) often emerge from non-government sources.

  Data management standards, organisational and technical, are now excellent in most universities and research institutions.

  The people of the analytical professions move much more freely now between NSIs and academia as their employers, taking their competences with them and spreading their skills amongst their peers.

- Decisions should then **recognise** that:

  Central, regional and local government will source analysis from the place best able to provide it - there is no closed shop for analysis.

  NSIs can not meet their wider objectives without enabling others to join in the description of the economy, society, and environment.

  NSIs have the privilege of lawful authority for original data collection. This privilege should benefit as many researchers as possible.

- These considerations lead us to **challenge some orthodoxes** for research data access methods:

  Fully anonymising micro-data before access will prevent good cross-cutting analysis and data linking.

  Removing local geographic identifiers will prevent sub-regional analysis.

  Sophisticated analysis can not be done through the letter-box of remote job submission.

  Access needs to be continuous for analysis supporting on-going policy development and evaluation.

  The people outside the NSI, after screening, can be as safe as the people inside the NSI, as are their facilities.

  The volume of research needed by public policy making means that checking every output can exhaust the resources of the NSI.

- **UK is working through this analysis and we have some conclusions:**

  Our new law concerns itself with the 'fitness and properness' of the researcher. A fit and proper researcher can have access to any data necessary for the described project.

  We will require researchers to check their own outputs (but we will provide standards to follow, and a support service)

  We will operate a three-tier model of facilities - data archive, academic laboratory/approved off-site environment, and on-site laboratory.

  We see little demand for remote job submission and have no plans to develop such a facility in the foreseeable future.

  We do see a role for remote access technologies in separating access from geography

Luisa Franconi, Istat, DCMT, Via Cesare Balbo 16, 00145 Roma, Italy, e-mail franconi@istat.it

The need to maintain a diversity of types of access: different users, different needs, different access

Masked microdata sets obtained through the application of different protection methods offer an alternative to tabular outputs. These could take the form of Public Use File (PUF) available for all users or Microdata File for Research (MFR) dedicated to scientific research. Besides these products, recently many agencies are offering services to allow access to microdata sets, obtained by suppressing all direct identifiers in the original microdata, for scientific research purposes. These services take the form of Data Analysis Centres where only the output of the analysis of each researcher (and not the input microdata) is reviewed for confidentiality checks. Finally, some agencies have started offering also remote access or remote executions services for scientific research purposes to their survey microdata, sometimes applying some disclosure techniques. Although this latter type of access is extremely appealing for both researchers and statistical agencies this could not substitute the microdata products mentioned before for different reasons. First, there is a reason stemming from the right of access. All citizens have the right of access to information so, if from one side it is important to give priority to research, from the other we cannot forget other categories of users who may need microdata. Among them there are students and the global project of teaching statistics to train next generations to make further use of statistical data in more and more aspects of society. There are also analysts, marketing experts or even lawyers who may not appeal to the scientific research goal but for whom the data are important. Therefore it is crucial for the society as a whole to allow different channels of microdata access not to discriminate different types of research or different needs of the citizens. Secondly, remote access requires, from the point of view of the agencies, dedicated staff with experience in reviewing output of researchers. Two problems may arise here. Staff is costly and, in times of strict financial laws for public administration, it is hard to ask agencies to invest on this sector although crucial for research. So, if the initial burden of setting up such systems of remote access can be taken by the statistical agencies, in the long term is absolutely necessary to gain additional forms of financing in order to achieve sustainability in management. Additionally, the staff we are looking for managing such systems is well trained and with experience. This is demanding for many reasons: first there aren't still available general rules and guidelines for output checking consequently there is lack of automatic ways of checking complex output. These experiences still need to be built. This is a long process and problems here need to be addressed. Resources are needed also in the creation of masked microdata files but, in this case, it is more a one case necessity rather than a continuous process. Therefore, although appealing, remote access should not be the only way to give access to microdata.

Future research: data utility and user needs

Among the several item in the research agenda I see two main issues that, in my opinion needs, to be tackled. The first issue is improving data utility as one dimension of quality. We are releasing nowadays microdata sets that wouldn't have been released few years ago. This is because there has been a big research effort in defining what was more essential for the statistical agencies, preserving confidentiality, in order to increase the quantity of microdata offered to users. Measures of risks have been defined and procedure to assess them have been implemented, protection methods developed. Now it is high time to seriously improve further the quality of the products we are offering. One essential dimension of quality in the area we are dealing with is data utility as is it a duty of the data provider to give practical, relevant and useful measures of goodness of the released microdata to guide the users in measuring reliability of their analysis. Thorough investigation of the effects of protection methods is still lacking and definitions of methods that preserve essential statistics should be fostered. This will allow to put into practice the right balance between risk and utility in microdata release. The second issue is trying to address better user needs. On example of a need that is emerging in Europe is the level of the geographical details in released microdata. More and more studies want to investigate phenomenon at very detailed geographical level to have a clear map of the different characteristics of small areas and how this areas influence social and economical behaviour. Most surveys though cannot possibly be significant at such level. As territorial information are extremely identifying protection methods that still produce useful data while protecting confidentiality need to be further investigated together with clear guidelines on the use of such information and their statistical errors. Another clear user need is improving access to different types of data such as enterprise microdata as well as panel and longitudinal microdata.

Addressing these needs and finding ways to allow better access to such data is certainly the current challenge.

Anco Hundepool, Statistics Netherlands, Division of Methodology and Quality, P.O. Box 4000, 2270 JM Voorburg, The Netherlands, Email: ahnl@cbs.nl

History

When PCs became available and statistical analysis of micro data was possible for researchers, the need for access to micro data grew. In the beginning the NSIs were very reluctant to give access to these individual databases because of confidentiality reasons, but gradually methods have been developed to assess the disclosure risk and also to avoid disclosure. Methods like global recoding, local suppression and several perturbative methods. Many of these methods have been implemented in μ-ARGUS.

Special databases have been developed for researchers, who, after signing heavy contracts and under special conditions, could analyse these research datafiles on their own computers. These projects have been very successful, but on the other hand the need for more detailed datafiles remained.

The next step was to open Research Data Centres (RDCs), also called OnSite laboratories, etc. Different names, but the principle was the same. Researchers were given access to the rich less protected datafiles on computers within the premises of the NSIs. Without any possibilities to bring any material outside of the centre, the researchers could analyse these datafiles. The drawback is that any output they want to take home, has to be checked by the NSI-staff for disclosure risk. This is a heavy burden, but on the other hand it has made many research projects possible. Another drawback is that researchers have to travel many times to the premises of the NSI and the NSIs have to reserve expensive locations for these RDCs.

As modern information technology progresses, it is now possible to build secure connections over the internet, enabling Remote Access to Statistical Information. This has led to investigations to see whether these connections could be used as an alternative for the RDCs. The first prototypes for this have been built and the results look promising. Time is now ready to further investigate these possibilities and use it on a much wider scale. Giving access to databases to researchers in other countries or even to European databases now becomes within reach.

Conclusions

Modern researchers require easy access to the statistical databases for their research, while the NSIs have to preserve the confidentiality aspects of the respondents. Producing confidentialised microdata files for research has been the answer to these research needs in the eighties and nineties. Also safe research data centres, within the premises of the NSIs, have been an answer, but modern information technology makes it possible to build safe OnLine Access via the internet.

This approach is becoming very popular in a short period of time. The researcher can work from his own institute without travelling, while the NSI can still control the confidentiality of the output, as all results are still checked for confidentiality by the NSI. The production of protected micro data files will be obsolete very quickly, as all researchers will go for the Remote Access approach!

References

Hundepool, Anco, Josep Domingo-Ferrer, Luisa Franconi, Sarah Giessing, Rainer Lenz, Jane Longhurst, Eric Schulte Nordholt, Giovanni Seri, Peter-Paul De Wolf (2006), *CENEX handbook on Statistical Disclosure Control,* CENEX-SDC project, http://neon.vb.cbs.nl/cenex/CENEX-SDC_Handbook.pdf

Wolf, Peter-Paul and Anco Hundepool (2007), *Remote Access (not) at Statistics Netherlands*, ISI-session, Lisbon

V

Panel discussion on balancing data quality and confidentiality

WP.40
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
(Manchester, United Kingdom, 17-19 December 2007)

Topic (v): Panel discussion on balancing data quality and confidentiality

PANEL DISCUSSION

BALANCING DATA QUALITY AND CONFIDENTIALITY

Chair: Lawrence H. Cox, National Center for Health Statistics, United States of America

Quality and Confidentiality Issues Associated with Tabular Data

Lawrence H. Cox

We separate issues surrounding the interplay between data quality for tabular data into two categories: local data quality and global data quality.

Local data quality is focused on characteristics of individual or small sets of cells. These include individual cell values and/or associated time trends (e.g., income or number of employees), or of quantities computed from two or more cell values (e.g., income per employee). Preserving local quality amounts to preserving local characteristics while satisfying confidentiality demands—or, balancing the two.

Global data quality is focused on characteristics of the data set or subsets. These include distributional parameters such as mean and variance, statistics such as quantiles, and distributional shape. Preserving global quality amounts to preserving global characteristics while satisfying confidentiality demands—or, balancing the two.

Local and global quality are not mutually exclusive. Two examples of the overlap are measurement error and correlation. Preserving all or most individual values to within measurement error arguably is perfect local quality, and arguably can be expected to preserve distributional parameters, statistics, and shape. Ensuring nearly perfect correlation between original and masked values is expected to preserve both local and global quality. Other examples are preserving time trends, rank order statistics, and usability/analyzability. These are examples of preserving quality in a *univariate* setting, viz., involving a single variable such as income or number of employees.

Local and global quality can also be *multivariate*, involving two or more variables such as income and number of employees. In addition to preserving univariate quality for each variable, in a multivariate setting the objective is to preserve relationships between variables such as covariance and regressions.

We observe the following data quality characteristics of familiar SDL methods.

Complementary cell suppression For univariate data, CCS preserves local quality perfectly for unsuppressed cells, poorly for suppressed cells; inhibits analyzability; pokes holes in the distribution. For multivariate data, its negative effects are magnified significantly.

Random rounding and perturbation Can preserve local and global quality well for both univariate and multivariate data.

Controlled tabular adjustment Can preserve local and global quality well for both univariate and multivariate data; QP-CTA and MDI-CTA have nearly opposite data quality strengths and limitations.

Perturbing underlying microdata Insufficient information to judge.

Data Masking Procedures for Numerical Microdata

Krish Muralidhar, University of Kentucky, Lexington KY 40506

Recently, there have been considerable developments in procedures that are used to mask numerical microdata. In this discussion, we attempt to assess the relative strengths and weaknesses of these procedures for the most general case in which there exists a data set consisting of both categorical and numerical variables. Since our focus is on masking numerical microdata, all categorical variables are treated as non-confidential or assumed to have been masked and some (or all) of the numerical variables are deemed confidential.

The primary objective of any masking procedure is to prevent disclosure of confidential information. We use the following definition of disclosure: "If the release of the data allows an intruder to estimate either the identity of an individual or the value of a confidential variable with a greater level of accuracy than before the release of the data, then disclosure has occurred." This definition is a combination of the disclosure risk definitions of Dalenius (1974) and Duncan and Lambert (1986). The risk of disclosure resulting from the masking procedure can be assessed as follows: "(Assuming that aggregate information regarding the entire data set and the confidential variables are already available to the user) Does the release of the masked microdata improve the user's predictive ability?" We can show that the answer to this question is "No" if, given the non-confidential variables, the original and masked confidential variables are (conditionally) independent. In this case, we can conclude that the masking technique minimizes disclosure risk since the release of the masked data does not improve the predictive ability of the intruder.

The secondary objective of any masking procedure is to provide the highest level of data utility (or lowest level of information loss) when the masked data is used in place of the original data. In order to achieve the lowest information loss, it is necessary that the response to any arbitrary query using the masked data should be identical to that using the original data. In practice, this measure is difficult to quantify and we usually employ empirical measures. These include the univariate characteristics of the confidential variable, linear and non-linear relationship measures between the variables, ability to maintain subset characteristics, and the ability to reach the same inferences using the masked data as that using the original data. Depending on the specific context, it is possible that alternative measures of data utility are used. Once we have assessed the masking technique on the primary objectives (disclosure risk and data utility), then we can use additional objectives such as ease of implementation, ease of use, etc.

It is often argued that there is an implicit trade-off between disclosure risk and data utility. This is not always true. Those techniques that minimize disclosure risk do not have this inherent trade-off. Only techniques that do not satisfy the minimum disclosure risk requirement have a trade-off between disclosure risk and data utility. Finally, the evaluation approach that we have presented allows us to identify future research opportunities.

1. Dalenius, T. 1977. Towards a methodology for statistical disclosure control. Statistisktidskrift 5 429–444.
2. Duncan, G. T., D. Lambert. 1986. Disclosure-limited data dissemination. J. Amer. Statist. Assoc. 81 10–18.

Quality and risk for different classes of synthetic data

Jerome P. Reiter

Synthetic data come in two flavors: fully synthetic or partially synthetic. To construct fully synthetic data, the agency (i) randomly and independently samples units from the sampling frame to comprise each synthetic data set, (ii) imputes the unknown data values for units in the synthetic samples using models fit with the original survey data, and (iii) releases multiple versions of these data sets to the public. Partially synthetic data comprise the units originally surveyed with only some collected values replaced with multiple imputations. For example, the agency might simulate sensitive variables or quasi-identifiers for units in the sample with rare combinations of quasi-identifiers; or, the agency might replace all data for selected sensitive variables or quasi-identifiers.

Selecting one of these two strategies involves a complex trade-off between disclosure risk and data usefulness. For fully synthetic data, identity and attribute disclosure risks are in general very small because the original units are not released. However, the validity of inferences with the synthetic data depends fully on the validity of the imputation models. The released data can be a simple random sample, eliminating the need for analysts to worry about the typically complex sampling design. But, it may be necessary to have large synthetic sample sizes or numbers of implicates to get inferences with good properties. For partially synthetic data, risks are higher since the collected units remain on the file. However, maintaining some genuine data weakens the reliance on imputation models for valid inferences. Analysts must worry about the original complex sampling design for inferences. A small number of datasets may be adequate when replacing only a fraction of the data.

Synthetic data involve another trade-off that appears most obviously in partially synthetic data: selecting the number of datasets to release. Increasing the number of datasets generally improves inferences (e.g., lowering variances and making normal approximations more plausible). But, it also increases disclosure risks, as intruders' can refine their guesses at the true values with more information. One approach is to release different numbers of replacements for different values, for example few replicates of highly identifying variables and more replicates of weakly identifying variables.

I believe that our current measures of data usefulness do not reflect all aspects of inference. We often focus on point estimation, when the real target should be inferential validity (e.g., 95% confidence intervals cover at least 95% of the time). With a few exceptions, we focus on the first two moments when many analyses rely on other features of distributions. We do not account for complex sampling designs and weights when evaluating (and proposing) disclosure limitation procedures, when these are the norm rather than simple random samples. Finally, we evaluate posited models when we also should examine the model-building process (e.g., do we get the same transformations, interactions, and variable selection). Developing usefulness metrics that incorporate these features is a challenging and important area of research.

European Commission

Work session on statistical data confidentiality – Manchester 17-19 December 2007

Luxembourg: Office for Official Publications of the European Communities

2009 — 430 pp. — 21 x 29.7 cm

ISBN 978-92-79-12055-8

How to obtain EU publications

Our priced publications are available from EU Bookshop (http://bookshop.europa.eu), where you can place an order with the sales agent of your choice.

The Publications Office has a worldwide network of sales agents. You can obtain their contact details by sending a fax to (352) 29 29-42758.

KS-78-09-723-EN-N

ISBN 978-92-79-12055-8